University Utrecht

Artificial Intelligence

Master's Thesis

The Relevance of Carnap's Inductive Logic for Supervised Classification

Author:

Wessel de Jong

Supervisor:

Dr. Janneke van Lith

Dr. Sander Beckers

March 19, 2019

Abstract

Recent advances in the field of machine learning have cast new light on the long-standing debate within the philosophy of science between falsificationists and inductivists. As a result of many failed attempts at justifying inductive inference, falsificationism has tended to be preferred over inductivism. However, Gillies claims that the success of certain machine learning procedures necessarily leads to the conclusion that inductivism has shown to be successful [19]. In this thesis I will challenge this claim by examining how criticism voiced against the inductive logic of one of the central figures of the inductivism, Rudolf Carnap, might still be relevant for modern day machine learning procedures. Since a general claim about all machine learning procedures would be prone to counterexamples this thesis will be limited to only linear models for supervised classification. I will consider how three well-known critiques that strongly undermined Carnap's inductive methods might still apply to such classification models.

Contents

1 Introduction

Since its emergence in the late 1970s machine learning has been developed with considerable success. Recent years have shown an ever growing applicability of machine learning methods and with this growth came many triumphs. Such triumphs have had a tremendous impact on engineering, industry and the economy. In addition to the significant impact machine learning has had in these fields, its methodologies have transformed many fields of science. While machine learning procedures used to only be applied in isolated predictions problems, nowadays they play a crucial role in many scientific endeavours. With an ever growing presence of machine learning methods in various field of science, more and more researchers and philosophers have come to the realization that these methods might have the potential to transform science at a fundamental level.

One of the first to address the consequences of the successes achieved by machine learning for the way we conduct science was Donald Gillies [19, 21, 21]. Gillies states that the advances in machine learning have had important implications for the scientific method. Among other things, these advances might cast new light on long-standing philosophical debates. Similar conclusions have been drawn by Williamson, who argues that the growth of machine learning presents the philosophy of science with inescapable lessons for the study of the scientific method [46, 45]. Both Gillies and Williamson specifically refer to one debate in the philosophy of science for which advances in machine learning might have far-reaching consequences, namely the debate on inductivism versus falsificationism. This debate, also called the inductivist controversy, has played and still plays a central role in the philosophy of science.

The inductivist controversy is centered around the question of how scientist should make discoveries and how the goals of science can be achieved. This controversy has been carried on for many years between supporters of falsificationism and supporters of inductivism. The first state that science proceeds by first making conjectures, a procedure that cannot be automated, after which the predictions of this conjecture are tested by observing and experimenting. The second state that science proceeds by first making a large number of observations after which laws can be extracted in a manner that could potentially be automated. Thus, in contrast to falsificationism, which is claimed to solely apply deductive inference, inductivism

relies on inductive inferences where general truths are extracted from a set of particular observations. This puts proponents of inductivism in the cumbersome position of having to justify inductive inference by showing how the truth of its conclusion is guaranteed. After all, many examples can be given of such inferences leading to false conclusions.

In the early twentieth century logical positivists proposed that there actually was a way to justify inductive inference. As a founding father of logical positivism, Rudolf Carnap [7, 8], devoted himself to developing an inductive system that could be used as a tool for furthering the sciences [40]. The inductive logic he attempted to develop was supposed to present a mathematical method with which one could judge in a quantitative way how much an hypothesis is confirmed by available evidence and background knowledge [36]. Using this method multiple hypotheses could be compared on their degree of confirmation. Although we might never be able to judge a general hypothesis to be true based on a set of observations, we can find the hypothesis that is most confirmed by the available evidence. In this way we could still get to a general conclusion from a set of observations.

Over many years, Carnap's attempts of developing an inductive logic have been met with heavy critiques. One of the most prominent opposers of his project, Karl Popper, claimed the inductive method was a myth that could never be a way of practicing science [37]. Others revealed that Carnap's proposals allow for multiple inconsistencies and irrational reasoning [26]. Consequently, none of Carnap's theories of inductive logic are viewed by Carnap himself or others to be fully adequate for the purpose of grounding his conception of inductive logic [17]. And as a result of many more failed attempts at justifying inductive inference, falsificationism has tended to be preferred over inductivism.

However, it has been claimed that the many successes achieved by machine learning since its early days suggest that inductivism remains a plausible stance. Gillies [19] claims that the success of certain machine learning procedures necessarily leads to the conclusion that inductivism has shown to be successful. He states that:

> "*[..] programs have been written which enable computers, when fed with data, to generate suitable hypotheses for explaining that data. Moreover this new kind of computer induction has resulted in the discovery of im-*

portant and previously unknown scientific laws." ([20])

Not only do machine learning algorithms allow for evaluating the degree of confirmation for a given hypothesis but also for the generation of such a hypothesis. This even led to the discovery of new scientific laws based on which we can make future predictions. Such successes strongly undermine Popper's project for eliminating induction from science. Gillies therefore concludes that Popper's view of induction being a myth can no longer be maintained and that inductive inferences based on many observations have become an important part of scientific procedure [20].

Gillies' claims have prompted a fierce discussion on the extent to which the successes of machine learning procedures undermine anti-inductivist views [4, 30, 21, 34, 3, 33, 38, 14, 22]. This debate strongly resonates with earlier disputes between logical positivists and their opponents. Allen [2] even claims that the recent ubiquity of data based machine learning procedures forms a new uprising of ideas originating from logical positivists. He talks of a new dark age of positivism. Subsequently, he argues Gillies' claims to be mistaken since it has already been conclusively shown that the ideas of logical positivism are erroneous. Allen thus implies that the proclaimed misunderstandings of logical positivism also apply to the data based machine learning procedures Gillies considers. At the same time, he does not present an explication of how precisely machine learning procedures are supposed to fall victim to the pitfalls of ideas stemming from logical positivism.

In light of Gillies' claims and the growing presence of machine learning procedures in many fields of science, it becomes very interesting to actually present such an explication. This could result in new insights about the role machine learning might fulfill in scientific endeavours. Thereby, if the problems of the inductive methods presented by logical positivists indeed still apply to machine learning procedures, this would seriously challenge the claim that the inductive inferences central to such procedures could lead to the discovery of new scientific laws. In this thesis I will therefore examine the relevance of criticism voiced against the inductive logic presented by Carnap for modern day machine learning procedures. I will consider three well-known critiques that strongly undermined Carnap's inductive methods. For each critique I will examine the relevance for machine learning by first formulating how the presented problem might apply to its procedures, after which I will inspect

how these procedures are supposed to account for this problem.

The decision to let Carnap's inductive logic represent the inductive methods presented by logical positivists is based on the strong resemblance this work shows to machine learning procedures. In many ways Carnap's inductive logic forms a bases for modern day artificial intelligence [35]. And since it has been shown that this work might lead to irrational reasoning, it becomes even more urgent to examine if the same might hold for machine learning procedures. Furthermore, because of its many varieties, it is impossible to formulate a general claim about all machine learning procedures. Such a claim would always be prone to counterexamples. In this work I will therefore only consider linear models for supervised classification.

I will first present an elaborate introduction to both supervised classification and Carnap's inductive logic. The reader familiar with the workings of linear models for supervised classification is invited to skip over section 2 in which such models are considered. Section 3 will introduce Carnap's inductive logic, starting with a short introduction of his early work, followed by an extensive overview of his last published work on this topic. Sections 4 - 6 each set forth one of the three critiques I consider in this thesis. In section 4 I will show how considerations presented by Nelson Goodman pose a challenge for the way we decide to represent our data in a classification problem. In section 5 it will become clear that a famous paradox that presented a serious challenge for Carnap's inductive logic shows us that we have to introduce some strong assumptions to be able to apply supervised classification methods. Finally, section 6 is centered around another famous problem presented by Goodman, namely the grue problem.

2 Machine Learning

In this section I will present an introduction to supervised classification methods. I will only consider simple linear models since these models greatly simplify considerations presented in later chapters without any loss of the significance of these considerations. Although there is a Bayesian alternative for every method presented in this section, I will only consider methods based on a frequentist understanding of probability. Finally, I will mainly base myself on the excellent handbook by Bishop [6], the extensive exposition of classification procedures published by Jain, Duin and Mao [31] and finally the thorough introduction of statistical learning theory presented by Luxburg and Schölkopf.

2.1 Supervised Classification

Supervised classification is one of the most well-studied problems in machine learning. Its general aim is to find automated ways of classifying certain objects into a finite set of classes. In supervised classification we deal with two kinds of spaces: the input space X which represents the objects to be classified and the output space Y which represents the classes an object can be assigned to. The objects to be classified are represented by a set of training variables while the different classes form the target variable. Our task is to find a functional relationship of the form $f : X \to Y$, that is a function that predicts the value of the target variable for a certain object based on its values on the training variables. [43]. To be able to find such a function we are presented with a dataset comprising N observations $\{\mathbf{x}_n\}$ of the training variables, where $n = 1, \ldots, N$, together with the corresponding values on the target variable $\{y_n\}$ [6]. A classification algorithm takes such a dataset as input and outputs a classifier f.

The dataset used for finding a classifier is assumed to be generated from a joint probability distribution P on $X \times Y$. We further assume all data points $(\mathbf{x}_i, y_i)$ to be sampled independently from P. We usually consider P to be fixed over time. This means that we assume that the distribution that generated the data we have observed up till now will be the same distribution underlying later observations. Obviously, at the time of learning P is unknown. If P were known, then learning

would be trivial since we could easily formulate expectations of the target variable based on the observed values of the training variables. However, in a supervised classification problem we only have indirect access to P through the observed data points. Based on this limited information we try to find a classifier that given some new value $\hat{x}$ would accurately predict the corresponding target value $\hat{t}$.

To be able to find such a classifier we define a loss function. This function is used to measure the performance of a certain classifier f. The loss function tells us the "cost" of classifying $\mathbf{x} \in X$ as $y \in Y$ [43]. The simplest loss function for classification is the misclassification error which assigns a loss of 0 to a correct classification and a loss of 1 to an incorrect classification. Such a loss function only measures the error of a classifier on some individual data point. We therefore also define a risk function which measures the average loss of a classifier over data points generated according to the underlying distribution P [43]. Simply put, the risk determines how many points within the input space would be misclassified when using classifier f. Consequently, our goal in supervised classification is to find the classifier with the smallest risk value. However, since P is unknown it is impossible to determine the true risk value of a classifier. We therefore apply the expected loss on the available dataset, known as the empirical risk, as an estimate of this true value. Based on this estimate we compare different classifiers.

Finally, we have to decide what kind of classifiers to consider. Before we start learning we have to define a hypothesis space F containing all candidate classifiers. Such a space is partly determined by the learning algorithm we use and the way the input space X is defined. An optimal classifier would assign an observation $\mathbf{x}$ to the category y for which $P(Y = y|X = \mathbf{x})$ is highest, since this would most probably result in the lowest number of misclassifications. However, this would again require knowing P. Therefore, given all of the above definitions, the standard problem of supervised classification can be formulated as follows: given data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ which have been drawn independently from some unknown probability distribution P, and given some loss function, how can we construct a function $f : X \to Y$ which has an optimal risk value. Next we will see how we would approach this problem.

2.2 Data Representation

Before learning can start, we first have to define the input space X and the output space Y. This means that we have to decide on a suitable representation of the target variable we would like to predict and select a set of training variables based on which we will make these predictions. The representation that we will use strongly influences the learning processes of the classification algorithm. First of all, it affects the number of data points needed to get a robust estimation of the optimal classifier. Second, our decision may greatly influence the computational costs of the learning process. And last but certainly not least, the representation we will use partly determines which classifier will be judged most optimal. It follows that great care should be taken when making this decision.

In a supervised classification problem the target values correspond to one of K discrete classes $\mathcal{C}_k$, where $k = 1, \ldots, K$. Most commonly the classes are taken to be disjoint, so that each observation is only assigned to one class. Not only do we have to decide on the disjoint classes, but also on how to represent these classes. There are various ways to let the target value we like to predict represent the class labels. For a two class problem where we assign a data point to one of two classes, the most convenient is often a binary representation. Using such a representation there is a single target variable $y \in \{0, 1\}$, where $y = 1$ represents class $\mathcal{C}_1$ and $y = 0$ represents $\mathcal{C}_0$ [6]. When using probabilistic models, the values of y can be interpreted as the probability that the considered $\mathbf{x}$ belongs to class $\mathcal{C}_1$. If we are considering more than two classes , $K > 2$, often a 1-of-K representation is used. In this coding scheme $\mathbf{y}$ is a vector of length K such that if the class is $\mathcal{C}_j$, then all elements y_k of $\mathbf{y}$ are zero except element y_j, which takes the value 1. For example, when considering four classes, an observation belonging to the second class would be represented as

$$\mathbf{y} = (0, 1, 0, 0)^{\mathrm{T}}.$$

We could again interpret the values y_k of $\mathbf{y}$ as the probability that the class is $\mathcal{C}_k$. These are just two types of representation, mainly convenient for probabilistic methods of classification. For different methods, different kinds of representation might be more convenient.

Next we have to decide on a set of training variables and how to represent these. Therefore, before learning we often preprocess the original input variable(s) [6]. This preprocesing aims at transforming the original input variables in such a way that will hopefully make it easier to learn from the data. The preprocessing stage is often called feature selection. Features are numeric representations of the objects of observation. With feature selection we try to represent the original input variables by only those features that we suspect to be good predictors of the target value. This selection could be based on prior knowledge of the research topic or could be made by algorithms designed for this purpose. Given a set of features every observation is represented as a feature vector $\mathbf{x} = (x_1, \ldots, x_D)$, where D is the number of features selected. The elements of such a feature vector are the values of each feature for a specific observation.

The features we decide to use all corresponds to a dimension of the input space X, which will be called feature space from now on. Within this space every observation is represented by the point that corresponds to its feature values. Figure 1 shows a simple two dimensional feature space. In this example we represent our observations by using two features. The class to which a data point belongs is indicated by its color. This is a two-class problem where we could for example be predicting if someone has cancer based on the outcomes of two medical tests.

Representing the dataset within a feature space allows for a geometrical interpretation of classifiers. By applying a certain classification algorithm we try to find decision boundaries that partition the feature space in such a way that all points within a decision region $\mathcal{R}_k$ resulting from this partition belong to the same class $\mathcal{C}_k$.

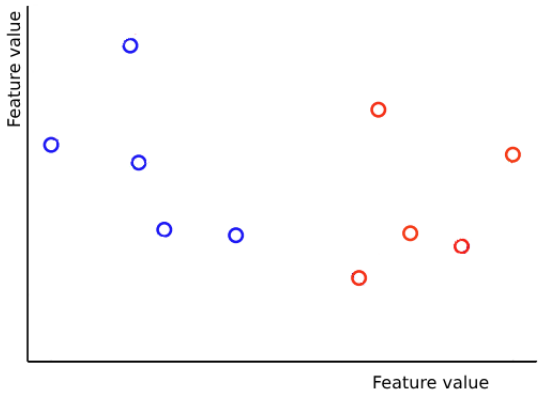

Figure 1: Feature Space (figure taken and modified from [6])

Based on such decision boundaries, an unobserved data point $\hat{x}$ is assigned to a certain class. Most commonly, decision boundaries are defined by a set of parameters denoted as $\mathbf{w}$ [6]. The hypothesis space F thus consists out of all possible instantiations of model parameters $\mathbf{w}$. The task of our learning algorithm is to search through this space to find the value of $\mathbf{w}$ that is best confirmed by the available data. As already mentioned above, the shape of the space F is partly determined by the classification algorithm used. This translates to the shape of the decision boundaries being depend on our choice of algorithm. In this thesis I will mainly consider linear models, which means that the decision boundaries are linear functions of the input vector $\mathbf{x}$. Figure 2 shows the decision boundary found in feature space by such a linear model. More elaborate classification methods like feed-forward neural networks can find highly non-linear decision boundaries as depicted in figure 3. However, it can be shown that such elaborate methods apply the same elementary principles as the classification methods considered here.

2.3 Training

Given an appropriate representation of the dataset, the training phase of the supervised learning algorithm can begin. In the training phase of our classification algorithm we try to find the decision boundaries that are best supported by the available data. However, keep in mind that our goal is to be able to classify unobserved data points which are likely to be different from the the data points in the

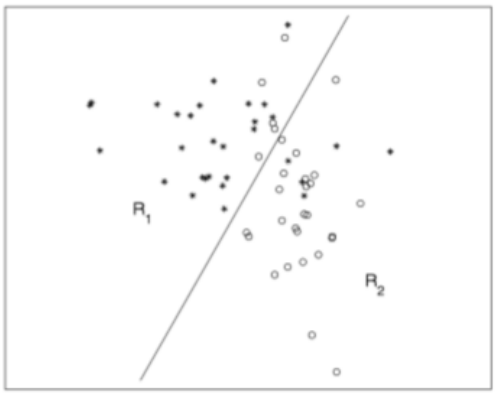

Figure 2: Linear classification model [31]

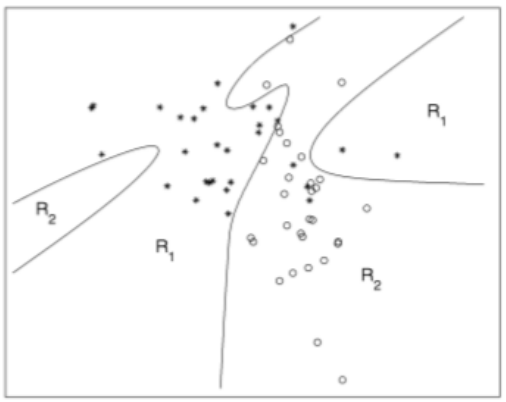

Figure 3: Non-linear classification model [31]

available dataset [31]. To ensure the model we learn will also generalize well for unseen data we split the available dataset in a training and a test set. Initially we use the training set to determine the best supported model parameters $\mathbf{w}$. Subsequently the test set is used to test the predictive capabilities of the best supported model on unseen data. Such a split should prevent overfitting, a phenomena whereby we only maximize the performance of our classifier on the training set resulting in inaccurate predictions for unseen data [6].

In this thesis I will consider two different approaches to finding the best decision boundaries during training; the probabilistic approach and the geometric approach [31]. Remember that the best classifier would assign a data point $\mathbf{x}$ to the class y for which $P(\mathcal{C}_k|\mathbf{x})$ is highest. Since P is unknown we can only acquire an estimate of these probabilities from the training data. The probabilistic approach does this by first determining the class-conditional probabilities $p(\mathbf{x}|\mathcal{C}_k)$ and the prior probabilities $p(\mathcal{C}_k)$ based on the training set, after which the Bayes' theorem is applied to find the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ for every class [6]. The posterior probabilities are then used to decide on the appropriate class. On the other hand, the geometric approach does not make use of probabilities at all. With this approach the decision boundaries are directly learned from the data points as represented within the feature space.

The probabilistic approach can be further subdivided into generative and discriminative models. Generative models first determine the class-conditional and prior probabilities for every class, after which the Bayes' theorem is applied to find the posterior class probabilities. To form an estimate of the class conditional probabilities we first have to make some assumptions about the probability densities for each class. We for example commonly assume Gaussian densities for continuous features. The assumptions we formulate depend on the dataset available and the prior knowledge of the investigator. Given such assumptions we express the class-conditional densities in a parametric form, after which we commonly apply a maximum likelihood procedure which determines the parameter values based on the training set. For the prior probabilities for every class it turns out that the maximum likelihood solution coincides with the ratio of the data points in the training set belonging to a class $\mathcal{C}$ and the total amount of data points. Finally, with Bayes'

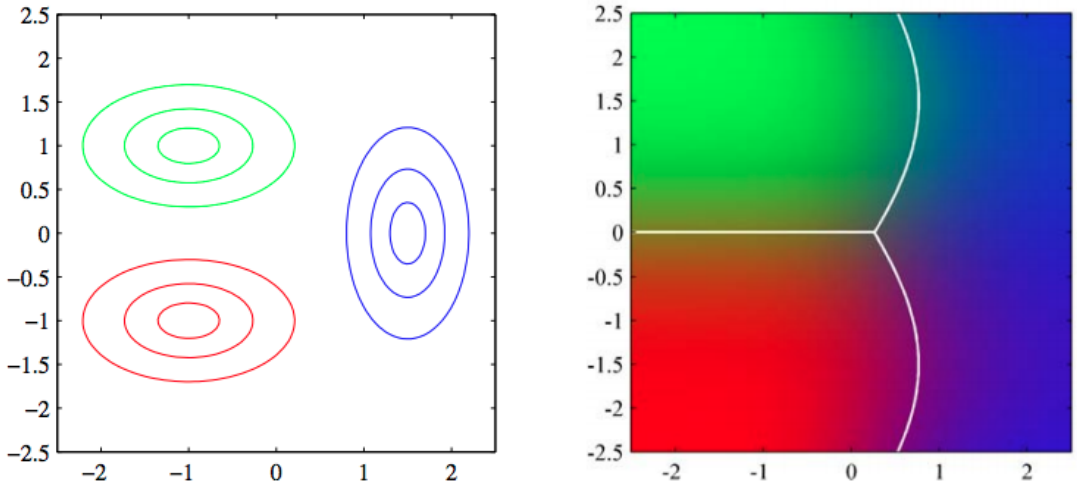

Figure 4: The left-hand plot shows the class-conditional densities learned from the training set for a three-class classification problem. We have assumed the classes to have a Gaussian distribution with class 1 (green) and class 2 (red) having the same covariance matrix in contrast to class 3 (blue). Based on these densities we find the posterior probabilities, as depicted by the colors in the right-hand plot. Note how the assumption of equal covariance matrices results in linear decision boundaries while different matrices result in quadratic decision boundaries (figure taken from [6]).

theorem we are able to find the posterior class densities for every class. Given the posterior probabilities we assign every point in feature space to the class which results in the lowest expected loss. This classification also establishes the decision boundaries (see figure 4). In this thesis I will not go into most of the technicalities of the statistical procedures considered, the reader is referred to the excellent overview presented by Bishop [6]

Discriminative models find posterior class probabilities by maximizing a likelihood function defined through the conditional distribution $p(\mathcal{C}_k|\mathbf{x})$. These models apply a direct approach, whereby we determine the posterior probabilities without first fitting class-conditional and prior probabilities [6]. A much applied example of such a model is the logistic regression classifier. This classifier maximizes the likelihood of the observed training set. This is done by an iterative method whereby we consider the likelihood of the training data given different sets of model parameters $\mathbf{w}$. By iteratively updating $\mathbf{w}$, optimal values can be found relatively easily. Based on this set of optimal model parameters $p(\mathcal{C}_k|\mathbf{x})$ can be determined for every $\mathbf{x}$.

In contrast to discriminative and generative models, which both apply a parametric method, there are also some much applied non-parametric density estimators. Such a method avoids the possibility of false assumptions about the actual densities that generated the data [6]. A well-known example of a non-parametric method is the k-nearest neighbor classifier [31]. I will not go into the details of this method. For now it suffices to mention that such a classifier assigns an input $\mathbf{x}$ based on its k nearest neighbors in feature space. A sphere is drawn around $\mathbf{x}$ until it includes k data points, after which $\mathbf{x}$ is assigned to the majority class within this sphere.

Before considering geometric models it is important to mention that classification by the parametric probabilistic models presented above can be subdivided in an inference stage and a decision stage [6]. In the inference stage we use the training data to learn the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$. Subsequently, in the decision stage we assign each $\mathbf{x}$ to a certain class based on these probabilities. The class to which a new $\mathbf{x}$ will be assigned depends on the applied loss function. Loss function are defined as seems appropriate for the aim of our investigation. If we simply aim for minimizing the chance of assigning $\mathbf{x}$ to a wrong class, we would apply the simple misclassification error as discussed above. This would result in assigning every $\mathbf{x}$ to the class $\mathcal{C}_k$ for which $p(\mathcal{C}_k|\mathbf{x})$ is highest.

However, often certain misclassifications might have severe consequences. Imagine the situation in which a cancer patient is classified as not having cancer. We would want to minimize the probability of making such an incorrect classification. By defining different loss values to different errors we can incorporate these considerations in the decision stage. We could for example assign much higher loss values to classifying a sick person as healthy than the other way around. By minimizing the expected loss function we would thus also minimize the number of such errors. Lastly, we could also decide not to classify those points for which $p(\mathcal{C}_k|\mathbf{x})$ is below some threshold, meaning our uncertainty about the true class is high. Neglecting such hard decisions might avoid making mistakes.

Finally, geometric models construct decision boundaries by optimizing the expected loss value directly from the feature space[31]. Such methods strongly depend on the loss function used. A well-known loss function is the mean squared error (MSE) which measures the squared 'distance' between the classifier output and the

true class labels as presented in the training data [43, 6]. By minimizing the MSE the optimal decision boundaries are found directly in feature space. A classical example of a geometric model is the perceptron classifier. This linear two-class method minimizes an alternative loss function known as the perceptron criterion. The perceptron learning algorithm iteratively adjusts the decision boundary by cycling through all training data points in turn. If it finds an incorrectly classified data point, the decision boundary is adjusted appropriately (see figure 5) [6]. This process is repeated until all data points are correctly classified. Although geometric models greatly simplify the learning process by mapping data point directly onto a certain class, there are many good reasons for wanting to compute the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$. Thereby, these models strongly depend on the choice of loss function used for learning.

2.4 Testing

Having found the model that is best supported by the training data, we test its predictive performance for unseen data with the test set. Again we could use a simple loss function and measure the performance of a classifier by simply determining the classification error on the test set. The number of misclassified data points serves as an estimate of the performance of our classifier on future observations. Based on this error rate we could also compare different models. We could for example train multiple models on the training set, after which we select the best model based on the misclassification error on the test set. However, often such a single number is not adequate to characterize the performance of a classifier [31]. As already mentioned above, different classification errors might have different consequences. Therefore, it is insightful to consider the kind of errors a model makes. Depending on the classification problem at hand it might for example be preferable to also evaluate a model on the number of false positives or false negatives it makes. Thereby, various information criteria have been proposed that attempt to capture the performance of competing models, such as the Akaike information criterion (AIC) [6]. This performance measure also compares models on their complexity by penalizing for the number of adjustable parameters of a model. Many more performance estimators have been proposed and since there doesn't seem to be one superior measure it is

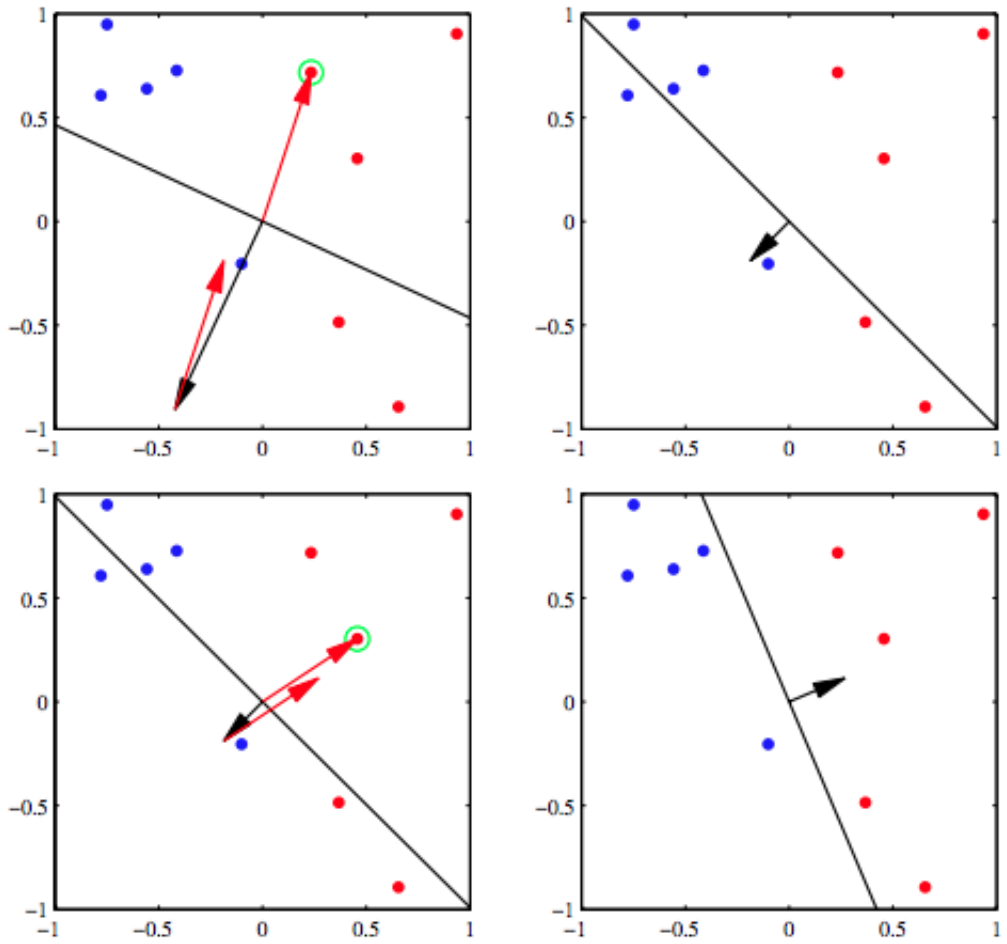

Figure 5: Convergence of the perceptron algorithm for a two-class classification problem. The upper left plot shows the initial model parameters $\boldsymbol{w}$ *(black arrow) and the corresponding decision boundary orthogonal to* $\boldsymbol{w}$*. The parameter vector points towards the 'red' decision region. An incorrectly classified data point is identified (green circle), after which its feature vector (red arrow) is added to* $\boldsymbol{w}$*. The upper right plot shows the new established decision boundary. The lower left plot shows the repetition of this process. Again adjusting the decision boundary stops the process since all data points are correctly classified, as shown in the lower right plot (figure taken from [6]).*

recommended to always consider multiple when deciding on the best model.

Lastly it should be mentioned that the reliability of performance measures strongly depends on the training and test set used. Often the available data for training and testing is limited. At the same time the training set should be large enough to ensure a high generalization ability of our model, the test set should be large enough to ensure reliability of the performance measures and both should be independent

[31]. Different methods have been proposed to ensure the reliability of performance measures when data is limited. Such methods divide the data into a certain number of partitions, after which different runs use one of these partitions as the test set and the remaining partitions as the training set. The resulting test results are then averaged to give an accurate description for the model performance.

Considering the above we see that supervised classification problems can be considered to be search problems where we try to find a set of model parameters within a hypothesis space. The model parameters that we judge to be best supported by the data define decision boundaries that divide the feature space into different classes. It follows that the model that is judged to be optimal strongly depends on the way we decide to represent the data. Such dependence strongly resembles elements of Carnap's inductive logic that have been heavily criticized, which I will consider now.

3 Carnap's Inductive Logic

Supervised classification methods present us with a way to reason from past observations to future expectations. Such reasoning, called inductive reasoning, has played a central role in the history of the philosophy of science. The term induction goes back to Aristotle whose conception of the manner involved moving only from particulars to universals [17]. With the introduction of precise mathematical accounts of the notion of probability and the description of more sophisticated inductive techniques this rather narrow way of thinking about inductive reasoning became outdated. Building further on the groundwork laid by these developments, the twentieth century has shown various attempts at forming a much more general framework for inductive reasoning. One of the most famous of such attempts has been presented by Rudolf Carnap (1891 - 1970). The inductive logic Carnap worked on aimed at characterizing a quantitative and logical relation which would express how much our past observations confirm possible future observations. The work he presented on this topic shows a strong resemblance to supervised classification. It has even been claimed that Carnap wrote the first artificial intelligence program [23].

Throughout the course of his life Carnap presented multiple systems of inductive logic. Here I will divide his work into two parts, based on the way observations are represented. Carnap's early work mainly corresponds to his "Logical Foundations of Probability" [9] and "The Continuum of Inductive Methods" [11]. In his early work Carnap represented observations as sentences of a formalized linguistic framework. Such frameworks play a central role in Carnap's inductive logic and much of his work can be considered as a further elaboration of this notion [41]. Over the years it became clear that the representation used in Carnap's early work was to restrictive to underlie his inductive reasoning. Therefore, in his last work, "The Basic System of Inductive Logic" [7, 8], he introduced a more elaborate way of representing observations. However, Carnap never fully succeeded in overcoming the problems already voiced against his earlier work.

In this section, I will first clarify how supervised classification methods and the inductive logic of Carnap formalize the same kind of reasoning. This direct correspondence can best be seen when considering Carnap's early work. I will therefore

present a quick overview of the early Carnap in which I will mainly highlight the similarities to supervised classification. Subsequently, I will present an extensive overview of the Basic System, which will form the starting point of my investigations in subsequent chapters. It will become clear that the geometrical concepts Carnap introduces in this work form predecessors to the feature spaces as applied in modern day supervised classification. However, since these geometrical concepts never really solved the problems of Carnap's inductive logic, it also becomes highly interesting to examine the susceptibility of supervised classification methods to these problems.

3.1 Early Carnap

Just as supervised classification methods, Carnap's work on inductive logic presents a system which guides us in forming beliefs regarding future observations based on some observational data we have already gathered. Such a system of inductive logic was supposed to serve as a basis for scientific investigations. Carnap considered inductive logic to be the study of confirmation [40]; we are interested in finding a way to express the degree to which some evidence E confirms hypothesis H. He equated confirmation to probability meaning that confirmation was captured by assigning conditional probability values to pairs of hypothesis and evidence [19]. In his early work, probability was considered to be a logical concept. The conditional probability ascribed to a hypothesis based on some evidence should be understood as a quantitative generalization of a logical entailment relation between both [17]. Degree of confirmation is thus seen as an objective relation between evidence and hypothesis. It follows that Carnap had to present a way in which a set of observations logically entails, or confirms, possible future observations.

The works of early Carnap present a range of functions which quantitatively determine how much a set of observations confirms future observations. Carnap claimed that these functions were able to capture different forms of inductive inference. One of these was the inference based on analogy, whereby we base future predictions on similarities between observations. It is in this kind of inductive reasoning that the strong resemblance between Carnap's inductive logic and supervised classification can be seen most clearly. To clarify this form of inductive inference

consider the following example. We are studying a rare kind of ladybug. Our first observation consists of a ladybug which is red and has six dots. Next we observe another ladybug but are only able to spot its color, namely red. Subsequently, we wonder what these observations tell us about the probability that the second ladybug also has six dots. In his early work Carnap aimed at presenting a system of inductive logic that could account for this kind of reasoning.

Carnap states that based on the observation that both ladybugs are red we should also have a higher expectation of the second ladybug having six dots [40]. Thus, the inductive effect an observation can have on future observations should be based on similarities between these observations. It follows that we need an account of what it means for two objects to be similar. Early Carnap determines the similarity between observations by tracking their number of shared properties. The two ladybugs in our toy investigation are judged to be similar because they share the property "red". To be able to make such similarity judgments we need a set of properties that will be used to describe our observations. Therefore, before we start making observations we have to decide on such a set. First, we need to determine which properties of the objects of our observations we deem to be inductively salient for the purpose of our investigation. Do we only observe the color and the number of dots of a ladybug, or should we also record its weight and size. Second, we need to decide on the predicates we will use to represent these properties. When we for example decide to record the color of our observations, do we then use a range of different colors or is a white/non-white distinction enough. Just as in supervised learning, before we can formulate predictions based on the evidence we have gathered, we should specify which properties of our observations we consider to be relevant for such predictions.

Given a set of properties that will determine the form of our observations, similarities are traced by applying Q-predicates. Q-predicates are full descriptions of any object by a conjunction of predicates, one for every property we have judged to be relevant for our investigation [40]. In our ladybug example, the Q-predicate for our first observation would consist out of a conjunction of the predicates "red" and "six dots". Note how these Q-predicates are the Carnapian version of modern day feature vectors as used in supervised classification. Given these Q-predicates, the principle of analogy Carnap wanted his inductive logic to account for states that

the probability that some object b has a certain property p, is increased by the information that one or more other objects, which share some other properties with b, have property p [12].

Unfortunately, Carnap has never succeeded in presenting a system of inductive logic that he judged to be capable of accounting for this kind of analogical effect. At the same time, the analogical reasoning Carnap wanted to account for appears to be at the roots of modern day supervised classification. To clarify this, again consider the ladybug investigation. Since the second ladybug we observed was red, we should also have a higher expectation of it having six dots. Being red thus raises the probability of having six dots. Such a statement is actually equivalent to the assumption that being red and having six dots are statistical dependent properties. Striving to account for the kind of analogical reasoning presented above can therefore also be seen as wanting to allow for statistical dependencies between properties of our observation. It is exactly this kind of reasoning that stands at the roots of supervised classification. Supervised classification methods indeed start from the assumption that there is such a statistical dependence between the properties of our observations. Since the exact form of this dependence is unknown, we apply earlier observations to acquire an estimate of this form. Subsequently, when considering to which class a new observation belongs, we assign the highest probability to the class to which earlier observations with similar feature values belonged. The underlying reasoning seems to be that if our new observation is similar in respect to a certain set of features to earlier observations, then it will most probably also be similar on another feature. Again, this is exactly the kind of analogical reasoning Carnap wanted to formalize in his early work.

Carnap never considered his early work to be successful since it was shown that it could lead to irrational reasoning. This was partly due to the strong linguistic focus of this work. This linguistic focus can, among other things, be seen in the way observations are represented in Carnap's early work. Carnap considered our observations to be part of a linguistic framework. These frameworks can best be described as language systems adopted in order to be used as a basis for scientific investigations [41]. Our observations were represented as sentences within such a linguistic framework. Consequently, confirmation was ascribed to such sentences.

However, Carnap came to the realization that the frameworks he applied in his early work were too restrictive. This can again best be seen when reconsidering the ladybug example. This time consider a slight variation of our ladybug example, as originally presented by Sznajder[40]. We are now not only interested in a ladybug's color and number of dots but also its size. The first ladybug we observe is big, red and has seven dots. The second is big, orange and has six dots. The third is big, blue and has fifteen dots. The second and third ladybug both only agree on their size to the first ladybug. Our first observation should therefore equally confirm both other observations. However, intuitively we would judge the first and second ladybug to be more similar since orange is more similar to red than blue and seven is closer to six than fifteen. We would therefore expect the first observation to more strongly confirm the second observation. However, the linguistic frameworks of Carnap's early work offered no way to capture such considerations. This resulted in our inductive reasoning being too strongly dependent on the way we define the linguistic frameworks we use for our investigation. This made Carnap reconsider the way observations should be represented. His later work presented a more elaborate way of such representation.

3.2 Late Carnap

In his two-part "A Basic System of Inductive Logic" (the Basic System from here on) [7, 8] Carnap continues the project of developing an inductive logic. Just as in his earlier work, his aim was to formalize the inductive reasoning needed for scientific investigations. Unfortunately, Carnap died while working on the second part of the Basic System and has never been able to finish his work. The Basic System was supposed to serve as the foundation for an inductive logic that could account for the problems voiced against Carnap's earlier work. With this goal in mind Carnap introduced a new way of representing our observations. The geometrical concepts of representation he presents should serve as a firm base for grounding the way in which different observations might confirm each other.

3.2.1 Confirmation Functions

In the Basic System Carnap moved away from the strong linguistic focus so characteristic of his earlier work [40]. This resulted in the linguistic frameworks of his earlier work becoming conceptual frameworks. In contrast to linguistic frameworks, conceptual frameworks are no longer just a formalized language. Carnap comes closest to defining conceptual frameworks by stating that such a framework is a universe of objects and a system of descriptive concepts that characterize the object [7]. These descriptive concepts, which we use to describe our observations, might be labeled with a language but do not necessarily have to belong to one. This makes our inductive reasoning less dependent on the language we define to describe our observations. However, conceptual frameworks do still serve the same purpose as linguistic frameworks, namely to model the process of forming beliefs regarding future observations based on some observational data we have already gathered. We are still interested in finding a way to express the degree to which some evidence E confirms a hypothesis H. In the Basic System Carnap presents a new set of functions to capture such confirmation, symbolically represented as $\mathscr{C}(H|E)$. And since we already noted that Carnap equated confirmation to probability [19], the confirmation functions assign real valued conditional probabilities to pairs of propositions; what is the probability of H given we have observed E.

In the Basic System Carnap also moved away from the logical concept of probability in which degree of confirmation is seen as an objective relation between propositions [40]. The Basic System introduced a more subjective concept of probability. Confirmation functions now represent credence functions that are rational to have before any evidence is accumulated [41]. While the initial credence functions are considered to be 'a trait of the underlying permanent intellectual character' of an agent, the further credences characterize the momentary state of an agent at a specific time with respect to his beliefs [7]. It thus seems to follow that the confirmation values are completely dependent on the agent to which they are ascribed. However, Carnap does not adhere to such a pure subjectivist interpretation of probability by accepting only those confirmation functions that are rational. Although many confirmation functions are possible, only those that are rational to have should be part of our system of inductive logic. Therefore, Carnap's aim is to find methods of de-

termining the "correct" or "acceptable" confirmation values of various propositions based on the conceptual structure of our investigation [29]. In the Basic System he examines multiple axioms and rationality constraints which could be added to his system of inductive logic as a way to restrict the range of possible confirmation functions to only those that are rational to have [41].

Confirmation functions are no longer applied to sentences, but to propositions [7]. Such a proposition is of the form individual a has property P [7]. Again, before observations can be made we should decide on the set of properties, which will be called attributes from now on, that will be used to describe our observations and in what form these observations will be recorded. Carnap only discusses monadic predicates as attributes that can be ascribed to individuals. Thereby, it is assumed that these predicates can only be observable properties of the observable individuals [7].

The attributes we will use for making observations are all classified into families. Each family consists of those attributes that belong to the same general kind, here called modality. An example would be the modality "color" to which the family of predicates "color names" would belong. Carnap does not present a formal definition of modalities. They can be best thought of as those respects in which objects can be judged similar or different [41]. Modalities can be both quantitative and qualitative. The Basic System only considers reasoning with predicates belonging to a single family. So, instead of reasoning from the color of a ladybug to its amount of dots, we are now only concerned with how a color observation confirms another color observation. We could for example ask ourselves how much observing a particular red ladybug confirms, or changes the probability, of next observing a yellow ladybug. To model the meaning of the predicates we selected for our investigation Carnap introduced a geometrical concept, namely attribute spaces. To each family of predicates belongs a specific attribute space which formalizes how making observations might confirm future observations.

3.2.2 Attribute Spaces

An attribute space is "an abstract, logical space whose points represent the elementary properties of the modality in question" [7]. If we would consider the modality

color, every point within the space represents a specific shade. However, since most investigations do not consider all shades of color, the attributes forming the family that is represented by the attribute space are often less specific than the particular points within this space. The attributes are therefore represented by a region within the attribute space. For example, the attribute "red" would be represented by a region while all points within this region correspond to a specific shade of red. Finally, all predicate families are disjoint and exhaustive, which means that every object has at least one attribute of a specific family and cannot have two attributes of such a family [40]. Consequently, all regions together form a partition of the entire space.

Carnap does not provide a precise recipe for choosing the shape of an attribute space [41]. He only suggests that the choice of attribute space should be informed by our scientific theories. Thereby, the number of dimensions of an attribute space seems to depend on the kind of modality the space corresponds to [40]. And since the classification of attributes into different modalities is not unique, an attribute space can be divided into dimensions in various ways [7]. Consider for example the space belonging to the family of the six colors red, orange, yellow, green, blue and violet. We could consider this space to be the three-dimensional space corresponding to the modality color, as is depicted in figure 6. In this case we would partition the space into six regions all belonging to a specific color from the family we consider. However, we could just as well split the attributes up into the three modalities hue, chroma and lightness and consider three one-dimensional spaces. Carnap seems to suggest that the shape of an attribute space should be chosen as seems fit to the scientific investigation at hand [40].

As we shall see, the structure of the attribute space determines how making a certain observation influences our expectation for future observations. Just as in his earlier work, Carnap bases this inductive effect on judgments of similarity between the attributes that make up our observations. Now if we observe a certain attribute, say "red", we would have the highest expectation of next observing the same attribute, another "red" thing, since the object most similar to our observation is the object of our observation itself. However, Carnap also states that our expectation of next observing a "yellowish-red thing" should increase because it is similar to our "red" observation. Thus, the inductive effect one observation can have on future

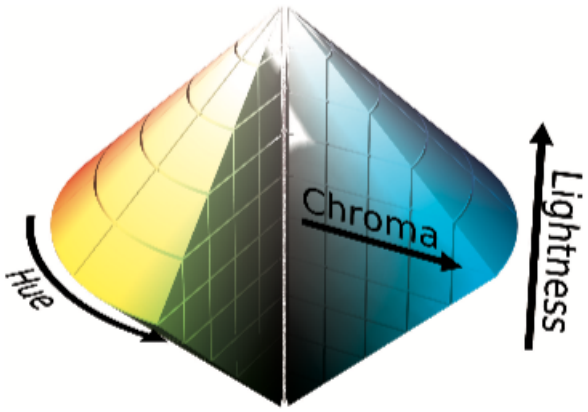

Figure 6: Color space (figure taken from [40])

observations is again based on the similarity between these observations. However, note that since the Basic System is only concerned with reasoning with one modality, we are now talking of another kind of similarity.

The degree of similarity between attributes is captured by applying a concept of distance within the attribute space [7]. This means that if X_1, X_2 and X_3 are regions within attribute space U, then the distance between X_1 and X_2 would be smaller than the distance between X_1 and X_3 if X_1 is judged to be more similar to X_2 than to X_3 [29]. Consider the attribute space in figure 7. This space could be seen as an one-dimensional subspace of the color space in figure 6. Since we would judge regions B_1 and B_2 to be more similar than B_1 and B_3, their regions are closer to each other. This also means that an observation of property B_1 would lead to a higher expectation of next observing B_2 than B_3. Sometimes it is possible to go beyond such comparative statements. In these cases we can define a quantitative distance function over the attribute space. Think for example of the attribute space belonging to the modality "length". The actual scale of length could serve as a distance measure over this space. The space could for example span over all lengths from 0 to 300 centimeters and be divided into 30 equal regions of 10 centimeter. The similarity between different regions could then be expressed as the normalized distance between their middle points. Notice how equating similarity with such distance measures invokes the strong assumption that similarities behave like distances, meaning they satisfy conditions like commutativity and the triangle condition [8].

Similarities between attributes are not always so clear as in the example above.

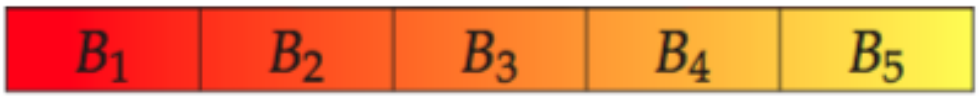

Figure 7: Attribute Space (figure taken and modified from [40])

Where then do the similarities, based on which we construct attribute spaces, come from? Carnap seems to consider similarity as similarity as perceived by a single agent [40]. However, he also discusses distance metrics based on judgments of groups of people. Here he talks of metrics introduced by psychologists who examined fields of our sensory qualities, like color and sounds, and came up with specific representations and difference measures. Thereby, similarity judgments could be based on background theory. The color space in figure 7 seems to be based on observational similarities between colors, just as the three-dimensional space in figure 6. However, a physicist, who will represent colors by the wavelength of a corresponding electromagnetic wave, would come up with different similarity judgments and thus a different attribute space. Such a shift would also mean a transition to attributes from a more reliable theoretical language instead of an observational language [7].

3.2.3 Similarity Influence

As mentioned above, to formalize how we can learn from observational evidence Carnap tries to invoke rationality constraints which restrict the set of possible confirmation functions to only those functions that are rational to have. In this line Carnap introduces a set of basic axioms for the confirmation functions [29]. However, there still exist many unacceptable functions consistent with the basic axioms. Consequently, additional rationality constraints are needed.

Some of these rationality constraints are introduced by the way we define an attribute space. It turns out that from the definition of our attribute space we can determine some values for the confirmation functions $\mathscr{C}$. Since we defined this function as a conditional probability function we should first assign prior probabilities to observing all different attributes. These prior probabilities can also be read of from the attribute space. Carnap introduces the confirmation function $\mathscr{M}$ as a

measure of the initial probabilities [7]. He suggests to relate the prior probability of an attribute to the width of its corresponding region within the attribute space [8]. The width is understood as the size (volume) of this region and could be measured using a width function which assigns normalized width values to regions within an attribute space. Such a function could be based on a distance measure defined over the attribute space. If there is no quantitative concept of width available (there is no quantitative distance function), attributes for which there is no reason to expect the occurrence of the one more than the other, should get assigned an equal sized region within the attribute space. If there is such a reason, the attribute with the higher expectancy should get assigned a bigger region. Finally, if a quantitative width function is available, prior probabilities should correspond to the value of this function. Returning to figure 7, we would assign each attribute $B1 - B5$ an equal prior probability of $\frac{1}{5}$ based on the equal widths of their regions.

To see how some $\mathscr{C}$-values can be determined from the attribute space, again consider the ladybug example. This time the investigator only observes predicates belonging to a single family, say color. Further suppose the investigator uses the attribute space depicted in figure 7. The similarity influence of observing a certain color, for example B_3, on future observations of a different color, $\{B_1, B_2, B_4, B_5\}$, is defined based on this space. Carnap emphasizes that this similarity influence only has secondary significance in determining confirmation values [8]. The confirmation values should primarily be determined by the number of observations of a specific attribute and the total number of observations. Thereby, if we have made a large number of observations, the similarity influence should only be small. However, the rate at which this influence is supposed to fade out is left as an open question [40].

To capture how observing a ladybug with color B_3 affects the probability of observing the other colors, Carnap introduces the η-parameter [8]. The η-parameters measure how much the knowledge that a P_h object was observed influences the belief that a P_j object will be observed [40]. This parameter is defined as the ratio of the probability of observing a P_j object given the evidence consisting of a single P_h observation and the initial probability of observing a P_j object. Given attribute indices j and h we have,

$$\eta_{jh} = \frac{\mathscr{C}(P_j a_2 \mid P_h a_1)}{\mathscr{M}(P_j a_2)} \tag{1}$$

and thus

$$\mathscr{C}(P_j a_2 \mid P_h a_1) = \eta_{jh} \mathscr{M}(P_j a_2) \tag{2}$$

with individual observations a_i. Because we do not yet know the exact form of the confirmation functions (C), the exact values of the η-parameters are unknown. However, since the η-parameters are used to capture the similarity influence between predicates and similarities are represented as distances within the attribute space, we could define some comparative restrictions on the η-values based on these distances. Carnap states similar tentative rules for the η-values as for the prior probabilities. If we have not defined a quantitative distance function over our attribute space and we have no reason for regarding attribute P_j as more similar or as less similar to P_m than to P_n, then we take $\eta_{jm} = \eta_{jn}$. If we are able to make a comparative judgment that P_j is more similar to P_m than to P_n without a quantitative distance function, then $\eta_{jm} > \eta_{jn}$. Finally, if a normalized distance function d is available, the η-values should be determined based on a function f such that $\eta_{jm} = f(d_{jm})$ [8]. The details of such a function are out of scope for this thesis. Here it is sufficient to observe that, in line with above similarity considerations, such a function should ascribe increasingly higher η-values to decreasing distances. This should result in a maximum η-value for $d = 0$, ensuring that observing a red ladybug would result in the highest probability of next observing another red ladybug.

Bear in mind that 2 only applies to situations where we have made a single observation. Using functions for distances and widths in our attribute space we are able to determine γ- and η-values for all considered predicates, based on which we can determine $\mathscr{C}$-values for our first and second observation [8]. However, given all γ- and η-values, values for the confirmation functions for later observations are not uniquely specified. Remember that the Basic System mainly consists out of initial considerations for the confirmation functions to be applied in a system of inductive logic. Carnap therefore only presents some tentative principles for considering bigger sets of observations. These principles could narrow the freedom of choice of confirmation functions for higher $\mathscr{C}$-values and should be considered as tentative

steps towards the final aim of a general rule for the determination of such values [8]. Important for this thesis is to consider that Carnap conjectures that the widths and distances belonging to the different regions within the attribute space should be sufficient as a basis for all $\mathscr{C}$-values. Such a conjecture puts heavy weight on the way we decide to define our attribute space. As we shall see in the next section, such reliance might lead to various problems.

4 Representation

In the previous section we saw that one function of attribute spaces was to represent the similarities between the predicates we consider in our investigation. Based on such similarities we could express how much making certain observations influences our expectation for future observations. By choosing to read similarities directly from the attribute space as distances, we put considerable significance on the way we define this space. Two different attribute spaces both representing the same set of observations might lead to different predictions. This can easily be seen when considering the attribute space for color observations. As we saw, color observations might be represented in both a one- or three-dimensional space. The distance between two colors might be different in both spaces, resulting in different predictions based on the same set of observations. This puts significant importance on the way we decide to represent our observations.

Similar dependence on the way we represent our data can be recognized in supervised classification methods. This can most clearly be seen when considering the set of features we use to represent our observations. The predictions we make for unseen data strongly depend on the feature space we use for making these predictions. This is most straightforward for geometric models. These models learn decision boundaries directly from the feature space and thus rely heavily on the definition of such spaces. If we choose to represent the same data in figure 5 by different features, the perceptron algorithm will obviously find a different optimal decision boundary. This might result in classifying the same observation in different classes, depending on the feature space used. The same point can be made for probabilistic models. Applying generative models with different features to represent the same data will most likely result in different class-conditional densities having the highest likelihood. The same holds for discriminative models, where the likelihoods also depend on the way we represent the data. Again the same data might lead to different predictions.

Of course we are immediately inclined to note that it is obvious that the result of our inductive endeavour strongly depends on the way we represent the available data. This does not only seem to apply to supervised classification but to statistical methods in general. It could therefore be argued that dependence on the way data

is represented is a broader problem of practising science and does not speak directly to supervised classification. At the same time, a well-known critique voiced by the philosopher Nelson Goodman (1906 - 1998) revealed how such dependence might lead to irrational reasoning in the Basic System. Goodman argued that it is not the possibility of different representations leading to different predictions that forms the problem, but rather our inability of pinpointing the relevant representation. In light of the strong correspondences of supervised classification methods to the Basic System it becomes highly relevant to consider if such an argument might also apply to classification methods. This would significantly challenge the strong claims made by Gillies about supervised classification methods. In this section I will therefore examine how the pitfalls of the Basic System revealed by Goodman might still be relevant for supervised classification methods.

4.1 Similarity

Similarities play an important role in Carnap's Basic System. Through similarities Carnap formalized the inductive effect one observation can have on others. In his "Seven Strictures on Similarity" Goodman dismissed this kind of reasoning by arguing that similarity is both a philosophically and scientifically flawed notion [25]. The seven strictures Goodman presents aim at showing how similarity is a slippery concept which should be treated with great care. The issues Goodman brings forth seem to strongly undermine the bases on which Carnap builds his system of inductive logic.

In the same period of time the Basic System was presented most theorizing about similarity was characterized by the geometrical model of similarity [15]. Carnap was one of the proponents of such an understanding of similarity, as can be concluded from the way we reason with similarities in the Basic System. The geometrical model states that similarities can be represented by means of a metric space. Given a distance function over such a space we can measure the degree of similarity between different objects represented in the space. This gives us a straightforward way of modelling comparative similarity judgments, but it also makes the strong assumption that similarity behaves like a distance function.

Although not explicitly mentioned by Goodman, the geometrical model clearly

seems to be implicit in his critique [15]. The considerations Goodman brings forward purport to show that this model is seriously flawed. These considerations are mainly supposed to prove the high context-sensitivity of similarity judgments:

> "*[C]omparative judgments of similarity often require not merely selection of relevant properties but a weighting of their relative importance, and variation in both relevance and importance can be rapid and enormous. Consider baggage at an airport check-in station. The spectator may notice shape, size, color, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. Which pieces of baggage are more alike than others depends not only upon what properties they share, but upon who makes the comparison, and when*" ([25])

This example addresses similarity as captured by the possession of common characteristics. However, the same context-sensitivity holds for late Carnapian similarities:

> "*[S]uppose we have three glasses, the first two filled with colorless liquid, the third with a bright red liquid. I might be likely to say the first two are more like each other than either is like the third. But it happens that the first glass is filled with water and the third with water colored by a drop of vegetable dye, while the second is filled with hydrochloric acid–and I am thirsty. Circumstances alter similarities*" ([25])

Both examples purport to show that there are many different ways in which two objects can be judged similar. Whichever property of our observations we deem to be relevant for making similarity judgments strongly depends on the context in which these judgments are made. Goodman comes to the conclusion that there cannot be one appropriate way to judge the similarity between two objects. For this reason, similarities as captured by the possession of common characteristics cannot account for our inductive reasoning. After all, our predictions given the same set of observations would then be different in a different context and there is no way to justifiably distinguish between these predictions.

Goodman's claims significantly undermine the use of attribute spaces in two ways. First of all, similarity judgments between the objects of our observations

are captured as distances within attribute spaces. Given that our similarity judgments are sensitive to the context in which these judgments are made, it follows that these distances will also depend on context. A different context would thus result in a different attribute space. Consequently, in a different context, the same set of observations will lead to different future predictions. Carnap counters these considerations by suggesting to use inter-subjective similarity metrics that resulted from psychological research. Such metrics have for example been introduced for color and for the pitch of musical notes and are supposed to be insensitive to context [8]. Goodman dismisses this kind of similarity metric on the ground that it creates rather than reflects a measure of sensory similarity [25]. The validity of such metrics could never be tested since there is no satisfactory way to obtain judgments of sensory similarity that are qualified to stand as criteria for appraising the result that follows from the intended psychological research. The same holds for using the already existing scale of the relevant modality as a similarity metric. In the Basic System we would for example use the difference in length as a way to capture similarity when considering the lengths of our observations. Again, we could ask ourselves why the ordering imposed by such a scale should be preferred over others. After all, there might be many justifiable ways in which we could judge similarities between persons of different length.

Secondly, and even more pressing for the Basic System, Goodman challenges Carnap's claim that the modality represented by an attribute space represents the inductively salient respect in which our future observations will be similar to the past. When we define an attribute space for our investigation we assume that future observations will be similar to past observations by having similar values on the modality represented by the space. Imagine for example that we have observed five ladybugs. We might wonder how these ladybugs might tell us anything about possible later ladybug observations. Should we expect future ladybugs to have the same number of dots, or will they be similar in their weight. We have to pick one of such properties in order to form predictions about future observations. Apart from mentioning that we should be informed by our scientific theories, Carnap does not offer much help in making this decision. At the same time, Goodman's claims are supposed to show there is actually no justified ground on which we could pinpoint one

property as the appropriate one. Similarities between future and past observations may come in many forms and it is unclear how to determine along which, among countless lines of similarity, time will run. And since Carnap states that it is up to the investigator to determine the relevant modality for his or her investigation, he does not really seem to account for these considerations.

4.2 Similarity in supervised classification

This second challenge remains as relevant for supervised classification as it was for the Basic System. In a classification problem we would be presented with a set of ladybug observations. These observations are divided into certain classes and our aim is to predict the class a future observation belongs to. To be able to do this we define a set of features based on which we might determine the appropriate class. We judge ourselves capable of determining the appropriate class since we believe that future observations will be similar to past observations in such a way that if we have observed certain feature values corresponding to a specific class, then future observations also having these feature values will be similar by belonging to the same class. This entire statement conceals multiple assumptions that are challenged by Goodman's critique.

First of all, Goodman might ask on what grounds we define the set of features we deem to be relevant for forming future predictions. Say that we have indeed observed multiple ladybugs and that we have divided these ladybugs into two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Furthermore, we consider the number of dots of these ladybugs as the appropriate feature based on which we can assign an observation to one of these classes. Subsequently, we spot a before unseen ladybug and assign it to class $\mathcal{C}_1$. At the same time, it could just as well be so that if we had considered the size of the ladybugs as the salient feature, we would have assigned the ladybug to class $\mathcal{C}_2$. In the context of judging size as the relevant feature we would thus end up with a different prediction for the same observation. Why then do we consider ourselves justified in considering the number of dots as the relevant feature? Formally Goodman's critique would confront supervised classification methods with the question of why we are justified to assume that future observations are similar to past observations because they share the property that their feature values fall within a certain region

within feature space.

While Carnap does not present any help in answering this question, supervised classification methods actually do so in the form of elaborate feature selection procedures. These procedures are supposed to find the most appropriate feature representation for a specific classification problem. Based on the resulting feature set we should be able to accurately formulate predictions about future observations. Lets next consider how these procedures are supposed to justify our selection of features.

4.2.1 Feature Selection

In most supervised classification problems we acknowledge that we have little knowledge about which features are relevant for predicting the class an observation belongs to [42]. To overcome this problem we usually introduce many candidate features to ensure a complete representation of the problem domain. This way we increase the probability that the relevant features are at least among the set of features we initially consider. At the same time, this will inevitably result in the inclusion of many features which are redundant or irrelevant for the target variable. We therefore often apply automated feature selection procedures which aim at selecting a subset of m highly discriminant features from the originally presented set of d features.

It is important to distinguish feature selection from feature extraction. Where the former refers to algorithms that aim to select the best subset of the input feature set, the latter refers to algorithms creating new features based on transformations or combinations of the original feature set [31]. Often the extraction of features precedes feature selection: we first construct features out of the sensed data, after which only a subset of the extracted features is used by our classification algorithm. Although it may certainly be interesting to examine the relevance of Goodman's critique for feature extraction methods, in this paper I will only focus on feature selection methods. I will therefore assume that we are presented with a set of features, which might or might not be the result of an automated extraction procedure.

A feature selection method can typically be characterized by the following four steps: subset generation, subset evaluation, stopping criterion and result validation [42]. The way these steps are executed directly influences which feature set will be selected. In the first step some search strategy is used to select a candidate

feature subset. This subset is evaluated based on some evaluation criterion in the second step. Subsequently, the subset that best fits the evaluation criterion after the stopping criterion has been met will be chosen. This subset is usually validated using prior knowledge of the problem domain. The most straightforward approach to go through these steps would be to consider all possible subsets of the original feature set and then select the best fitting subset [31]. However, because the number of possible subsets grows combinatorially, such an approach is practically impossible. For this reason we commonly apply methods that find a solution by only testing some of the possible feature subsets. These models can broadly be categorized into filter models, wrapper models and embedded models [42].

Filter models select a feature set based on general characteristics of the data in the training set [42]. These characteristics are supposed to represent some kind of relevance measure that determines how well a feature or a set of features is capable of distinguishing between the different class labels assigned to the training data points. Interestingly, Carnap was one of the first to investigate the notion of relevance [5]. His work still seems to be at the roots of most modern day relevance measures. Filter models usually offer an efficient way to find a set of features. At the same time, these models totally ignore the effects of the selected features on the performance of our classification algorithm.

Wrapper models do not directly apply a measure of relevance [5]. These models utilize a specific classifier to evaluate the quality of a set of features [42]. First a specific subset of features is selected. Based on this subset we train a classification algorithm on the training set. Subsequently, the remaining dataset is divided into a test set and a validation set. The classifier resulting from the first step is then evaluated on the validation set. Usually the classification error is used as an evaluation criterion [31]. The subset of features based on which our classifier makes the least number of misclassifications is selected as the best subset. Naturally, using a different classification algorithm or a different evaluation criterion would result in a different subset of features being selected. Although wrapper models often result in better performance of our classifier in comparison to filter models, they tend to be computationally expensive when considering a large number of features [42].

Finally, embedded models combine the best of both worlds. These models first

apply relevance measures as is done in filter models. Based on these measures several subsets of the original feature set are selected. Out of these candidates the subset with the lowest classification error given some classification algorithm is selected. Embedded models commonly achieve both efficient feature selection and good performance of the classification algorithm [42].

All three models consider a feature relevant based on its capability of distinguishing between the class labels we have assigned to the data points in the available dataset. Filter models do this by capturing this capability in a relevance measure, wrapper models do this by directly testing the performance of a learned classifier. We justify our selection of features on the ground that based on these features we are capable of accurately discriminating the predefined classes in the observations we have already made. It is then up to our learning algorithm to assign weights to the features in this set such that we may accurately classify future observations. In the ladybug example we would select both the number of dots and size as relevant features since they allowed us to discriminate between the classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Based on the ladybugs we have already observed we decide if size or number of dots should be conclusive when both suggest contrasting predictions.

It follows that our selection of relevant features is relative to the way we have divided the data into classes. Only in the context of predefined classes are we able to determine which features are relevant for our inductive investigation. This might seem trivial, but note how this is exactly the point Goodman makes. Given a different context, here understood as another way of assigning data points to class labels, we judge different sets of features to be relevant for making similarity judgments between our observations.

4.2.2 Class Selection

Goodman would naturally follow up his first question by asking on what grounds the way we assigned our observations to classes is justified. After all, the answer to his first question of how we decided on the relevant set of features depended on the way we labeled our observation. Thereby, given the same set of objects, there might be many alternative ways in which we could divide these objects into classes. On what grounds do we then distinguish between these alternative ways of classifying

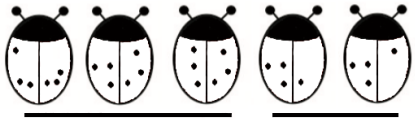

Figure 8: Class assignment based on number of dots.

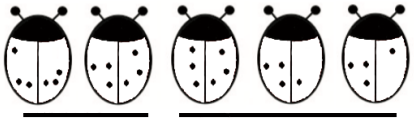

Figure 9: Class assignment based on weight.

our observations? The weight of Goodman's critique shifts from justifying why we prefer one set of features over another to justifying why we designate one way of assigning our observations to classes over another.

To be able to divide our observations into classes, we have to pinpoint some specific properties based on which they could be assigned to different classes. These classes would then consist out of the observation that are similar on the specified properties. But this is exactly the kind of task Goodman considers to be impossible. There might be many properties based on which we could define such classes. Take the five ladybugs we have observed. They could be divided into classes based on their number of dots, resulting in the class assignment depicted in 8. But at the same time we could also do this based on their weight, which would result in the class assignment depicted in 9. Whichever property we deem to be relevant for classifying our observations, there will always be a justifiable alternative that would lead to different classes. And even if we have decided on a certain property, on what grounds do we assign a specific data point to one of these classes? What are the criteria for assigning a class label to a single observation? Inevitably we have to make some equally justifiable decisions in this process. What then were our considerations when making these decisions? Our only option is to rely on our prior knowledge of the problem domain. We necessarily have to introduce some bias into our learning process by assigning weights to a specific set of properties when dividing objects into classes.

Again, there turns out to be a well-known theorem in the supervised classification literature formalizing exactly this point, namely Watanabe's ugly duckling theorem [44]. The ugly duckling theorem asserts the impossibility of classification without considering certain features or aspects of objects as more important than others [32]. This is proven by considering a case in which 2^n ducklings are represented by n binary

features, such as having black feathers or a fat body. Based on these features the ducklings are classified into positive or negative classes. Furthermore, the positive class is represented by Boolean functions of these binary features. Given this set-up it can easily be shown that there is an equal amount of ways to discriminate between an ugly duckling and a normal duckling as there is to discriminate between two normal ducklings. The main point is that every duckling resembles a normal duckling and an ugly duckling equally. Given enough properties of these ducklings there will always be equally justifiable ways to divide them into classes. This result follows from considering every feature to be equally important. To be able to classify a duckling as ugly we need to select some arbitrary feature that would make a duckling ugly.

Does this mean that we are never really justified in the way we have assigned our observations into classes? Not at all! We always have to introduce prior knowledge or make some assumptions in order to investigate anything at all. Often we have good reasons to assign certain objects to specific classes. Thereby, the classes we will use in our investigation are often dictated by the initial goal of our investigation. The above mainly purports to show that classes are not naturally in the data, we put them in there. Accordingly, the knowledge resulting from supervised classification procedures is as good as the knowledge we put into these procedures. The success of a classification learning algorithm stands or falls with the truthfulness of the assumptions we necessarily have to introduce when applying such an algorithm. Thereby, it certainly applies that we have to add something more than just data in order to ever acquire valuable scientific knowledge from a computer program.

5 Principle of Indifference

In the previous section we examined how we decided on the way we represent our observations in a supervised classification problem. In this section I will assume that we have decided on such a representation. Our observations already have been assigned to class labels and based on these labels we have found a suitable set of features. We will see next that to be able learn from our observations, we have to introduce significant assumptions about how the classes we defined are distributed over the observation space defined by the features we selected.

5.1 Bertrand's paradox

Joseph Louis François Bertrand (1822-1900) was a French mathematician who presented a well-known paradox for the Principle of Indifference (POI). This principle states that whenever we do not have information distinguishing between a set of mutually exclusive and jointly exhaustive alternatives, we should assign the same probability to each alternative [40]. Because of the complexity of the paradox originally presented by Bertrand, the point it tries to make can better be transmitted by a simpler example presented by van Fraasen [18]. Suppose we know of a box factory which produces cubes with side-lengths ≤ 2 cm. Further suppose that this is all we know about the cubes. We ask ourselves what probability we would assign to the next box produced having side-lengths ≤ 1 cm. Following the Principle of Indifference we would assign a probability of $\frac{1}{2}$ to this event. Next we do not consider the side-lengths but the side-areas. Now we ask ourselves what probability we would assign to the next box having side-areas ≤ 1 cm^2. Again following the Principle of Indifference we would assign a probability of $\frac{1}{4}$. Note however that the event of a cube with side-lengths between 0 and 1 cm or side-areas between 0 and 1 cm^2 is the same event. The Principle of Indifference would thus prescribe inconsistent probabilities to the same event.

This paradox weighed heavily on Carnap's earlier work on inductive logic. In his earlier work Carnap applied the POI directly on the predicates of our linguistic framework. Since there was no way to distinguish between these predicates, we should assign an equal prior probability to every predicate P_j. The following example

shows how this leads to a similar paradoxical situation as the one presented by van Fraasen. Suppose we decide to record the color of our observations and only make use of a "red"/"non-red" distinction. Since we use two predicates, we would assign both predicates a prior probability of $\frac{1}{2}$. On the other hand, if we would have decided to use the predicates "red", "orange" and "not-red-or-orange" we would assign each predicate a prior probability of $\frac{1}{3}$. Now the problem arises that, given no information, using a different partition of the modality to be observed leads to inconsistent prior probability assignments to the same predicate ("red"). Which of these different partitions should we then prefer over the other, and why? Carnap does not offer much help in answering this question. However, despite these objections against the POI, he kept believing its basic idea to be sound. Therefore, in the Basic System he tries to apply the principle in such a way that the above problem is avoided.

With the introduction of attribute spaces Carnap presented a way to differentiate between predicates of the same family. By representing predicates in an attribute space they can be distinguished based on their size and position within this space. This means that the predicates do not form a set of indistinguishable alternatives anymore. Applying the POI on the level of the predicates is no longer an option. This way, inconsistent probability assignments when considering different ways of carving up the modality should be evaded. Consider for example the attribute space we used for our ladybug investigation (figure 7). We assign a prior probability of $\frac{1}{5}$ to the predicate "red" (B_1) based on the width of its regions. Now suppose that we decide to use a different carving-up of our attribute space and collapse the regions $B_2 - B_5$ to one big region. This corresponds to the case in which we would only use the predicates "red" and "non-red". Our attribute space would then look like the space depicted in figure 10. As can be seen, the region assigned to "red" remains the same and thereby the prior probability assigned to it as well, namely $\frac{1}{5}$. Thus, the possibility of assigning inconsistent prior probabilities to the same predicate seems to be avoided by the introduction of attribute spaces [16].

The way that Carnap conceived of the γ-rule reveals how the POI is still part of his inductive logic. The γ-rule presupposes the idea that indifference considerations should be applied on the level of the points of the attribute space rather than on the predicates of the language [41]. The attribute space presents us with the privileged

Figure 10: Alternate attribute space for ladybug investigation (figure taken and modified from [40])

set of alternatives which should be considered equiprobable, namely the elementary properties of the modality this space represents. To clarify this point, again consider the color space. The points making up this space represent all the possible ways in which an object can have color. The space thus captures all that is empirically possible, it defines the set of shades of color we can observe in our investigation. It is precisely this set of shades of color to which the POI should be applied. Every specific shade of color we might observe should be considered equiprobable, since there is no information to distinguish between them. The indifference considerations applied in the Basic System can then be formalized as follows: because we have no reason to think that certain elementary properties of the modality we consider will be observed more often than others, initial probabilities should be equally distributed over the points of the attribute space representing this modality [40]. Subsequently, we can divide these points into a family of predicates in various ways via attribute partitions. It follows that initial probabilities for regions defined by such partitions will indeed be equal to the widths of such regions.

With this alternate way of applying the POI Carnap wanted to address the objections to this principle. In his earlier work the POI could lead to inconsistent probability assignments since it was unclear how one set of equiprobable alternatives should be preferred over another. Attribute spaces are supposed to solve this problem by presenting the privileged set of alternatives. However, this solution just shifts the problem from deciding between different partitions of our modality to deciding between different ways of representing such a partition in the attribute space. Consider the attribute space in figure 11. This space represents all elementary properties of the modality in question, as denoted by the integers from 1 to 10, and thus captures all the possible observations we can make in our investigation. Next, we

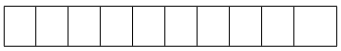

Figure 11: Attribute space.

Figure 12: Initial probability of $\frac{1}{6}$ assigned to "red".

Figure 13: Initial probability of $\frac{1}{4}$ assigned to "red".

have to divide this space into regions of a certain width which represent the predicates we have defined in our conceptual framework. These predicates will form the subjects of our predictions. Suppose that we have only defined two predicates P_1 and P_2. Figure 12 shows a possible way of distributing the predicates over the attribute space. P_1 is represented by the red area and P_2 by the green area. Given this space we could ask ourselves what the probability is of next observing a P_2 thing given that we have observed the specific property 2 which we consider to be a P_1 thing. In this case, this would be similar to asking what the probability is of next observing the property 7, 8, 9 or 10. Off course the answer to this question strongly depends on the way the predicates we defined are distributed over the attribute space. And since Carnap does not present a clear recipe for determining the width corresponding to a predicate, we seem to be free to distribute the predicates over our space in a different way. We could for example use the space depicted in 13. First of all, this space would lead to different initial probabilities for both predicates. Presuming that all separate points in our attribute space should be considered equiprobable, P_1 should get assigned a prior probability of $\frac{1}{6}$ when using the space in figure 12 and a prior probability of $\frac{1}{4}$ when using the space in figure 13. Second, this space also captures different judgments about which elementary properties we consider to be a P_2 thing. Both attribute spaces would therefore lead to different predictions given the same set of evidence. Which of these attribute spaces should we then prefer, and on what grounds? The same problem presents itself in a different form.

5.2 No free lunch theorem

This problem can also be voiced against modern day supervised classification methods. And as it turns out, the high relevance of this problem is actually well-known. To explicate the formulation of this problem in a supervised classification setting

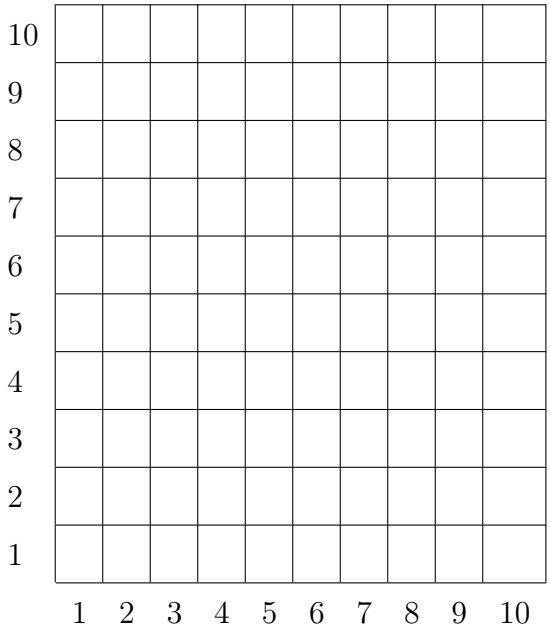

Figure 14: Feature space.

consider a feature space $\mathbf{X}$ defined by two features. These features can take values in the range of integers from 1 to 10. Figure 14 presents a grid representing this space. Every area in this grid corresponds to a specific feature vector and thus to a specific observation. Note how these areas can be conceived of as corresponding to the individual points an attribute space spans over. Further suppose that we have defined two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$. Whereas in Carnap's Basic System every possible observation would be assigned to a specific predicate, we now assume that every feature vector has a certain probability of belonging to a class. Given this set-up we would like to answer similar questions as we did in the Basic System. We could for example ask ourselves what the probability is of next observing a specific feature vector belonging to $\mathcal{C}_2$ given that we have observed a certain feature vector belonging to class $\mathcal{C}_1$. Although formulating the questions we ask in supervised classification in this way clearly shows the correspondence to Carnap's Basic System, remember that we would always base our predictions on more than just one observation. Eventually our goal is to find the best classifier capable of answering such questions based on a set of observations.

Now consider what it means for a classifier to be the best classifier. Intuitively we would say that the best classifier should make the most accurate predictions for future observations. Such a classifier should not only result in the lowest number of

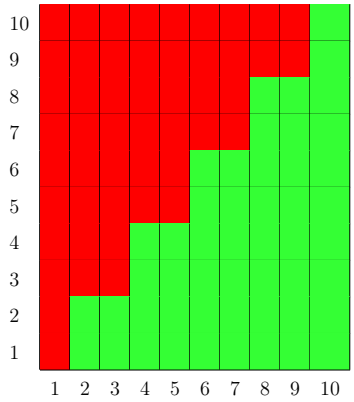

Figure 15: Probability distribution P_1.

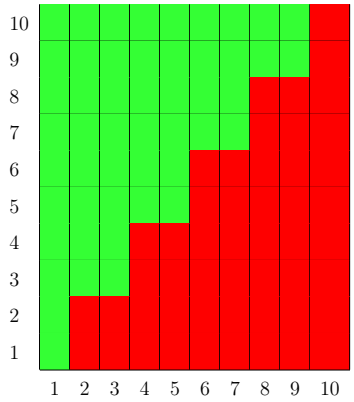

Figure 16: Probability distribution P_2.

misclassifications on the test set, but on any set of new observations. To be able to estimate the number of misclassifications a classifier will make on a new set of observations we need to know the true class probabilities $P(\mathcal{C}_k|\mathbf{x})$ for every possible $\mathbf{x}$. Based on these probabilities, we could estimate the probability our classifier would misclassify any new observation. But remember that we are agnostic about the joint probability distribution P over the feature space X and the classes $\mathcal{C}_k$. Consequently, our best option is to find a classifier that performs well on any distribution P. The best classifier would then formally be defined as the classifier that achieves the lowest number of misclassifications averaged over all possible distributions P. However, as we shall see next, given this definition, there will not be any classifier outperforming another classifier [43].

Suppose that we are presented with a small training set $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ and that these observations form a subset of the feature space in figure 14. Further suppose that all points in the feature space that are not part of the training set form the test set. We apply a classification algorithm on the training set and find a classifier. Formally this classifier would be the best classifier if it acquires the lowest number of misclassifications averaged over all possible distributions underlying the data we have observed. Figure 15 depicts such a possible distribution P_1 with the red and green areas indicating the observations for which the probability is highest of belonging to class $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively. Distribution P_1 corresponds to a certain labelling of the data points in our test set. Now imagine that our classifier would perfectly label all points in this test set. Subsequently, it can easily be seen that there will always be another possible distribution for which our classifier will make the largest possible

error. This distribution would corresponds to an inverse labelling of the data points in the test set. In our example this would be the distribution depicted in figure 16. Thus, on average over both possible distributions our classifier would not perform better than random guessing. In the same way it can be shown that for any given error R there exists a probability distribution P such that the error of our classifier on the test set will be R. It follows that averaged over all possible distributions P, every classifier will achieve the same test error: whenever there is a distribution where the classifier performs well, there is a corresponding "inverse" distribution on which the classifier performs poorly [43]. And even more troublesome, no classifier will be judged to perform better than random guessing on the test set.

This problem is well known in the supervised classification literature and usually stated in the form of the so-called no free lunch theorem [43]. The no free lunch theorem builds on the argument that we consider average performance over all probability distributions and that we consider these distributions to be equally likely. By considering all distributions, we also consider those distributions where the labels are assigned to data points "without any system". Such a distribution could for example be constructed by tossing a coin and assigning true labels to data points based on the outcome. Knowing the labels of the training points then does not seem to add anything. We therefore need some way to exclude these kinds of distributions from the space of probability distributions under considerations.

This is exactly the same problem Carnap's Basic System was faced with. The Basic System is centered around the question: given an observation and a specific distribution of prediction labels over our attribute space, what is the probability of next observing another kind of observation. Inconsistent predictions resulted from different ways of distributing the labels of our predictions over our attribute space. We therefore needed a way to prefer one of such distributions. Supervised classification seems to solve this problem by aiming to learn the true distribution of labels, which should lead to the best predictions. We now ask ourselves: given this set of observations, from which this distribution of classes over our feature space follows, what is the probability that we will next make an observation belonging to a certain class. To be able to make such predictions, we need to specify a privileged set of distributions among which the true one can be. So again we need some way

to favour some distributions of our labels over others.

The no free lunch theorem breaks down by formulating assumptions about the shape of the probability distribution underlying the data we have observed, as for example done in the generative methods discussed in 2. Such restrictions on the space of probability distributions come in various forms. The main point of consideration here is that our predictions about the future based on a set of observations crucially depend on the assumptions we make about these observations. To be able to learn we have to assume that one out of a restricted set of probability distributions underlies the labelling of our observations. Apart from mentioning that we should be informed by our scientific theories when formulating these assumptions, Carnap does not offer much help. On the other hand, in a supervised classification problem we have a set of observations at our disposal. Based on this set we could formulate an educated guess about the distribution underlying our observations. Thereby, we might exclude many possible distributions based on prior knowledge of the problem domain. Berthrand's paradox and the no free lunch theorem therefore mainly seem to show the need to explicate the grounds based on which we form the strong assumptions needed to be able to learn from past experience.

6 The New Riddle Of Induction

In the final section of this thesis I will examine the relevance of another famous critique by Nelson Goodman for supervised classification. In "Fact, Fiction and Forecast" Goodman introduced his widely discussed "New Riddle of Induction" which raised a serious problem for Carnap's hope of developing an inductive logic [24]. The riddle Goodman presents is based on a simple device for generating alternative hypotheses that fit past data perfectly well but which have very different implications for the future [39]. We will discover that for this reason we do not only have to introduce restrictions on the set of possible distributions underlying our data but also on the class of candidate functions capable of modelling these distributions.

6.1 Grue

In the "New Riddle of Induction" Goodman wonders how to formulate rules that define the difference between valid and invalid inductive inferences [24]. Formulating these rules would come down to defining the relation that obtains between an evidence statement S_1 and a hypothesis S_2 if and only if S_1 may properly be said to confirm S_2 in any degree. In his search for such a definition, Goodman quickly comes to the conclusion that confirmation of a certain hypothesis depends rather heavily upon features of the hypothesis itself. This conclusion is reached by the following considerations about pieces of copper and men who are third sons. Observing a specific piece of copper conducting electricity would confirm the statement that other pieces of copper also conduct electricity, and thus the hypothesis that all pieces of copper do so. At the same time, finding out that a man in some room is a third son, would not at all confirm that all men in that same room, or anywhere, are also third sons. As noted by Goodman, the crucial difference between the two hypotheses that all pieces of copper conduct electricity and that all men are third sons is that the former is a lawlike statement while the latter is just an accidental generality. Thus, Goodman concludes that only lawlike hypotheses may be confirmed. Hence, we have to come up with some way to discriminate between lawlike and accidental hypotheses. However, this is not as simple as it might seem, as made clear by Goodman's famous grue problem. This problem formed a big threat for Carnap's

inductive logic. For this reason I will also present the problem in the setting of the Basic System.

Goodman asks us to imagine that we are investigating emeralds and that we want to form future predictions based on the color of the emeralds we have observed. We define a attribute space representing this modality and start making observations. Suppose that all emeralds we observe before a certain time t are green and are thus represented in the space by the predicate green. Since Carnap has stated that confirmation for a future observation should mainly be determined by the amount of times we have already made the same observation, at time t our observations most strongly confirm the hypothesis that the next emerald we observe will also be green. So far everything appears to be fine.

Now consider the predicate grue that applies to all things examined before t just in case they are green but to other things just in case they are blue [24]. Observing the same emeralds would now result in a set of evidence statements asserting that the observed emerald is grue. In this case, at time t our observations would most confirm that the next emerald will be grue. The predictions that the next emerald will be green and that the next emerald will be grue are thus equally confirmed by our evidence statements of the same observations. However, if the next emerald is grue, then it is blue and not green. And although we know that it is actually the green hypothesis that is confirmed by our observations, following the Basic System both the grue and green hypotheses would be equally confirmed. The same set of observations seems to equally confirm inconsistent hypotheses.

Lets state the same problem in a supervised classification setting to clarify the relevance of this problem for its methods. This time, to stay true to Goodman, we will not consider ladybugs, but emeralds. Up till time t we have stumbled upon multiple emeralds. Given these observations, we would like to predict if the emeralds we might find in the future are green or blue. As dictated by this goal we first divide the emeralds into two classes, corresponding to green and blue emeralds. Furthermore, from the emeralds we have already observed we learn that we might be able to accurately predict the color of an emerald by the longitude and latitude of the location where it is found. Consequently, we represent all emeralds in the feature space defined by the longitudes (X_1) and latitudes (X_2) we consider relevant

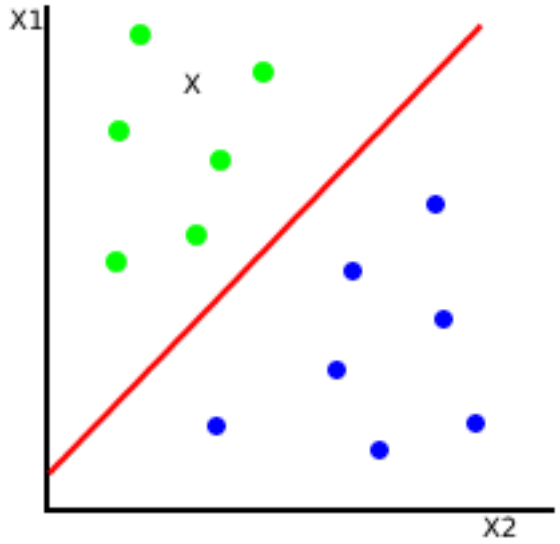

Figure 17: Feature space for emerald investigation.

for our investigation. Subsequently, the observed emeralds are distributed over a training and test set and we apply a classification algorithm like the ones discussed in section 2 to find the classifier that best fits the data. This classifier is depicted in figure 17 including some data points out of the test set. Given this classifier we would predict the before unseen emerald e, corresponding to the feature vector depicted by X, to be green.

It can easily be shown that if we would have used grue instead of green to describe our observations, the classifier best fitting the data would divide the feature space into decision regions in the exact same manner. The only difference would be that the green region would now be considered the grue region. Accordingly, the emerald X would have been classified as being grue, and hence not green. Even more troublesome, based on our observations there would be no way to discriminate between these inconsistent predictions. The classifier predicting e to be grue will fit the same observations equally well as the classifier predicting e to be green. And since t might be any moment in the future, there is actually an infinite number of alternative hypotheses which would all be equally well confirmed by the same data. Based on the data there is no justification for judging one of these hypotheses to be more confirmed than the others.

An immediate response to this example could be to state that we are not really faced with a problem because we would always be able to rule out the grue classifier

based on background knowledge. For example, we may already have the knowledge that the color of an emerald is determined by environmental factors such as the composition of the rock in which it is found. Furthermore, knowing that these factors remain constant over a large area would make it very unlikely that all emeralds found in the same area are grue. It might even be the case that based on the way emeralds are usually excavated we would consider it highly unlikely that emeralds are grue. However, the problem can easily be formulated in a scenario in which we could not rely on background knowledge. Consider for example the less known scenario presented by Goodman in which marbles are drawn from a bowl without any information about the process by which marbles are selected. Our goal is to predict the color of the next marble based on its size and weight. Following the same steps as in the emerald investigation, we would again have to conclude that there is now way to distinguish between inconsistent predictions based on the observations we made up to some moment in time t. For all we know, the sampling process might be biased towards selecting green balls of certain weight and size before t and blue balls of similar weight and size after t [31]. This example directly confronts us with the question of how to distinguish between inconsistent predictions equally well confirmed by the data when we do not have any prior knowledge of the problem domain.

A popular way of attacking the problem is based on excluding grue-like hypotheses because these hypotheses typically involve some spatial or time restrictions, or reference to some particular individual [24]. We could for example only consider classifiers that would not involve a specific moment in time like t. A grue classifier would always have to refer to some t and should thuerefore be excluded from consideration. This approach is based on restricting the space of possible classifiers considered by our classification algorithm to only those classifiers of a specific kind. Carnap took a similar approach by stating that lawlike hypotheses could only consist out of purely qualitative properties [10]. Purely qualitative properties would then be properties that can be expressed without the use of any individual constant. Such properties do not need to refer to any specific place, time or individual. Grue would certainly not fall into this category. But as Goodman argues, it is not so clear why grue hypotheses would and green hypotheses would not have to involve a time

reference.

Consider "green" and "blue" and "grue" and "bleen", where "bleen" applies to emeralds (or marbles) examined before time t just in case they are blue and to other emeralds just in case they are green [24]. Sure enough, following the above approach we would have to exclude classifiers capable of capturing the concepts "grue" and "bleen" because they involve a time reference. But, as made clear by Goodman, this is only so because we initially describe our observations as "green" and "blue". If we start with "green" and "blue" then "grue" and "bleen" would indeed be explained by "green", "blue" and a time reference. But, if we would have initially described objects as "grue" and "bleen", this would be the other way around. The concept "green" would then apply to grue emeralds observed before t and to other emeralds if they are bleen. Which classifiers would have to be excluded is completely relative to the way we initially describe our observations. Stating that "green" and "blue" are the only right way to have initially described our observations would be completely begging the question. Attacking this problem by making some distinction between the concepts green and grue thus does not seem to work.

6.2 VC dimension

A more recent analysis of the grue problem presented by Steel suggests that the problem can be solved by emphasizing some epistemically significant characteristic of hypotheses beyond fit with the data [39]. Up till this point, we only considered evaluating classifiers based on their fit with the data. We considered the best classifier to be the classifier with the lowest error on the observed data. Unfortunately, we discovered that this did not provide us with any ground to prefer the green classifier over the grue classifier. We therefore attempted to formulate restrictions on the set of candidate classifiers in such a way that the grue classifier would be excluded. Another approach would be to introduce an alternative learning procedure that balances error on the data against something else - often called simplicity [28]. The goal of our learning procedure would then be to find the classifier that minimizes a value that captures the balance between fit to the data and simplicity. This is exactly the kind of approach examined by Steel. He shows that the grue problem might be solved by expressing the simplicity of a hypothesis in terms of a central concept in

Figure 18: Shattering three points (figure taken from [27])

Figure 19: Not shattering four points (figure taken from [27])

the theory underlying most classification algorithms, namely the VC dimension.

Remember that in a classification problem we always consider a set of possible classifiers out of which we want to select the one most confirmed by the data. VC dimensions are used to express the richness of such sets of classifiers [27]. This is done by means of the concept of shattering [39]. A set of classifiers F shatters some set of N data points in feature space if and only if for any labeling of those points, some rule in F is capable of perfectly classifying the points so labeled [39, 27]. The VC dimension of this set of classifiers is then defined as the largest number N such that some set of N points in feature space is shattered by rules in F. To clarify this concept, consider that we want to determine the VC-dimension of the set of linear classifiers for the case where we want to classify data points into two classes. Now consider some set of three data points and all the possible ways these points might be divided into two classes, such as depicted in figure 18. For every possible labeling there is a linear classifier capable of correctly classifying all three points. The set of linear classifiers thus shatters some set of three data points. In a similar way it can be shown that there is no set of four point shattered by the set of linear classifiers. Figure 19 shows some possible sets of four labelled data points for which there is no linear classifier that would correctly classify all points. It follows that the VC dimension of the set of linear classifier is 3.

Steel shows us that the grue problem as originally presented by Goodman can be solved by always conjecturing the hypothesis that is consistent with the data from the set of hypotheses with the lowest VC dimension [39]. This solution is based on the assertion that the VC dimension of the set of hypotheses containing the grue hypothesis is higher than the VC dimension of the set of hypotheses containing the green hypothesis. Although both hypotheses are equally consistent with the data, the green hypothesis is associated with a lower VC dimension and should therefore

be preferred. Thus, following the learning procedure explicated by Steel, we would always favor the green hypothesis over the grue hypothesis.

The exact details of this solution in the context of supervised classification are unknown to me. However, I suspect a solution based on the VC dimension to be mistaken on the grounds that the VC dimension appears to be a language dependent concept. I conjecture that the VC dimension, or any other way of capturing simplicity of a hypothesis, is an equally relative matter as simplicity itself. That simplicity is a relative matter has already been made clear by noticing that green and blue are only considered to be simpler than grue and bleen because we explain these concepts in terms of green and blue. However, green and blue could just as well be explained in terms of grue and bleen, which makes grue and blue simpler. Given the way the VC dimension is defined, it appears to be highly probable that a similar example can be formulated concerning this concept. This is endorsed by noting that the notion of shattering, based on which the VC dimension is defined, presupposes some preferred language in which the data points are expressed. After all, to be able to determine if a classifier might be able to correctly classify some data points, we have to assume some language in which these data points are expressed. A simple translation of the preferred language might result in a lower VC dimension for the set of classifiers containing the grue classifier in comparison to the set of classifiers containing the green classifier. Consequently, we would be back at the original problem of stating why we would prefer the concept green over the concept grue.

6.3 Why Bother?

We might wonder why the grue problem is relevant for modern day supervised classification at all. We might for example state that it is obvious that if we artificially define our primitive concepts, then the world's actual behavior becomes more cumbersome to describe [1]. We should not be surprised if we would then make wrong predictions. In practice it is very unlikely that we will ever use a grue-like classifier when we make predictions. After all, we will most probably never be in a situation in which we do not have any background knowledge to refute such classifiers. If we are able to make accurate predictions about the world around us based on the

primitive concepts we normally employ, why would we bother about artificial concepts like grue? Such a question seems to imply that we do not need a consistent theory of induction at all. Although in practice we will most probably not encounter the artificial examples considered in this section, they are all well defined examples that present a significant challenge for modern day inductive endeavours. With a growing application of supervised classification methods in various fields of science it becomes untenable to claim that we do not need to worry about such examples. Accounting for the grue problem therefore remains as relevant for supervised classification methods as it once was for Carnap's inductive logic. So far, we do not seem to have encountered any satisfactory answer to this problem

Finally, our best hope of such an answer comes from Goodman himself. The solution Goodman presents does not attempt to justify our decision of "green" over "grue", but provides an account of how we actually choose such predicates for induction [13]. According to Goodman we favor predicates like "green" over predicates like "grue" because in the past we have successfully used hypotheses involving "green" much more often than hypotheses involving "grue". Since we have used the predicate "green" more often than "grue", we would say that "green" is better entrenched [24]. Based on this notion of entrenchment we are capable of discriminating between a green classifier and a grue classifier. Because the former is better entrenched, we should always prefer it over the latter when both are equally confirmed by the data. Goodman thus proposes that a solution to the grue problem can be found in the language we use and have been successfully using for a long time.

7 Conclusion

The aim of this thesis has been to examine how criticism voiced against Carnap's inductive logic might still be relevant for modern day machine learning methods. Carnap's last published work on inductive logic, "The Basic System of Inductive Logic", formed the starting point of my investigation. Thereby, since a general claim about the entire field of machine learning would be prone to counterexamples, I decided to only focus on linear supervised classification methods. By highlighting the resemblance between the reasoning applied in such methods and in Carnap's work on inductive logic, it became possible to formulate strong criticism against supervised classification methods that before had been aimed at the very core of Carnap's inductive logic. I have considered three such critiques that all purported to display the inadequacy of Carnap's system of inductive logic.

In section 4, I considered Goodman's "Seven Strictures of Similarity" in which he dismissed any kind of reasoning based on similarities by arguing that similarity is a flawed notion. This significantly undermined Carnap's Basic System since the inductive reasoning formalized in this system relied heavily on similarity judgments between observations. In the context of the Basic System, the by Goodman proclaimed context-sensitivity of similarity judgments, led to the formulation of inconsistent predictions in different contexts. I argued that this point remains highly relevant for supervised classification methods by revealing how such methods also crucially depend on similarity judgments between the objects of our observations. To be able to apply supervised classification methods we have to define a way in which our observations will be assigned to classes. But, as Goodman made clear, there are actually many justifiable ways to divide objects into classes. We therefore have to introduce some prior knowledge into the learning procedure by specifying some arbitrary property based on which our observations could be assigned to classes. In this way, Goodman's critique showed us that the knowledge resulting from supervised classification procedures strongly depends on the decisions we make when assigning data points to classes.

Subsequently, in section 5, I assumed that we already defined the classes and the set of features based on which we represent our observations. Through the paradox presented by Bertrand it became clear that if we consider every possible probability

distribution that could be underlying the data we have observed as equally probable, then there would be no grounds to judge some classifier as better than others. This conclusion unveiled an interesting correspondence between Bertrand's paradox and the heavily discussed no free lunch theorem. To the best of my knowledge this correspondence has not yet been noted in the supervised classification literature. Both Bertrand's paradox and the no free lunch theorem show us that we have to introduce some restrictions on the set of possible probability distributions to be able to learn from our observations. This conclusion emphasized that the predictions that follow from applying supervised classification methods strongly depend on the assumptions we necessarily have to introduce to be able to apply these methods at all.

In the final section of this thesis, Goodman's grue problem revealed that there will always be inconsistent classifiers equally confirmed by our observations. We therefore have to introduce some restrictions on the set of classifiers we consider in a classification problem to ensure classifiers based on artificial concepts like grue are excluded. This result fits nicely with the conclusion that followed from considering Bertrand's paradox. First we have to define some restrictions on the set of possible distributions underlying the data we have observed, after which we only want to consider a specific set of classifiers capable of modelling these distributions. However, many attempts of formulating grounds based on which we could induce restrictions on the set of candidate classifiers turned out to be insufficient to rule out grue-like classifiers.

All three critiques mainly form a strong reminder that supervised classification methods can never be a stand-alone method for the discovery of new scientific laws. Without the careful introduction of prior knowledge about the domain in which a classification method is applied, no valuable knowledge can come out of such a method. The knowledge we derive from supervised classification methods is as good as the knowledge we introduce into them. No matter how big the dataset at our disposal, without a careful formulation of the assumptions necessary to be able to learn form this set, there is no sense in applying any automated learning procedure. This point becomes highly relevant in light of the growing presence of supervised classification methods in many scientific endeavors.

Finally, besides examining machine learning procedures within the context of the inductivist controversy, I have tried to convey the importance of philosophical consideration of the methods applied within this field. Many machine learning researchers seem to have dismissed foundational considerations of the field in favor of more practical considerations. Consequently, many modern machine learning methodologies have become completely void of any direct correspondence to philosophical concepts. At the same time, I would argue that there are tremendous opportunities for mutually beneficial interactions between philosophy and machine learning. For example, in this thesis it has become clear that the philosophy of science can learn a lot from advances in machine learning and its implications for the scientific method. At the same time, we have seen that well known problems within machine learning might find their roots in less know philosophical works, such as was the case for the no free lunch theorem and the ugly duckling theorem. These works might cast a new light on these problems as they present themselves for machine learning. Both examples show that philosophy and machine learning can benefit enormously from establishing a long-term dialogue.

References

[1] Scott Aaronson. "Why philosophers should care about computational complexity". In: *In Computability: Gödel, Turing, Church, and beyond (eds.* Citeseer. 2012.

[2] John F Allen. "Bioinformatics and discovery: induction beckons again". In: *BioEssays* 23.1 (2001), pp. 104–107.

[3] John F Allen. "Hypothesis, induction and background knowledge. Data do not speak for themselves. Replies to Donald A. Gillies, Lawrence A. Kelly and Michael Scott". In: *BioEssays* 23.9 (2001), pp. 861–862.

[4] John F Allen. "In silico veritas: Data-mining and automated discovery: the truth is in there". In: *EMBO reports* 2.7 (2001), pp. 542–544.

[5] David A Bell and Hui Wang. "A formalism for relevance and its application in feature subset selection". In: *Machine learning* 41.2 (2000), pp. 175–195.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[7] Rudolf Carnap. "A Basic System of Inductive Logic, Part I". In: *Studies in Inductive Logic and Probability, part I*. Ed. by Richard Jeffrey and Rudolf Carnap. University of California Press: Los Angeles, 1971, pp. 34–165.

[8] Rudolf Carnap. "A Basic System of Inductive Logic, Part II". In: *Studies in Inductive Logic and Probability, part II*. Ed. by Richard Jeffrey and Rudolf Carnap. University of California Press: Los Angeles, 1980, pp. 7–155.

[9] Rudolf Carnap. "Logical foundations of probability". In: (1962).

[10] Rudolf Carnap. "On the application of inductive logic". In: *Philosophy and phenomenological research* 8.1 (1947), pp. 133–148.

[11] Rudolf Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, 1952.

[12] Rudolf Carnap. "Variety, analogy, and periodicity in inductive logic". In: *Philosophy of Science* 30.3 (1963), pp. 222–227.

[13] Daniel Cohnitz and Marcus Rossberg. "Nelson Goodman". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University, 2016.

[14] Edoardo Datteri, Hykel Hosni, and Guglielmo Tamburrini. "Machine learning from examples: A non-inductivist analysis". In: *Logic & Philosophy of Science* 3.1 (2005), pp. 1–31.

[15] Lieven Decock and Igor Douven. "Similarity after goodman". In: *Review of philosophy and psychology* 2.1 (2011), pp. 61–75.

[16] Lieven Decock, Igor Douven, and Marta Sznajder. "A geometric principle of indifference". In: *Journal of Applied Logic* 19 (2016), pp. 54–70.

[17] Branden Fitelson. "Inductive logic". In: *Philosophy of science: An encyclopedia* (2005), pp. 384–393.

[18] Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989.

[19] Donald Gillies. "Artificial Intelligence and Scientific Method". In: (1996).

[20] Donald Gillies. "The problem of induction and Artificial Intelligence". In: (2003).

[21] Donald A Gillies. "Popper and computer induction". In: *BioEssays* 23.9 (2001), pp. 859–860.

[22] Piotr Giza. "Automated discovery systems and the inductivist controversy". In: *Journal of Experimental & Theoretical Artificial Intelligence* 29.5 (2017), pp. 1053–1069.

[23] Clark Glymour. "Android Epistemology: Computation, Artificial Intelligence". In: *Introduction to the Philosophy of Science*. Ed. by Merrilee H. Salmon. Hackett, 1992, p. 364.

[24] Nelson Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

[25] Nelson Goodman. *Problems and Projects*. Indianapolis: Bobbs-Merrill, 1972.

[26] Teddy Groves. "Let's reappraise Carnapian Inductive Logic!" PhD thesis. University of Kent, 2015.

[27] Gilbert Harman and Sanjeev Kulkarni. *Reliable reasoning: Induction and statistical learning theory*. MIT Press, 2012.

[28] Gilbert Harman and Sanjeev Kulkarni. "Statistical learning theory as a framework for the philosophy of induction". In: *Philosophy of statistics*. Elsevier, 2011, pp. 833–847.

[29] Risto Hilpinen. "Carnap's new system of inductive logic". In: *Synthese* 25.3-4 (1973), pp. 307–333.

[30] Robin Holliday. "The incompatibility of Popper's philosophy of science with genetics and molecular biology". In: *Bioessays* 21.10 (1999), pp. 890–891.

[31] Anil K Jain, Robert PW Duin, and Jianchang Mao. "Statistical pattern recognition: A review". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000), pp. 4–37.

[32] Toshihiro Kamishima et al. "Efficiency Improvement of Neutrality-Enhanced Recommendation." In: *Decisions@ RecSys*. Citeseer. 2013, pp. 1–8.

[33] Douglas B Kell and Stephen G Oliver. "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era". In: *Bioessays* 26.1 (2004), pp. 99–105.

[34] Lawrence A Kelley and Michael Scott. "On John Allen's critique of induction". In: *Bioessays* 23.9 (2001), pp. 860–861.

[35] Hannes Leitgeb. "Logic in general philosophy of science: old things and new things". In: *Synthese* 179.2 (2011), pp. 339–350.

[36] Mauro Murzi. "Rudolf Carnap (1891-1970)". In: *Internet Encyclopedia of Philosophy*.

[37] Karl R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, 1962.

[38] Neil R Smalheiser. "Informatics and hypothesis-driven research". In: *EMBO reports* 3.8 (2002), pp. 702–702.

[39] Daniel Steel. "Testability and Ockham's razor: How formal and statistical learning theory converge in the new riddle of induction". In: *Journal of Philosophical Logic* 38.5 (2009), pp. 471–489.

[40] Marta Sznajder. *Inductive Logic on Conceptual Spaces*. 2017.

[41] Marta Sznajder. "What conceptual spaces can do for Carnap's late inductive logic". In: *Studies in History and Philosophy of Science Part A* 56 (2016), pp. 62–71.

[42] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review". In: *Data classification: algorithms and applications* (2014), p. 37.

[43] Ulrike Von Luxburg and Bernhard Schölkopf. "Statistical learning theory: Models, concepts, and results". In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

[44] Satosi Watanabe. *Pattern recognition: human and mechanical*. New York: John Wiley & Sons, Inc., 1985.

[45] Jon Williamson. "A dynamic interaction between machine learning and the philosophy of science". In: *Minds and Machines* 14.4 (2004), pp. 539–549.

[46] Jon Williamson. "The philosophy of science and its relation to machine learning". In: *Scientific Data Mining and Knowledge Discovery*. Springer, 2009, pp. 77–89.