Discretization in the context of (non-)monotonicity

MSc Thesis Artificial Intelligence

Judith van Dulst, 3470822
Supervisor: Silja Renooij
12/02/2019
Universiteit Utrecht

Contents

1 Introduction

Agents in Artificial Intelligence have to be able to act in environments with uncertainties. For example, they may have to interact with other agents of which they cannot be certain of how they will act. For agents to be able to reason about these uncertainties they can use probabilistic reasoning. Bayesian networks are models of joint probability distributions over sets of variables and are used for probabilistic reasoning. In Bayesian networks we almost always use discrete variables. However, there are many domains in which variables are actually continuous. One way to capture these continuous variables in a Bayesian network is by discretization of the variables for which various methods exist. With discretization, a continuous interval is divided into subintervals or *bins*, by placing so-called *cut points*. Most discretization methods focus on discretizing a single variable and typically assume a data set with data points for this variable, which are subsequently divided into bins. Variables, however, are often related to other variables and this relation may be affected by discretization of one of the variables. In practice, relations between variables often are monotone, which means that higher values for one of the variable gives a higher probability for higher values for the other variable. Or conversely, higher values for the input variable gives a higher probability for lower values for the output variable [9]. Monotonicity is a property that we would like to preserve during discretization. Therefore, we would like to know to which extent these properties of (non-)monotonicity are affected by discretization. Although there are various papers related to monotonicity ([7], [6], [10]) and to discretization methods ([2], [4]), to the best of our knowledge there is no research describing the relation between discretization and monotonicity.

Some widely used discretization methods, such as *Equal Frequency* and *Equal Width* focus only on the variable that is to be discretized. Other popular methods, such as *ChiMerge* [3] and *MDLP* [1] take other variables, typically a class variable, which gives the class to which the observation belongs, into account when discretizing a variable. For *Equal Frequency* and *Equal Width* the number of bins should be predetermined b the user. For these methods we will discuss how the number of bins should be chosen to best preserve monotonicity and we will give guidelines for this. For *ChiMerge* and *MDLP*, we will investigate how they handle (non-)monotonicity by examining some special cases theoretically and some other (non-)monotone probability distributions with experiments. The discretization methods prescribe between which values a cut point should be placed. With exception of *Equal Width*, they use data points as the start and end points of the bins. We will examine if there are other options to choose the placement of the cut points that may give a better discretization.

In Chapter 2 we will present the preliminaries. In Chapter 3 we will discuss monotonicity and we will examine when monotonicity is preserved or induced upon discretization independent of which discretization method is used. In this chapter we will also propose and study measures for a degree of monotonicity. In Chapter 4, 5, 6 and 7 we will subsequently examine the discretization methods

Equal Frequency, *Equal Width*, *ChiMerge* and *MDLP*. For each method we will discuss the advantages and disadvantages. In Chapter 6 and 7 we will theoretically study how *ChiMerge* and *MDLP* respectively work for some special cases. In Chapter 8 we will examine how *ChiMerge* and *MDLP* work for some (non-)monotone probability distributions by conducting experiments. Finally, in Chapter 9 we will propose other suggestions to determine where a cut point should be placed than what the methods prescribe.

2 Preliminaries

In this thesis we consider the discretization of stochastic variables based upon datasets containing samples of their values. We assume that datasets are of finite size, and as a result the number of different values recorded for a continuous variable can be large, but is finite. We therefore assume that any variable can be represented by a discrete stochastic variable $X = \{x_1, ..., x_n\}$ for which there exists some ordering $\prec$ on its values, such that $x_1 \prec ... \prec x_n$. The probability distributions for these stochastic variables are estimated from the same datasets.

Definition 1. A *dataset* is a multiset $D_X = \{x_i, 1 \leq i \leq n\}$ which consists of all data points for stochastic variable X, with n subsets $D_{x_i} = \{x_i \in D_X\}$.
Similarly, multiset $D_{X,C} = \{(x_i, c_j) | 1 \leq i \leq n, 1 \leq j \leq m\}$ is a dataset consisting of all data points for the combination of variables X and C, with subsets $D_{x_i,c_j} = \{(x_i, c_j) | (x_i, c_j) \in D_{X,C}\}$

From the dataset D_X, the probabilities for X can be estimated as follows:

$$\Pr(X = x_i) = \frac{|D_{x_i}|}{|D_X|}$$

And the conditional probabilities for C:

$$\Pr(C = c_j | X = x_i) = \frac{|D_{x_i,c_j}|}{\sum_{k=1}^{m} |D_{x_i,c_k}|}$$

We illustrate the probability estimation with the following example dataset $D_{X,C} = \{(1,1),(1,1),(2,1),(2,2),(2,2),(2,2),(3,1),(3,1),(3,1),(3,2)\}$. The probabilities for X are estimated as follows:

$$\begin{aligned} \Pr(X = 1) &= \tfrac{2}{10} \\ \Pr(X = 2) &= \tfrac{4}{10} \\ \Pr(X = 3) &= \tfrac{4}{10} \end{aligned} \qquad \begin{aligned} \Pr(C = 1 | X = 1) &= \tfrac{2}{2} \\ \Pr(C = 1 | X = 2) &= \tfrac{1}{4} \\ \Pr(C = 1 | X = 3) &= \tfrac{3}{4} \\ \Pr(C = 2 | X = 1) &= 0 \\ \Pr(C = 2 | X = 2) &= \tfrac{3}{4} \\ \Pr(C = 2 | X = 3) &= \tfrac{1}{4} \end{aligned}$$

The purpose of discretization is to reduce the number of possible outcomes of a variable X. To this end, discretization methods typically divide the complete

range of values of X into so-called bins B_i, $i \in \{1,2,...,t\}$, where t is the number of bins. The bins are intervals of values of the original variable. For example, assume that values of the stochastic variable X can be any percentage. Then, X is a continuous variable. One way to get a discrete variable is to divide the values in the following bins $B_1 = [0,25)$, $B_2 = [25,50)$, $B_3 = [50,75)$ and $B_4 = [75,100]$. The original domain is cut into intervals and all values that lie in these intervals are together considered as a single value for the newly discretized variable.
We will call the points at which these intervals are cut off *cut points*. Each discretization method differs in how it is decided where these cut points are placed. These cut points can be values in the continuous domain of X. They are not necessarily data points.
The choice for the value for t is for some methods free and needs to be predetermined by the user and for other methods the value of t is determined by the method itself.

Some discretization methods determine all the cut points in one step, some methods start by determining one cut point and add new cut points step-by-step and yet other methods start by determining many cut points and then remove some of these cut points.
Regardless of the actual implementation of a discretization method, each method can be seen as a process that starts out with n bins, one for each data point, and iteratively combines two bins into a new one. Each such iteration, which we will refer to as a *discretization step*, basically creates a new variable with a new set of values.

Definition 2. A *discretization step* for a variable X maps its values $x_1 \prec ... \prec x_n$ to the values $x'_1 \prec ... \prec x'_{n-1}$ of a new variable X' such that for some $i \in \{1,...,n-1\}$ we have that:
$\forall k \in \{1,...,i-1\} : x'_k = x_k$.
$x'_i = x_i \vee x_{i+1}$
$\forall l \in \{i+1,...,n-1\} : x'_l = x_{l+1}$

For example, suppose we have a variable X with $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ and $x_5 = 5$ and suppose the values x_3 and x_4 are merged together. Then, X' consists of the values $x'_1 = 1$, $x'_2 = 2$, $x'_3 = 3 \vee 4$ and $x'_4 = 5$.

We will use this definition to study to what extent various methods can preserve characteristics of the data.

Note that our definition for a discretization step does not coincide exactly with the combining of bins. For our purpose, this definition is, however, easier to use. For each value of X' the bin B_i will then be defined such that the smallest value of the bin B_i is less or equal to the smallest value of x'_i, the greatest value of the bin B_i is greater or equal to the greatest value of x'_i, the bins do not overlap and all the bins together cover $[x'_1, x'_{n-1}]$ The values of X' are $x'_1,...,x'_{n-1}$ and for the actual implementation of the data after discretization, the corresponding bins will be used.
In this thesis, we will look at discretization in models for discrete probability distributions. These can, for example, be represented by Bayesian Networks.

However, all remarks that are made hold for other models as well. We solely use Bayesian Networks for the representation. In Bayesian networks we are mostly concerned with the connection between multiple variables. In this thesis, we will look at the effects of discretization on parts of the Bayesian network. A Bayesian network is defined in Definition 3 [8].

Definition 3. A *Bayesian network* consists of a set of variables and a set of directed edges between variables. The variables together with the directed edges form an acyclic graph.
To each variable A with parents $B_1,...,B_n$ a conditional probability table $\Pr(A|B_1,...,B_n)$ is attached.

We will focus solely on parts of the Bayesian network. We define these parts as relations

Definition 4. A *relation* $X \to C$ from stochastic variable X to stochastic variable C is associated with an arc $X \to C$ in a Bayesian network where X is a parent of C, as shown in Figure 1.

Figure 1: Relation with two variables

Definition 5. A *relation* $X \to C|Y_1,...,Y_n$ from variable stochastic X to stochastic variable C in the context of a specific value assignment of the stochastic variables $Y_1,...,Y_n$ is associated with an arc $X \to C$ in a Bayesian network where $X, Y_1,...,Y_n$ are parents of C, as shown in Figure 2

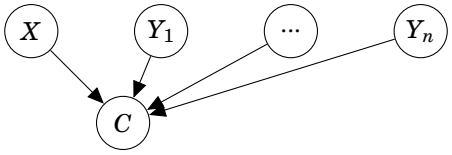


Figure 2: Relation with more than two variables

In this thesis, we will also compare distributions on differently discretized variables. For this we need a measure, and we will use the Kullback-Leibler divergence[5]

Definition 6. Consider a variable X and the set of conditional distributions $P(X)$ and $Q(X)$. Then, we define the *Kullback-Leibler divergence* between P and Q as:

$$KL(Q||P) = \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)}$$

When we are working with sets of conditional distributions for the same conditioning variable, we will use this Kullback-Leibler divergence as well. We will do this by summing over all conditioning contexts:

Definition 7. Consider two variables X and Y and the set of conditional distributions $P(Y|X)$ and $Q(Y|X)$. Then, we define the total Kullback-Leibler divergence between P and Q as:

$$KL^{\Sigma}(Q,P) = \sum_{x \in X} KL(Q_x||P_x)$$

3 Monotonicity

3.1 Monotonicity

When a relation is monotone in distribution, it is either isotone in distribution or antitone in distribution.

Definition 8. Let X and C be as before. A relation $X \rightarrow C$ is *isotone in distribution* if for every $c_k \in \{c_1, ..., c_m\}$ and every $x_i \in \{x_1, ...x_{n-1}\}$ it holds that

$$\Pr(C \leq c_k | X = x_i) \geq \Pr(C \leq c_k | X = x_{i+1})$$

Definition 9. Let X and C be as before. A relation $X \rightarrow C$ is *antitone in distribution* if for every $c_k \in \{c_1, ..., c_m\}$ and every $x_i \in \{x_1, ...x_{n-1}\}$ it holds that

$$\Pr(C \leq c_k | X = x_i) \leq \Pr(C \leq c_k | X = x_{i+1})$$

Both Definitions 8 and 9 are generalized for a relation of the form $X \rightarrow C|Y_1, ..., Y_l$ by conditioning all probabilities involved on values for $Y_1, ..., Y_l$.

3.2 Preserving monotonicity with discretization methods

By discretization of the variables in a Bayesian network, you can lose characteristics of the data, one such characteristic is (non-)monotonicity. In this section we will look at when (non-)monotonicity is preserved and when it is induced upon discretization.

3.2.1 Relations with two variables

Consider the Bayesian network $\mathscr{B}$ with variable X with values $x_1, ..., x_n$ such that $x_1 \prec ... \prec x_n$, $n > 1$ and variable C with values $c_1, ..., c_m$ such that $c_1 \prec ... \prec c_m$, $m > 1$ and the structure as shown in Figure 1.

Suppose the probabilities $\Pr(c_j|x_i)$ are estimated from a dataset $D = \{(x_i, c_j)|1 \leq i \leq n, 1 \leq j \leq m\}$, with subsets $D_{i,j} = \{(x_i, c_j)|(x_i, c_j) \in D\}$ for specific i and j, by frequency counting. That is,

$$\Pr(c_j|x_i) = \frac{|D_{i,j}|}{\sum_{k=1}^{m} |D_{i,k}|}$$

Theorem 1. *If the relation $X \to C$ is isotone in distribution and a discretization step is performed on X, where some x_i and x_{i+1} are placed in a bin together, the newly created relation $X' \to C$ will be isotone in distribution as well.*

Proof. Suppose the relation $X \to C$ is isotone in distribution. Then, by definition, for every $c_k \in C$ it holds that

$$Pr(C \leq c_k|X = x_{i-1}) \geq Pr(C \leq c_k|X = x_i) \geq Pr(C \leq c_k|X = x_{i+1}) \geq Pr(C \leq c_k|X = x_{i+2})$$

Let $\sum_{k=1}^{m} |D_{i,k}| = a$ and $\sum_{k=1}^{m} |D_{i+1,k}| = b$ with $a, b \in \mathbb{N}$. Then $P(X = x_i) = a$ and $P(X = x_{i+1}) = b$. This gives us

$$\begin{aligned}
\Pr(C \leq c_k|X = (x_i \vee x_{i+1})) &= \sum_{l=1}^{k} \Pr(C = c_l|X = (x_i \vee x_{i+1})) \\
&= \sum_{l=1}^{k} \frac{\Pr(C = c_l \wedge X = (x_i \vee x_{i+1}))}{\Pr(X = (x_i \vee x_{i+1}))} \\
&= \sum_{l=1}^{k} \frac{\Pr(C = c_l \wedge X = x_i) + \Pr(C = c_l \wedge X = x_{i+1})}{\Pr(X = x_i) + \Pr(X = x_{i+1})} \\
&= \sum_{l=1}^{k} \frac{\Pr(C = c_l|X = x_i)\Pr(X = x_i) + \Pr(C = c_l|X = x_{i+1})\Pr(X = x_{i+1})}{\Pr(X = x_i) + \Pr(X = x_{i+1})} \\
&= \frac{a \cdot \sum_{l=1}^{k} \Pr(C = c_l|X = x_i) + b \cdot \sum_{l=1}^{k} \Pr(C = c_l|X = x_{i+1})}{a + b} \\
&= \frac{a \cdot \Pr(C \leq c_k|X = x_i) + b \cdot \Pr(C \leq c_k|X = x_{i+1})}{a + b} \\
&\geq \frac{a \cdot \Pr(C \leq c_k|X = x_{i+1}) + b \cdot \Pr(C \leq c_k|X = x_{i+1})}{a + b} \\
&= \Pr(C \leq c_k|X = x_{i+1}) \\
&\geq \Pr(C \leq c_k|X = x_{i+2})
\end{aligned} \tag{1}$$

and, analogously

$$\begin{aligned}
\Pr(C \leq c_k|X = (x_i \vee x_{i+1})) &= \frac{a \cdot \Pr(C \leq c_k|X = x_i) + b \cdot \Pr(C \leq c_k|X = x_{i+1})}{a + b} \\
&\leq \frac{a \cdot \Pr(C \leq c_k|X = x_i) + b \cdot \Pr(C \leq c_k|X = x_i)}{a + b} \\
&= \Pr(C \leq c_k|X = x_i) \\
&\leq \Pr(C \leq c_k|X = x_{i-1})
\end{aligned} \tag{2}$$

Performing the discretization step on X by placing the values x_i and x_{i+1} in a bin together, Equations 1 and 2 give us:

$$Pr(C \leq c_k | X' = x'_{i-1}) \geq Pr(C \leq c_k | X' = x'_i) \geq Pr(C \leq c_k | X' = x'_{i+1})$$

For all other values of X' the monotonicity is maintained as it did for X. We conclude that the newly created relation $X' \rightarrow C$ is isotone in distribution as well. □

Theorem 2. *If the relation $X \rightarrow C$ is antitone in distribution and a discretization step is performed on X, where some x_i and x_{i+1} are placed in a bin together, the newly created relation $X' \rightarrow C$ will be antitone in distribution as well.*

Proof. Suppose the relation $X \rightarrow C$ is antitone in distribution. Thus, for every $c_k \in C$ holds

$$Pr(C \leq c_k | X = x_{i-1}) \leq Pr(C \leq c_k | X = x_i) \leq Pr(C \leq c_k | X = x_{i+1}) \leq Pr(C \leq c_k | X = x_{i+2})$$

Analogously to the proof for Theorem 1 we find

$$\begin{aligned} \Pr(C \leq c_k | X = (x_i \vee x_{i+1})) &= \sum_{l=1}^{k} \Pr(C = c_l | X = (x_i \vee x_{i+1})) \\ &= \frac{a \cdot \Pr(C \leq c_k | X = x_i) + b \cdot \Pr(C \leq c_k | X = x_{i+1})}{a+b} \\ &\leq \frac{a \cdot \Pr(C \leq c_k | X = x_{i+1}) + b \cdot \Pr(C \leq c_k | X = x_{i+1})}{a+b} \\ &= \Pr(C \leq c_k | X = x_{i+1}) \\ &\leq \Pr(C \leq c_k | X = x_{i+2}) \end{aligned} \tag{3}$$

And,

$$\begin{aligned} \Pr(C \leq c_k | X = (x_i \vee x_{i+1})) &= \frac{a \cdot \Pr(C \leq c_k | X = x_i) + b \cdot \Pr(C \leq c_k | X = x_{i+1})}{a+b} \\ &\geq \frac{a \cdot \Pr(C \leq c_k | X = x_i) + b \cdot \Pr(C \leq c_k | X = x_i)}{a+b} \\ &= \Pr(C \leq c_k | X = x_i) \\ &\geq \Pr(C \leq c_k | X = x_{i-1}) \end{aligned} \tag{4}$$

Performing the discretization step on X by placing the values x_i and X_{i+1} in a bin together, Equations 3 and 4 give us:

$$Pr(C \leq c_k | X' = x'_{i-1}) \leq Pr(C \leq c_k | X' = x'_i) \leq Pr(C \leq c_k | X' = x'_{i+1})$$

For all other values of X' the monotonicity is maintained as it did for X. We conclude that, the newly created relation $X' \rightarrow C$ is antitone in distribution as well.

□

Theorem 3. *If the relation $X \rightarrow C$ is monotone in distribution and a discretization step is performed on C, the newly created relation $X \rightarrow C'$ will be monotone in distribution as well.*

Proof. Suppose the relation $X \rightarrow C$ is isotone in distribution. Thus, for every $x_l \prec x_{l+1}$ and every $c_i \in C$, it holds that

$$\Pr(C \leq c_i | X = x_l) \geq \Pr(C \leq c_i | X = x_{l+1})$$

We know that $c_i < c_{i+1}$. Thus, from $C \leq (c_i \vee c_{i+1})$ follows that $C \leq c_{i+1}$. Therefore,

$$\begin{aligned} \Pr(C \leq (c_i \vee c_{i+1}) | X = x_l) &= \Pr(C \leq c_{i+1} | X = x_l) \\ &\geq \Pr(C \leq c_{i+1} | X = x_{l+1}) \\ &= \Pr(C \leq (c_i \vee c_{i+1}) | X = x_{l+1}) \end{aligned} \tag{5}$$

Performing the discretization step gives

$$\Pr(C' \leq c'_i | X = x_l) \geq \Pr(C' \leq c'_i | X = x_{l+1})$$

For all other values of C' the monotonicity is maintained as it did for C.We conclude that the relation $X \rightarrow C'$ is isotone in distribution as well.

If the relation $X \rightarrow C$ was antitone in distribution, the relation $X \rightarrow C'$ would be antitone in distribution as well. This can be proven in the same way as isotonicity by replacing each $\geq$ by $\leq$.

□

Combining Theorem 1, Theorem 2 and Theorem 3 gives us the following property:

Property 1. If a relation $X \rightarrow C$ is monotone in distribution, discretization of either X or C will preserve the monotonicity.

However, if the relation $X \rightarrow C$ is non-monotone in distribution, discretization of either X or C will not necessarily preserve non-monotonicity for the newly created relation. This gives us the following property:

Property 2. If a relation $X \rightarrow C$ is non-monotone in distribution, discretization of either X or C can cause monotonicity of the newly created relation.

The following two examples will show Property 2.

The first example will show that if the relation $X \rightarrow C$ is non-monotone in distribution and a discretization step is perfomed on X, the relation $X' \rightarrow C$ can be monotone in distribution.

Consider the relation $X \to C$ with $X = \{x_1, x_2, x_3, x_4\}$, $C = \{c_1, c_2\}$ and probabilities as given in Table 1. Note that the relation is not monotone.

Table 1: Probabilities of the values of C given the different values for X

	x_1	x_2	x_3	x_4
c_1	0.3	0.5	0.2	0.1
c_2	0.7	0.5	0.8	0.9

Now, perform one discretization step on X by taking $x'_1 = x_1 \vee x_2$.
Because the prior probabilities for the values of X are unknown, the probabilities $\Pr(C \leq c_1 | X' = x'_1)$ and $\Pr(C \leq c_2 | X' = x'_1)$ can not be computed. We can, however say something about how they relate to the other probabilities.

From $\Pr(C \leq c_1 | X = x_1) \leq \Pr(C \leq c_1 | X = x_2)$ and the proof for Theorem 2 it follows that

$$\Pr(C \leq c_1 | X = (x_1 \vee x_2)) \geq \Pr(C \leq c_1 | X = x_1)$$

Thus, from this and the probabilities in Table 1 it follows that after the discretization step on X, where x_1 and x_2 are placed in a bin together, we have:

$$\Pr(C \leq c_1 | X = (x_1 \vee x_2)) \geq 0.3 \geq \Pr(C \leq c_1 | X = x_3) \geq \Pr(C \leq c_1 | X = x_4)$$

After the discretization step, this gives:

$$\Pr(C \leq c_1 | X' = x'_1) \geq \Pr(C \leq c_1 | X' = x'_2) \geq \Pr(C \leq c_1 | X' = x'_3)$$

And, because $\Pr(C \leq c_2 | X = (x_1 \vee x_2)) = \Pr(C \leq c_2 | X = x_3) = \Pr(C \leq c_2 | X = x_4) = 1$, we have that, after the discretization step it holds that:

$$\Pr(C \leq c_2 | X' = x'_2) \geq \Pr(C \leq c_2 | X' = x'_2) \geq \Pr(C \leq c_2 | X' = x'_3)$$

It follows that the newly created relation $X' \to C$ is isotone in distribution. This shows that discretization of X can induce monotonicity of the newly created relation.

The second example shows that if the relation $X \to C$ is non-monotone in distribution and a discretization step is performed on C, the relation $X \to C'$ can be monotone in distribution.
Consider $X = \{x_1, x_2\}$, $C = \{c_1, c_2, c_3, c_4\}$ and probabilities as given in Table 2:

Table 2: Probabilities of the values of C given the different values of X

	x_1	x_2
c_1	0.1	0.3
c_2	0.4	0.1
c_3	0.3	0.3
c_4	0.2	0.3

This relation $X \to C$ is not monotone. Performing a discretization step on C by taking $c'_1 = c_1 \vee c_2$ will give the probabilities as shown in Table 3:

Table 3: Probabilities of the values of C given the different values of X

	x_1	x_2
c'_1	0.5	0.4
c'_2	0.3	0.3
c'_3	0.2	0.3

From these probabilities, it follows that:

$$\Pr(C \leq c'_1 | X = x_1) \geq \Pr(C \leq c'_1 | X = x_2)$$
$$\Pr(C \leq c'_2 | X = x_1) \geq \Pr(C \leq c'_2 | X = x_2)$$
$$\Pr(C \leq c'_3 | X = x_1) \geq \Pr(C \leq c'_3 | X = x_2)$$

Therefore, the newly created relation $X \to C'$ is isotone in distribution. This shows that discretization of C can induce monotonicity of the newly created relation.

In fact, if discretization makes both X and C binary-valued, then the relation $X \to C$ is necessarily monotone. This is shown in Theorem 4.

Theorem 4. *If both X and C are binary variables, the relation $X \to C$ is monotone.*

Proof. Assume that X and C are binary variables, so $X = \{x_1, x_2\}$ and $C = \{c_1, c_2\}$. This means that either $\Pr(C \leq c_1 | X = x_1) \geq \Pr(C \leq c_1 | X = x_2)$ or $\Pr(C \leq c_1 | X = x_1) \leq \Pr(C \leq c_1 | X = x_2)$.
From $\Pr(C \leq c_2 | X = x_1) = \Pr(C \leq c_2 | X = x_2) = 1$ it follows that both $\Pr(C \leq c_2 | X = x_1) \geq \Pr(C \leq c_2 | X = x_2)$ and $\Pr(C \leq c_2 | X = x_1) \leq \Pr(C \leq c_2 | X = x_2)$.
Then, if $\Pr(C \leq c_1 | X = x_1) \geq \Pr(C \leq c_1 | X = x_2)$, the network is isotone in distribution and if $\Pr(C \leq c_1 | X = x_1) \leq \Pr(C \leq c_1 | X = x_2)$, the network is antitone in distribution. □

3.2.2 Relations involving more than two variables

This section will be about relations involving more than two variables. First, consider the relation $X \rightarrow C|Y$. For this relation the following property holds:

Property 3. If the relation $X \rightarrow C|Y$ is monotone for a given value of Y, the relation $X \rightarrow C$ can be non-monotone, isotone or antitone.

The following example shows this property. Consider the relation $X \rightarrow C|Y$ with $X = \{x_1, x_2, x_3, x_4\}$, $C = \{c_1, c_2, c_3\}$, $Y = \{y_1, y_2\}$ and the probabilities for the values of C given the values for the combinations of X and Y as shown in Table 4:

Table 4: probabilities for the values of C given the combinations of values for X and Y

	y_1				y_2			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
c_1	0.9	0.8	0.3	0.2	0.2	0.4	0.6	0.7
c_2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
c_3	0.0	0.1	0.6	0.7	0.7	0.5	0.3	0.2

From these probabilities it follows that the relation $X \rightarrow C|Y = y_1$ is isotone in distribution and the relation $X \rightarrow C|Y = y_2$ is antitone in distribution. Given the probabilities for the values of Y and the probabilities from Table 4, the probabilities for the values of C given only the values of X can be calculated. Taking $\Pr(Y = y_1) = 0.5$ shows that the relation $X \rightarrow C$ can be non-monotone, as shown in Table 5:

Table 5: Probabilities of the values of C given the different values of X when $\Pr(Y = y_1) = 0.5$

	x_1	x_2	x_3	x_4
c_1	0.55	0.60	0.45	0.45
c_2	0.10	0.10	0.10	0.10
c_3	0.35	0.30	0.45	0.45

Taking $\Pr(Y = y_1) = 0.8$ shows that the relation $X \rightarrow C$ can be isotone, as shown in Table 6:

Table 6: Probabilities of the values of C given the different values of X when $\Pr(Y = y_1) = 0.8$

	x_1	x_2	x_3	x_4
c_1	0.76	0.72	0.36	0.30
c_2	0.10	0.10	0.10	0.10
c_3	0.14	0.18	0.54	0.60

And, taking $\Pr(Y = y_1) = 0.2$ shows that the relation $X \to C$ can be antitone, as shown in Table 7:

Table 7: Probabilities of the values of C given the different values of X when $\Pr(Y = y_1) = 0.2$

	x_1	x_2	x_3	x_4
c_1	0.34	0.48	0.54	0.60
c_2	0.10	0.10	0.10	0.10
c_3	0.56	0.42	0.36	0.30

From this example, we can conclude that if the relation $X \to C|Y$ is monotone, the relation $X \to C$ can be non-monotone, isotone and antitone.

From these examples, we can also deduce the next property:

Property 4. The fact that the relation $X \to C$ is isotone, does not imply that the relation $X \to C|Y$ is isotone as well. Also, the fact that the relation $X \to C$ is antitone, does not imply that the relation $X \to C|Y$ is antitone as well

This property can be found from the probabilities from Table 6. In this example, the relation $X \to C$ is isotone. However, the relation $X \to C|Y = y_2$ was antitone in distribution. In the example of Table 7 the relation $X \to C$ is antitone, while the relation $X \to C|Y = y_1$ was isotone.

The monotonicity of the relation $X \to C$ given Y could be a characteristic of the relation that should be maintained. Discretization of X or C will maintain this characteristic, but it may be lost with discretization of Y.

Corollary 1. *If the relation $X \to C|Y = y_i$ is isotone/antitone in distribution for some $y_i \in Y$, the relation $X' \to C|Y = y_i$ will also be isotone/antitone in distribution after one discretization step.*

Proof. If the relation $X \to C|Y = y_i$ is isotone in distribution, the fact that the relation $X' \to C|Y = y_i$ will also be isotone in distribution can be proven in the same way as Theorem 1 by replacing each $\Pr(C = c_j|X = x_k)$ by $\Pr(C = c_j|X = x_k \wedge Y = y_i)$.

If the relation $X \to C|Y = y_i$ is antitone in distribution, the fact that the relation $X' \to C|Y = y_i$ will also be antitone in distribution can be proven by making the same replacement in the proof for Theorem 2 □

Corollary 2. *If the relation $X \to C|Y = y_i$ is isotone/antitone in distribution for some $y_i \in Y$, the relation $X \to C'|Y = y_i$ will also be isotone/antitone in distribution after one discretization step.*

Proof. This can be proven by replacing each $\Pr(C = c_j|X = x_k)$ by $\Pr(C = c_j|X = x_k \wedge Y = y_i)$ in the proof for Theorem 3. □

Property 5. The fact that the relation $X \to C|Y = y_i$ is monotone in distribution for some $y_i \in Y$, does not imply that the relations $X \to C|Y = (y_i \vee y_{i+1})$ and $X \to C|Y = (y_{i-1} \vee y_i)$ are monotone in distribution as well.

This property is shown in the next example. In this example the relation $X \to C|Y = y_1$ is isotone in distribution and the relation $X \to C|Y = y_2$ is antitone in distribution. Assume $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2, y_3\}$, $C = \{c_1, c_2\}$ and the following probabilities for C given X and Y as shown in Table 8:

Table 8: Probabilities for the values of C given the combinations of values for X and Y

	y_1			y_2			y_3		
	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
c_1	0.7	0.5	0.1	0.2	0.3	0.9	0.2	0.7	0.6
c_2	0.3	0.5	0.9	0.8	0.7	0.1	0.8	0.3	0.4

Now, perform a discretization step on Y with $y'_1 = (y_1 \vee y_2)$. Assume that $\Pr(Y = y_1 | Y = y_1 \vee Y = y_2) = 0.5$ and $\Pr(Y = y_2 | Y = y_1 \vee Y = y_2) = 0.5$. This gives the following probabilities, as shown in Table 9:

Table 9: Probabilities for the values of C given the combinations of values for X and Y

	y'_1			y'_2		
	x_1	x_2	x_3	x_1	x_2	x_3
c_1	0.45	0.40	0.50	0.20	0.70	0.60
c_2	0.55	0.60	0.50	0.80	0.30	0.40

From the probabilities in Table 9 it follows that the relation $X \to C|Y' = y'_1$ is non-monotone. The monotonicity in the relation $X \to C|Y = y_1$ and in the relation $X \to C|Y = y_2$ is lost by discretization of Y.

3.3 Degrees of monotonicity

Now, we have seen that monotonicity is preserved when data is discretized, but how well is it preserved? To check this, we need to be able to measure some degree of monotonicity. The next example will give an idea of what we want to measure:

Let $X = \{10, 20\}$ and $C = \{c_1, c_2\}$.

Network 1:

$\Pr(C \leq c_1|X = 10) = 0.4$ $\Pr(C \leq c_1|X = 20) = 0.4$
$\Pr(C \leq c_2|X = 10) = 0.7$ $\Pr(C \leq c_2|X = 20) = 0.7$
$\Pr(C \leq c_3|X = 10) = 1$ $\Pr(C \leq c_3|X = 20) = 1$

Network 2:
$\Pr(C \leq c_1|X = 10) = 0.8$ $\Pr(C \leq c_1|X = 20) = 0.2$
$\Pr(C \leq c_2|X = 10) = 0.9$ $\Pr(C \leq c_2|X = 20) = 0.3$
$\Pr(C \leq c_3|X = 10) = 1$ $\Pr(C \leq c_3|X = 20) = 1$

To investigate the degree of monotonicity we define step size between two probabilities:

Definition 10. The *step size* between two probabilities from conditional distributions over C given consecutive values x_i and x_{i+1} of X is given by $|\Pr(C = c_j|X = x_{i+1}) - \Pr(C = c_j|X = x_i)|$ for some $c_j \in C$.

We might say that network 2 has a higher degree of monotonicity, because the step size for $C = c_1$ given consecutive values $X = 10$ and $X = 20$ is larger than it is for network 1.

3.3.1 Proposals for measuring the degree of monotonicity

In the previous example, we would, intuitively, say that network 2 has a higher degree of monotonicity. However, with other networks, it is much harder to decide which one has a higher degree of monotonicity. To decide this, we need a formal definition for the degree of monotonicity of a network.

To choose such a definition, we can look at the step sizes in the network. The first proposal is to find the greatest step size, defined by $GR(X,C)$, between these two probabilities in the network $\mathscr{B}$:

Definition 11.

$$GR(X,C) = max(abs(\Pr(C \leq c_k|X = x_{j+1}) - \Pr(C \leq c_k|X = x_j))|k \in \{1,...,m-1\}, j \in \{1,...,n-1\})$$

Another proposal is to find the average step size, defined by $AV(X,C)$, between these two probabilities in the network $\mathscr{B}$:

Definition 12.

$$AV(X,C) = \sum_{\substack{k \in \{1,...,m-1\} \\ j \in \{1,...,n-1\}}} \frac{|\Pr(C \leq c_k|X = x_{j+1}) - \Pr(C \leq c_k|X = x_j)|}{(n-1) \times (m-1)}$$

Now, we can look at how these degrees of monotonicity will change after discretization.

Theorem 5. *If $\mathscr{B}_1$ is isotone in distribution and two values $x_i, x_{i+1} \in X$ are replaced by a new value $x_{i,i+1}$ with $x_{i,i+1} = x_i \vee x_{i+1}$ and there exists a j such that, for some $k \in \{1,...,m\}$ holds*

$$GR(\mathscr{B}_1) = |\Pr(C \leq c_k | X = x_{j+1}) - \Pr(C \leq c_k | X = x_j)|$$

with $j \neq i$. Then, $GR(\mathscr{B}_2)$ of the newly created $\mathscr{B}_2$ can not be smaller than $GR(\mathscr{B}_1)$ of the original $\mathscr{B}_1$.

Proof. To show that $GR(\mathscr{B}_2)$ can not be smaller than $GR(\mathscr{B}_1)$, we have to show that $GR(\mathscr{B}_2)$ is at least as great as $GR(\mathscr{B}_1)$.

Because $\mathscr{B}_1$ is isotone, it follows that $\Pr(C \leq c_k | X = x_{j+1}) \leq \Pr(C \leq c_k | X = x_j)$ and thus

$$GR(\mathscr{B}_1) = \Pr(C \leq c_k | X = x_j) - \Pr(C \leq c_k | X = x_{j+1})$$

.

We distinguish between the following three cases: $j = i+1$, $j = i-1$ and $j \neq i+1 \wedge j \neq i-1$.

First, assume $j = i+1$. Thus,

$$GR(\mathscr{B}_1) = \Pr(C \leq c_k | X = x_{i+1}) - \Pr(C \leq c_k | X = x_{i+2})$$

From Theorem 1 it follows that $\Pr(C \leq c_k | X = x_{i,i+1}) \geq \Pr(C \leq c_k | X = x_{i+1})$. Thus,

$$GR(\mathscr{B}_1) \leq \Pr(C \leq c_k | X = x_{i,i+1}) - \Pr(C \leq c_k | X = x_{i+2})$$

From the definition of $GR(\mathscr{B})$ it follows that

$$GR(\mathscr{B}_1) \leq \Pr(C \leq c_k | X = x_{i,i+1}) - \Pr(C \leq c_k | X = x_{i+2}) \leq GR(\mathscr{B}_2)$$

Now, assume $j = i-1$. Thus,

$$GR(\mathscr{B}_1) = \Pr(C \leq c_k | X = x_{i-1}) - \Pr(C \leq c_k | X = x_i)$$

From Theorem 1 it follows that $\Pr(C \leq c_k | X = x_i) \geq \Pr(C \leq c_k | X = x_{i,i+1})$. Thus,

$$GR(\mathscr{B}_1) \leq \Pr(C \leq c_k | X = x_{i-1}) - \Pr(C \leq c_k | X = x_{i,i+1})$$

From the definition of $GR(\mathscr{B})$ it follows that

$$GR(\mathscr{B}_1) \leq \Pr(C \leq c_k | X = x_{i-1}) - \Pr(C \leq c_k | X = x_{i,i+1}) \leq GR(\mathscr{B}_2)$$

Lastly, assume $j \neq i+1 \wedge j \neq i-1$. Then, in $\mathscr{B}_2$, the values j and $j+1$ still exist and it follows directly that

$$GR(\mathscr{B}_1) = \Pr(C \leq c_k | X = x_j) - \Pr(C \leq c_k | X = x_{j+1}) \leq GR(\mathscr{B}_2)$$

□

4 Equal Frequency

Equal Frequency is a discretization method where the bins are constructed such that each bin contains (approximately) the same number of data points. The number of bins is chosen by hand.

Algorithm 1 AlgorithmEF(D_X, t)

N is the number of data points in D_X.
$D_X = \{x_1, ...x_N\}$ are the data points.
t is the number of bins in which you want to divide the data points.
Output: c_i with $i \in \{1, ..., t-1\}$ are the cut points for your bins.

1: Define each c_i as the smallest element of X with $c_i > x_k$ where $k = \lceil i \times \frac{N}{t} \rceil$
2: The first bin is given by the interval $[x_1, c_1)$
3: Each i-th bin for $i \in \{2, ..., t-1\}$ is given by the interval $[c_{i-1}, c_i)$
4: The last bin is given by the interval $[c_{t-1}, x_N]$

The following example with dataset $D_X = \{1,2,3,4,5,6,7,8,9,10\}$ will demonstrate the algorithm as shown in Algorithm 1.

1. In this example, t=3.
2. Since $t = 3$, there will be 2 cut points, c_1 and c_2. First, calculate c_1. Since $i = 1$, this will give $k = \lceil 1 \times \frac{10}{3} \rceil = 4$ and $x_4 = 4$. The smallest element of X greater than 4 is 5. Thus, $c_1 = 4$.
   In the same way we can calculate c_2. This gives $k = \lceil 2 \times \frac{10}{3} \rceil = 7$, $x_7 = 7$ and the smallest element of X greater than 7 is 8. Thus, $c_2 = 8$.
3. The first bin is $[1,4)$
4. The second bin is $[4,8)$
5. The last bin is $[8,10]$

4.1 Problems with *Equal Frequency*

4.1.1 Problem due to the number of bins

There are two problems that make it impossible to divide a dataset into bins with exactly equal frequencies.
The first is when the number of bins is not a divider of the number of data points. Look for example at the following dataset: $D_X = \{1,3,5,6,7,9,12\}$. Set $t = 2$, which is not a divider of $N = 7$. Following the algorithm, we get the following discretization:

interval	$[1,7)$	$[7,12]$
number of data points	4	3

This problem cannot be solved as the data points cannot be distributed evenly

over the bins. However, if all data points are dissimilar, following the algorithm will guarantee that the number of data points in the bins will differ by a maximum of 1. This problem can only be prevented by choosing the number of bins such that it is a divider of the number of data points.

4.1.2 Problem due to recurring data points

The second problem happens with recurring data points. When there are no recurring data points the number of data points in the bins will differ by a maximum of 1, as seen with the previous problem, there is no such guarantee with recurring data points. The following (extreme) example will demonstrate what might happen.
Take $D_X = \{1,2,3,4,5,5,5,5,5,6,7,8,9,10,11\}$. There are 15 data points. Suppose that $t = 3$. When there would be no recurring data points, dividing 15 data points into 3 bins with *Equal Frequency* would give 3 bins of size 5. This would give the following cut points for X: $D_X = \{|1,2,3,4,5|,5,5,5,5,6|,7,8,9,10,11|\}$. When data points are the same, they can never lie in different intervals. Therefore, there can not be a cut point between two 5's. Following the algorithm, this cut point is placed after the last 5. Thus, setting $t = 3$ will give the following discretization:

bin	$B_1 = [1,6)$	$B_2 = [6,7)$	$B_3 = [7,11]$
number of data points	9	1	5

It is clear that in this case, the algorithm will not have anything to do with *Equal Frequency* anymore. The data points are divided quite unevenly. This depends heavily on the number of bins. Setting $t = 4$ for the same dataset will give the following discretization:

bin	$B_1 = [1,5)$	$B_2 = [5,6)$	$B_3 = [6,9)$	$B_4 = [9,11]$
number of data points	4	5	3	3

By choosing $t = 4$ instead of $t = 3$, we see that the differences in size of the bins are much smaller.

4.2 Choosing the number of bins

Previously, we found that, because of the problem due to recurring data points, choosing the number of bins might greatly influence how evenly the data points are divided. In this section we will discuss how we can check in advance how well a chosen number of bins will divide the data points.

4.2.1 Avoiding bad cut points

Consider two datasets of the same size and with no recurring data points. If both datasets are divided into the same number of bins, the size bin B_i will be equal for both datasets for each i.
The next example shows this. Consider the following three datasets D_{Xa},D_{Xb}

and D_{Xc}. All datasets consist of 14 data points. Both D_{Xa} and D_{Xb} do not have recurring data points and D_{Xc} does have recurring data points. Suppose that all three datasets are divided into 4 bins. The vertical bars correspond with start and endpoint of a bin:

$$D_{Xa} = \{|1,2,3,4,|5,6,7,|8,9,10,11,|12,13,14|\}$$

$$D_{Xb} = \{|2,5,7,9,|13,15,20,|25,38,39,43,|45,50,51|\}$$

$$D_{Xc} = \{|1,2,3,4,4,4,|5,|6,7,8,9,|10,11,12|\}$$

We see that for each i the bin B_i has the same size for X_a and X_b, but not for X_c. If bin B_1 would have been size 4, like for X_a and X_b, the cut point would have been within the sequence of 4's. Therefore, the cut point is placed after the last 4. To avoid these inequalities in bin size, these shifted cut points should be avoided as much as possible. The following algorithm can test how many of these "bad" cut points there will be when choosing a certain number of bins, where the number of these bad cut points is denoted by z:

1. Take $Y = \{y_1, ..., y_N\}$ with $y_i = \begin{cases} 1 & \text{if i<N and } x_i = x_{i+1} \\ 0 & \text{else} \end{cases}$

2. Calculate z by $z = \sum_{i \in \{1,...,t\}} y_{\lceil i \times \frac{N}{t} \rceil}$

Using this algorithm for D_{Xc} with $t = 4$ would give:

1. $Y = \{0,0,0,1,1,0,0,0,0,0,0,0,0,0\}$
2. $z = y_4 + y_7 + y_{11} + y_{14} = 1 + 0 + 0 + 0 = 1$

Thus, choosing $t = 4$ for X_c would give 1 bad cut point.

This algorithm returns the number of bad cut points. However, some cut points are worse than others. Look, for example, at the next three datasets, which all consist of 14 data points and are all divided into 3 bins:

$$D_{Xa} = \{|1,2,3,4,5,|6,7,8,9,10,|11,12,13,14|\}$$

$$D_{Xb} = \{|1,2,3,4,5,5,|6,7,8,9,|10,11,12,13|\}$$

$$D_{Xc} = \{|1,2,3,4,5,5,5,5,5,|6,|7,8,9,10|\}$$

In X_b and X_c the cut point would have been between the first and second 5 if there were no recurring numbers. With the algorithm for both X_b and X_c this cut point would count as a bad cut point even though in the discretization of X_b the cut point is only shifted 1 place and with the discretization of X_c, it is shifted 4 places. As the discretization of X_c is more unequal than the discretization of

X_b, it seems that some cut points result in larger differences in bin sizes.

In the previous algorithm, all bad cut points were given by a 1. To also take into consideration how bad a cut point is, we will give each cut point a number based on how many places the cut point will shift. If, for example, the cut point would only have to shift 1 place it would still be given a 1, but if the cut point would have to shift 2 places it would be given a 2, etc.

1. Take $Y = \{y_1, ..., y_N\}$ with $y_i = |\{x_j \in X | j > i \wedge x_j = x_i\}|$
2. Calculate the frequency table for the values $y_{\lceil i \times \frac{N}{t} \rceil}$ with $i \in \{1, ..., t\}$.

In this table can be found how many bad cut points a given number of t will give and how bad they are.
Using this algorithm for X_c with $t = 3$ would give:

1. $Y = \{0,0,0,0,4,3,2,1,0,0,0,0,0,0\}$
2. $y_5 = 4$, $y_{10} = 0$ and $y_{14} = 0$. So the frequency table becomes:

number of cut point	0	4
frequency	2	1

Of course, a lot of datasets don't consist of integers, but of decimal numbers and there may not be as many recurring numbers. However, it may be preferred that data points that differ very little from each other end up in the same bin. In this case, both previous algorithms may be used with instead of counting a cut point as a bad cut point when two data points are exactly the same, counting them as a bad cut points when the difference between two data points is smaller than a given difference.

4.3 Advantages and disadvantages of *Equal Frequency*

4.3.1 Advantages of *Equal Frequency*

- One advantage of *Equal Frequency* is that it is quite easy to compute the discretization with this algorithm.

4.3.2 Disadvantages of *Equal Frequency*

- A disadvantage of *Equal Frequency* is that does not take the class-variable into consideration. The next example will show why this can be a disadvantage.
  Assume, we look at the relation $X \rightarrow C$, where $X = \{1,2,3,4\}$ and $C = \{1,2\}$ with dataset $D = \{(1,1),(2,1),(3,1),(4,2)\}$. Discretization on X into 2 bins with the *Equal Frequency* algorithm would give the following two bins: $B_1 = \{1,2\}$ and $B_2 = \{3,4\}$. However, in this case it would be much more

logical to place the values $1,2,3$ into the first bin and 4 in the second bin as these all have the same qualifiers.

- Another disadvantage is that the algorithm can work poorly when the data consists of many recurring data points.

5 Equal Width

In this section we consider the *Equal Width* discretization method. With *Equal Width*, the cut points are chosen such that each interval has the same length. This means that only the smallest and the largest data points have any influence on the discretization. For this discretization method, the number of bins t should be predetermined. In this chapter we will look at how *Equal Width* works and we will investigate if it is possible to choose the number of bins such that certain properties of the data are preserved. We will start by giving a short overview of the advantages and disadvantages of this discretization method. After this we will give a guideline to overcome these disadvantages. Lastly, we will investigate how well this guideline works.

Algorithm 2 AlgorithmEW(t, X)

$X = \{x_1, ..., x_N\}$ is a stochastic variable.
t is the number of bins in which you want to divide the domain of X.
1: Calculate the length of each interval: $k = \frac{x_N - x_1}{t}$
2: The first bin is given by the interval $[x_1, x_1 + k)$
3: Each i-th bin for $i \in \{2, ..., t-1\}$ is given by the interval $[x_1 + (i-1) \times k, x_1 + i \times k)$
4: The last bin is given by the interval $[x_1 + (t-1) \times k, x_N]$

We will illustrate the *Equal Width* discretization algorithm, as shown in Algorithm 2, using the example stochastic variable $X = \{1,2,3,4,5,6,7,8,9,10\}$. For this example, we choose $t = 3$.

1. The length of each interval is then given by $k = \frac{10-1}{3} = 3$
2. The first bin is given by the interval $[1,4)$
3. The second bin is given by $[1 + 1 \times 3, 1 + 2 \times 3) = [4,7)$
4. The third and last bin is given by the interval $[1 + 2 \times 3, 1 + 3 \times 3] = [7,10]$

5.1 Advantages and disadvantages of *Equal Width*

In this section we will discuss the advantages and disadvantages of *Equal Width* that we already can deduce from the algorithm.

5.1.1 Advantages of *Equal Width*

- The biggest advantage of *Equal Width* is that it is very easy to apply this method. With only knowing the smallest and largest value of X and the number of bins, all the bins can be determined.

5.1.2 Disadvantages of *Equal Width*

- *Equal Width* does not take the class-variable into consideration. We have already seen why this is a disadvantage with *Equal Frequency*. It does not even take any data points into consideration, apart from the smallest and largest value.
- The fact that *Equal Width* doesn't consider actual data points can cause empty bins (bins without data points). For example, consider the data set $D_X = \{1,2,5,6\}$ assume we choose $t = 3$. Then, the bins would become $B_1 = [1,3)$, $B_2 = [3,4)$ and $B_3 = [5,6]$. Then there are no data points that lie in the second bin. This is not a problem if these values actually don't exist in X, but it is a problem if we just did not find data points for these values.

5.2 Choosing the number of bins

In the previous section we have seen an advantage and two disadvantages of the *Equal Width* algorithm. The algorithm is quite simple, but we can see from the disadvantages, that this comes with a price. In this section we will propose a guideline that tries to minimize the effects of these disadvantages. This guideline helps in choosing the number of bins and the algorithm remains the same. Therefore, the advantage of *Equal Width* is preserved.
With *Equal Width*, the choice for the number of bins is completely free. From the literature we found that often we do not know what a good value for this number of bins is [4]. One option for choosing the number of bins in *Equal Width* could be choosing as many bins as possible with the limitation that the number of bins should not be more than desired in the specific application. This can be motivated by domain knowledge or by the observation that choosing too many bins can, for example, make the model too complex. In general, more bins results in less information loss. The disadvantages of the *Equal Width* method are that it does not take the class-variable and the actual data points into consideration. When we do want to take the class-variable into consideration we need to look at relations between two stochastic variables, as done in Definition 4. Here, we propose a guideline for choosing the number of bins for the discretization of X in the relation $X \to C$ that does take this class-variable into consideration:

Guideline 1. Choose the bins such that consecutive values of X with a similar relative frequency of occurrence for C given X in the data are placed in the same bin.

Here, similar is not yet well-defined. To measure the similarity of these relative frequencies of occurrence we choose the standard deviation of the mean

frequency of occurrence.

To illustrate the use of this guideline, we look at an example. Consider the relation $X \rightarrow C$ with $X = \{1,2,3,5,7,8,9,12\}$ and $C = \{1,2\}$. The data points in $D_{X,C}$ and their frequencies are shown in Table 10:

Table 10: frequencies for the combinations of values for X and C

	$C=1$	$C=2$
$X=1$	3	7
$X=2$	7	3
$X=3$	5	5
$X=5$	9	1
$X=7$	8	2
$X=8$	7	3
$X=9$	1	9
$X=12$	2	8

Since the guideline works with relative frequencies we need to convert the frequencies from Table 10 to relative frequencies (and therefore to estimated probabilities). This gives the relative frequencies for C given X as given in Table 11:

Table 11: Probabilities $\Pr(C|X)$ for the combinations of values for X and C

	$C=1$	$C=2$
$X=1$	0.3	0.7
$X=2$	0.7	0.3
$X=3$	0.5	0.5
$X=5$	0.9	0.1
$X=7$	0.8	0.2
$X=8$	0.7	0.3
$X=9$	0.1	0.9
$X=12$	0.2	0.8

From Table 11 we can see that, for example, the relative frequencies of occurrence given $X = 7$ and $X = 8$ are closer to each other than the relative frequencies of occurrence given $X = 8$ and $X = 9$. So, when we follow the guideline for this example, we would rather choose the bins such that $X = 7$ and $X = 8$ would be in the same bin instead of choosing the bins such that $X = 8$ and $X = 9$ would be in the same bin.

A reason to follow this guideline is to get probabilities $\Pr(C|X')$ for the newly created variable X' that are closer to the probabilities $\Pr(C|X)$ for the original value than when we would not follow this guideline.

This example shows one case where it can be useful to follow the guideline, when we want to minimize the distance between the original distribution and the distribution after discretization.

We will again look at the example as given in Table 11. Using Algorithm 2 for discretization of X with $t = 3$ as the number of bins gives the newly created variable X' with the following bins: $B_1 = [1, 4\frac{2}{3})$, $B_2 = [4\frac{2}{3}, 8\frac{1}{3})$ and $B_3 = [8\frac{1}{3}, 12]$. Using Algorithm 2 for discretization of X with $t = 4$ as the number of bins gives the newly created variable X'' with the following bins: $B_1 = [1, 3.75)$, $B_2 = [3.75, 6.5)$, $B_3 = [6.5, 9.25)$ and $B_4 = [9.25, 12]$. The standard deviations of the probabilities $\Pr(C = 1|X)$ ($C = 2$ would give the exact same standard deviations) within each bin are given in Figure 3. For example, the standard deviation for bin B_1 for X' is calculated as follows. First, the mean is calculated: $\frac{0.3+0.7+0.5}{3} = 0.5$. Then, the standard deviation becomes $\sqrt{\frac{|0.3-0.5|^2+|0.7-0.5|^2+|0.5-0.5|^2}{3}} = 0.163$

Bin	standard deviation
B_1	0.163
B_2	0.082
B_3	0.05
Average	0.098

(a) Standard deviations of the probabilities $\Pr(X|C = 1)$ in the bins corresponding with X'

Bin	standard deviation
B_1	0.163
B_2	0
B_3	0.309
B_4	0
Average	0.118

(b) Standard deviations of the probabilities $\Pr(X|C = 1)$ in the bins corresponding with X''

Figure 3: standard deviations

From Figure 3 we can see that the average standard deviation for $t = 3$ is smaller than the average standard deviations for $t = 4$. Thus, when we compare $t = 3$ and $t = 4$, the guideline would prescribe $t = 3$. We will now show the effect of choosing either $t = 3$ or $t = 4$ by comparing the distributions for both choices with the original distributions.

To compare the distribution over C given X and the distributions over C given X' we will use the Kullback-Leibler divergence. However, the Kullback-Leibler divergence compares two distributions on the same variable. Therefore, we will map the distributions Pr over C given X' to distributions Pr$'$ over C given X. Recall that, the values of X' correspond with bins and that each value of the original value X lies in one of the bins of X'. Assume that $x'_k \in X'$ corresponds with bin B_j and that $x_i \in B_j$ for some $x_i \in X$. We will map $\Pr(C|X')$ to $\Pr'(C|X)$ such that $\Pr'(C|X = x_i) = \Pr(C|X' = x'_k)$

Using this we can rewrite the distributions on X' as seen in Table 12.

Table 12: Probabilities $\Pr(C|X')$ and $\Pr'(C|X)$ given in terms of values for X and C

$\Pr(C\|X')$	$C=1$	$C=2$
$X'=B_1$	0.5	0.5
$X'=B_2$	0.8	0.2
$X'=B_3$	0.15	0.85

$\Pr'(C\|X)$	$C=1$	$C=2$
$X=1$	0.5	0.5
$X=2$	0.5	0.5
$X=3$	0.5	0.5
$X=5$	0.8	0.2
$X=7$	0.8	0.2
$X=8$	0.8	0.2
$X=9$	0.15	0.85
$X=12$	0.15	0.85

The overall distance between the distributions $\Pr(C|X)$ from Table 11 and the distributions $\Pr'(C|X)$ from Table 12 is calculated by $KL^{\Sigma}(\Pr',\Pr) = \sum_{x\in X} KL(\Pr_x||\Pr'_x)$. For example, for $X=1$ this gives:

$$D_{KL}(\overset{'}{\underset{1}{\Pr}}||\underset{1}{\Pr}) = \sum_{c\in C} \overset{'}{\underset{1}{\Pr}}(c)\log\frac{\Pr'_1(c)}{\Pr_1(c)}$$

$$= \overset{'}{\underset{1}{\Pr}}(1)\log\frac{\Pr'_1(1)}{\Pr_1(1)} + \overset{'}{\underset{1}{\Pr}}(2)\log\frac{\Pr'_1(2)}{\Pr_1(2)}$$

$$= 0.5\log\frac{0.5}{0.3} + 0.5\log\frac{0.5}{0.7} \approx 0.08718$$

The overall distance between these distributions then becomes 0.26511.
We will now compare this distance with the distance for the discretization where we use 4 bins instead of 3. The probabilities for the newly created relation $X'' \to C$ are given in Table 13. These distributions $\Pr(C|X'')$ are again mapped to distributions $\Pr''(C|X)$ the same way as above. This mapping is also given in Table 13. The distance between these distributions then becomes 1.0097.

Table 13: Probabilities $\Pr(C|X'')$ and $\Pr''(C|X)$ for the combinations of values for X and C

$\Pr(C\|X'')$	$C=1$	$C=2$
$X''=B_1$	0.5	0.5
$X''=B_2$	0.9	0.1
$X''=B_3$	$\frac{8}{15}$	$\frac{7}{15}$
$X''=B_4$	0.2	0.8

$\Pr''(C\|X)$	$C=1$	$C=2$
$X=1$	0.5	0.5
$X=2$	0.5	0.5
$X=3$	0.5	0.5
$X=5$	0.9	0.1
$X=7$	$\frac{8}{15}$	$\frac{7}{15}$
$X=8$	$\frac{8}{15}$	$\frac{7}{15}$
$X=9$	$\frac{8}{15}$	$\frac{7}{15}$
$X=12$	0.2	0.8

We see that when we choose $t = 3$ we get a new distribution that is closer, in terms of having a smaller overall KL distance, to the original distribution than when we choose $t = 4$. Thus, in this example we benefit from choosing to follow the guideline. This means that there are datasets where choosing to follow the guideline can be useful. However, this example is an artificial example that we created specifically for this purpose. Often, it will not be possible to follow this guideline. We will now look into this a bit deeper.

Two reasons why we can not always follow this guideline are:

1. The data must allow being split according to this guideline. In our example, when we chose $t = 3$, the data was divided such that the relative frequency of occurrence of values within the same bin were quite similar. It is not always possible to find a value for t such that this occurs.

2. In general, when we follow Algorithm 2, as soon as 1 cut point is chosen, the number of bins is immediately completely determined and therefore all the other cut points are determined as well. Thus, the freedom of choice is very limited, even though the guideline might suggest more freedom of choice.

To illustrate this second case, consider again the example from Table 11. Here, the guideline prescribes a cut point between 3 and 5. Choosing $t = 3$ ensures that there is a cut point between 3 and 5. However, this also automatically creates a cut point between 8 and 9. In this example, this creates only bins where the relative frequencies of occurrence are quite similar. If we would have the same example, with the exception that
$Pr(C = 1|X = 12) = 0.9$ instead of $\Pr(C = 1|X = 12) = 0.2$, the relative frequencies of occurrence within the last bin would much less similar. Thus, only changing one probability would make it impossible to follow the guideline.

We have seen that there are cases where following the guideline could give a better result. We will also investigate how this guideline handles monotonicity. From Theorem 1 and Theorem 2, we know that if the relation $X \rightarrow C$ is monotone, then, after discretization, the newly created relation $X' \rightarrow C$ will be monotone as well. Thus, whether the guideline is followed or not, monotonicity will be preserved.
We will investigate non-monotone relations where monotonicity occurs on intervals of the domain. This happens in the next example. Consider $C = \{1,2\}$ and $X = \{1,2,3,4,5,6,7,8,9,10,11,12\}$. The probabilities are given in Table 14

Table 14: Probabilities $\Pr(C|X)$ for the combinations of values for X and C

	$C=1$	$C=2$
$X=1$	0.3	0.7
$X=2$	0.5	0.5
$X=3$	0.7	0.3
$X=4$	0.6	0.4
$X=5$	0.4	0.6
$X=6$	0.2	0.8
$X=7$	0.4	0.6
$X=8$	0.8	0.2
$X=9$	0.9	0.1
$X=10$	0.7	0.3
$X=11$	0.5	0.5
$X=12$	0.1	0.9

From Table 14, we can see that the relation $X \rightarrow C$ is non-monotone, but monotone on the the following domains for X: [1,3], [4,6], [7,9] and [10,12]. We could completely preserve this property by placing the cut points between 3 and 4, 6 and 7, and 9 and 10. However, this would mean that all those intervals would then only consist of 1 value. Of course, this would that the monotonicity is preserved on these intervals, but all relations $X \rightarrow C$ where X consists of 1 value are monotone. So, it seems that this approach does not really preserve the properties of the original distribution. A consequence of this monotonicity on these intervals is that, if the probability $\Pr(C|X)$ is higher at the end of the interval than at the start of the interval, it will be higher at the start of the next interval than at the end of the next interval. (And, the other way around). To preserve this property, we should put the cut points inside these intervals, such that the last values of one interval and the first values of the next interval are put into the same bin. This, however would mean that we are putting consecutive values with a similar frequency of occurrence in the date into the same bin and therefore we are again following the guideline. So, trying to preserve non-monotonicity does not give us a better result than just using the guideline.

6 ChiMerge

ChiMerge [3] is discretization method used in classification problems where we have a relation $X \rightarrow C$. *ChiMerge* is used for the discretization of the observable variable X and the method takes the class-variable C into account, in contrast to *Equal Frequency* and *Equal Width*.
The algorithm is shown in Algorithm 3.

Algorithm 3 AlgorithmChiMerge($D_{X,C}$)

N is the number of data points.
$X = \{x_1, ..., x_n\}$ is the observable variable.
$C = \{c_1, ..., c_m\}$ is class-variable.
$D_{X,C}$ is the dataset for X and C.

1: Create bins such that each data point with a unique x-value is put into its own bin.
2: Compute the χ^2-value for each pair of adjacent bins. The χ^2-value is given by:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{m} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

where
A_{ij} = the number of data points in the i-th bin in the j-th class
R_i = the number of data points in the i-th bin
C_j = the total number of data points in the j-th class in both bins
E_{ij} = the expected frequency of A_{ij}, given by $\frac{R_i \times C_j}{R_1 + R_2}$

3: Merge the pair of adjacent bins with the lowest χ^2-value
4: Repeat step 2-3 until all χ^2-values exceed a given χ^2 - threshold

The *ChiMerge* method starts by creating bins, where each unique x-value gets its own bin. All the data points with this x-value are put into this bin. Then, the χ^2-value is calculated for each pair of adjacent bins. The pair of bins with the lowest χ^2-value is then merged. This process is repeated until all the χ^2-values exceed the threshold.
The threshold can be found from a table [3] based on the desired significance level and the number of degrees of freedom. The number of degrees of freedom is 1 less than the number of classes. The significance level is the probability that the χ^2-value is less than the threshold if X and C are independent. A higher significance level gives a higher threshold and therefore more bins are merged together.

The following example will illustrate steps 1-3 of the algorithm as shown in Algorithm 3.
Consider $X = \{1,2,3,4\}$ and $C = \{1,2,3\}$. The data points in $D_{X,C}$ and their frequencies are shown in Table 15

Table 15: frequencies for the combinations of values for X and C

	$C=1$	$C=2$	$C=3$
$X=1$	1	2	7
$X=2$	6	1	3
$X=3$	8	1	1
$X=4$	9	0	1

For step 1, we have to create bins such that each data point with a unique x-value is put into its own bin. This would give the following bins: $B_1 = [-\infty, 2)$, $B_2 = [2,3)$, $B_3 = [3,4)$, $B_4 = [4,\infty)$.
For step 2, we have to compute the χ^2-value for each pair of adjacent intervals. Here, we show the calculation for the bins B_1 and B_2:

$$\begin{aligned}\chi^2 &= \frac{(1-\frac{7\times10}{20})^2}{\frac{7\times10}{20}} + \frac{(2-\frac{3\times10}{20})^2}{\frac{3\times10}{20}} + \frac{(7-\frac{10\times10}{20})^2}{\frac{10\times10}{20}} \\ &+ \frac{(6-\frac{7\times10}{20})^2}{\frac{7\times10}{20}} + \frac{(1-\frac{3\times10}{20})^2}{\frac{3\times10}{20}} + \frac{(3-\frac{10\times10}{20})^2}{\frac{10\times10}{20}} \\ &\approx 5.50\end{aligned} \tag{6}$$

In the same way, the calculation for χ^2 for B_2 and B_3 will give $\chi^2 = 1.29$ and χ^2 for B_3 and B_4 will give $\chi^2 = 1.06$.
If we use a significance level of 0.9, the threshold will be 4.6. So, the first two bins that will be merged together are B_3 and B_4, as they have the lowest χ^2-value and this value does not exceed the threshold.

6.1 Advantages and disadvantages of *ChiMerge*

6.1.1 Advantages of *ChiMerge*

- An advantage of the *ChiMerge* method is that it does take the class-variable into consideration.

6.1.2 Disadvantages of *ChiMerge*

- A disadvantage of the *ChiMerge* method as opposed to the *Equal Frequency* and *Equal Width* methods is that the algorithm is harder to compute.
- Another disadvantage of the *ChiMerge* method is that it can not take monotonicity into consideration. When the χ^2-value is calculated, the order of the adjacent intervals is not take into account. Therefore, it would give the same value for the χ^2-value if the order of the adjacent intervals would be reversed. This makes it impossible to take monotonicity into account.

We will investigate how *ChiMerge* works for several probability distributions. Because it is hard to say something in general, we will investigate some simple distributions theoretically in this chapter. For the distributions that are more complicated to investigate we will use experiments in Chapter 8.
In this chapter we will investigate the zero influence relation and the deterministic relation.

Zero influence relation

Definition 13. The relation $X \to C$ is a *zero influence relation* if $\Pr(C = c_j|X = x_1) = \Pr(C = c_j|X = x_2) = ... = \Pr(C = c_j|X = x_n)$ for all $c_j \in C$

If $X \to C$ is a zero influence relation, the relation is isotone and antitone in distribution. To investigate what happens if we use *ChiMerge* for the discretization of X, where $X \to C$ is a zero influence relation, we first need to know what the χ^2-value of a pair of adjacent bins with the same probability distribution is.

Lemma 1. *If for two adjacent bins B_a and B_b it holds that $\Pr(C = c_j|x \in B_a) = \Pr(C = c_j|x \in B_b)$ for each $c_j \in C$, then the χ^2-value of B_a and B_b is 0.*

Proof. Assume that for two adjacent bins B_a and B_b it holds that $\Pr(C = c_j|x \in B_a) = \Pr(C = c_j|x \in B_b$ for each $c_j \in C$. For the first bin, B_a, $\Pr(C = c_j|X \in B_a)$ is calculated by $\frac{A_{1j}}{R_1}$, as seen in Algorithm 3 and for the second bin, B_b, $\Pr(C = c_j|X \in B_b)$ is calculated by $\frac{A_{2j}}{R_1}$. Since these two probabilities are equal, we have that $\frac{A_{1j}}{R_1} = \frac{A_{2j}}{R_2}$. As a consequence $R_2 \times A_{1j} = R_1 \times A_{2j}$. We use this to show that $A_{1j} = E_{1j}$ for every j:

$$\begin{aligned} A_{1j} &= \frac{A_{1j} \times (R_1 + R_2)}{R_1 + R_2} \\ &= \frac{R_1 \times A_{1j}}{R_1 + R_2} + \frac{R_2 \times A_{1j}}{R_1 + R_2} \\ &= \frac{R_1 \times A_{1j}}{R_1 + R_2} + \frac{R_1 \times A_{2j}}{R_1 + R_2} \\ &= \frac{R_1 \times (A_{1j} + A_{2j})}{R_1 + R_2} \\ &= \frac{R_1 \times C_j}{R_1 + R_2} \\ &= E_{1j} \end{aligned} \tag{7}$$

In the same way it can be shown that $A_{2j} = E_{2j}$. Thus, for every pair of adjacent bins, the χ^2-value will be equal to zero, since $A_{ij} - E_{ij} = 0$ for every i and j. □

Theorem 6. *If the relation $X \to C$ is a zero influence relation, upon discretization of X, ChiMerge will place all values of X in a single bin.*

Proof. Assume the relation $X \to C$ is a zero influence relation. With *ChiMerge*, first all data points with a unique x-value are put into their own bins. In step 2, the χ^2-value for each pair of adjacent bins is calculated. Because, $X \to C$ is a zero influence relation, it holds that $\Pr(C = c_j|x \in B_a) = \Pr(C = c_j|x \in B_b)$ for each $c_j \in C$ and every two adjacent bins B_a and B_b. From Lemma 1 it follows that the χ^2-value of each pair of adjacent bins is 0. This means that all the bins will be merged together resulting in only one bin.

□

It might seem that this is a flaw of *ChiMerge*, but if the relation $X \to C$ is a zero influence relation, it means that the value of X does not influence the value of C. Therefore, there is no relation between X and C and we would not want this relation in our model.

Deterministic relation

Definition 14. The relation $X \to C$ is a deterministic relation if for each value x_i of X, there is exactly one value of C associated with it. This means that each $\Pr(C = c_j | X = x_i)$ is either 0 or 1. In this case, we will call c_j the deterministic outcome for x_i.

Theorem 7. *If the relation $X \to C$ is a deterministic relation, upon discretization of X, ChiMerge will first merge every pair of adjacent bins that have the same deterministic outcome for C. After that, ChiMerge will merge every pair of adjacent bins with a different deterministic outcome but with a combined number of data points that is less than the threshold, as seen in Algorithm 3.*

To prove Theorem 7 we need the following Lemma:

Lemma 2. *If two adjacent bins have a different deterministic outcome, the χ^2-value for these two bins will be equal to the total number of data points in these two bins.*

Proof. Consider two adjacent bins with different deterministic outcomes. Assume without loss of generality that bin B_1 has deterministic outcome c_1 en bin B_2 has deterministic outcome c_2. Then, because the relation is deterministic, we

have $A_{11} = R_1$, $A_{22} = R_2$, $A_{12} = 0$, $A_{21} = 0$, $R_1 = C_1$ and $R_2 = C_2$.

$$
\begin{aligned}
\chi^2 &= \frac{(A_{11}-E_{11})^2}{E_{11}} + \frac{(A_{12}-E_{12})^2}{E_{12}} + \frac{(A_{21}-E_{21})^2}{E_{21}} + \frac{(A_{22}-E_{22})^2}{E_{22}} \\
&= \frac{(A_{11}-\frac{R_1\times C_1}{R_1+R_2})^2}{\frac{R_1\times C_1}{R_1+R_2}} + \frac{0-\frac{R_1\times C_2}{R_1+R_2})^2}{\frac{R_1\times C_2}{R_1+R_2}} + \frac{(0-\frac{R_2\times C_1}{R_1+R_2})^2}{\frac{R_2\times C_1}{R_1+R_2}} + \frac{(A_{22}-\frac{R_2\times C_2}{R_1+R_2})^2}{\frac{R_2\times C_2}{R_1+R_2}} \\
&= \frac{(R_1-\frac{R_1^2}{R_1+R_2})^2}{\frac{R_1^2}{R_1+R_2}} + \frac{(-\frac{R_1\times R_2}{R_1+R_2})^2}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{(-\frac{R_1\times R_2}{R_1+R_2})^2}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{(R_2-\frac{R_2^2}{R_1+R_2})^2}{\frac{R_2^2}{R_1+R_2}} \\
&= \frac{(\frac{R_1^2+R_1\times R_2-R_1^2}{R_1+R_2})^2}{\frac{R_1^2}{R_1+R_2}} + \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{(\frac{R_2^2+R_1\times R_2-R_2^2}{R_1+R_2})^2}{\frac{R_2^2}{R_1+R_2}} \\
&= \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_1^2}{R_1+R_2}} + \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_1\times R_2}{R_1+R_2}} + \frac{\frac{R_1^2\times R_2^2}{(R_1+R_2)^2}}{\frac{R_2^2}{R_1+R_2}} \\
&= \frac{R_2^2}{R_1+R_2} + \frac{R_1\times R_2}{R_1+R_2} + \frac{R_1\times R_2}{R_1+R_2} + \frac{R_1^2}{R_1+R_2} \\
&= \frac{(R_1+R_2)^2}{R_1+R_2} \\
&= R_1+R_2
\end{aligned}
\tag{8}
$$

□

Using Lemma 2 we can now prove Theorem 7.

Proof. Assume that the relation $X \to C$ is deterministic. If there exist adjacent bins that are associated with the same value of C, these bins will be merged first. This is because the χ^2-value of these bins will be 0, which follows from Lemma 1 and because *ChiMerge* merges the pair of adjacent bins with the lowest χ^2-value.

After this, if there still exists more than 1 bin, all adjacent bins will have a different deterministic outcome. From Lemma 2 it follows that the χ^2-value of each two adjacent bins will be equal to the total number of data points in these two bins. Thus, only if there exists a pair of adjacent bins with a combined number of data points that is less than the threshold, will they be merged together.

□

From Theorem 7 it follows that if the relation $X \to C$ is a deterministic relation, *ChiMerge* will merge all the pairs of adjacent bins that have the same deterministic outcome and the pairs of adjacent bins with a different deterministic outcome but with a combined number of data points that is less than the threshold. Since, the threshold is usually quite low, this will not happen often. For example, the threshold with a number of degrees of freedom of 10 and a significance level of 0.99 equals 23.21. This means that the two bins are merged

together, in this example, if the total number of data points is less than 24 and if there are 11 different outcomes for C for these two bins. This might occur with small data sets.

7 Minimum Description Length

MDLP [1] is a discretization method that, just as *ChiMerge*, takes the class-variable into account. *MDLP* is based on entropies. The entropy of a dataset S is given by

$$Ent(S) = -\sum_i \Pr(i)\log_2(\Pr(i))$$

The algorithm for *MDLP* is shown in Algorithm 4

Algorithm 4 AlgorithmMDLP($D_{X,C}$)

$X = \{x_1,...,x_n\}$ is the observable variable.
$C = \{c_1,...,c_m\}$ is the output variable.
$D_{X,C}$ is the dataset for X and C.

1: Put all the data points together in 1 bin.
2: For each possible cut point p calculate the information gain. The information gain is given by:

$$IG = Ent(S) - \frac{|S_1|}{|S|}Ent(S_1) - \frac{|S_2|}{|S|}Ent(S_2)$$

where
S is the dataset.
S_1 and S_2 are the two subsets of S that are created by cut point p.

3: Add the cut point p with the highest information gain, if it holds that

$$IG > \frac{1}{|S|}\left[\log_2(|S|-1) + \log_2(3^k-2) - (kEnt(S) - k_1Ent(S_1) - k_2Ent(S_2))\right]$$

where
k, k_1, k_2 are the number of classes in respectively S, S_1 and S_2.

4: If a cut point is added, repeat steps 2-3 recursively on the separate bins until no more cut points can be added.

The next example will illustrate steps 1-3 of the *MDLP* method. Consider $X = \{1,2,3,5,7,8,9,12\}$ and $C = \{1,2\}$. The data points in $D_{X,C}$ are shown in Table 10.
For step 1, one bin in created: $B_1 = [1,12]$.
For step 2, the information gain for each possible cut point is calculated. The first possible cut point is between 1 and 2. The information gain for this cut point is

$$IG = (-\frac{42}{80}\log_2(\frac{42}{80}) - \frac{38}{80}\log_2(\frac{38}{80}))$$
$$-\frac{10}{80}(-\frac{3}{10}\log_2(\frac{3}{10}) - \frac{7}{10}\log_2(\frac{7}{10})) - \frac{70}{80}(-\frac{39}{70}\log_2(\frac{39}{70}) - \frac{31}{70}\log_2(\frac{31}{70})) = 0.02130$$

The information gain for the other cut points are
between 2 and 3: 0.00060
between 3 and 5: 0.00108
between 5 and 7: 0.01633
between 7 and 8: 0.06466
between 8 and 9: 0.14518
between 9 and 12: 0.04588

The cut point with the highest information gain is the cut point between 8 and 9. For step 3, we have to check if this information gain is higher than the threshold. This threshold is

$$\frac{1}{80}[\log_2(79) - \log_2(3^2 - 2) - (2 \times 0.99820 - 2 \times 0.93407 - 2 \times 0.60984)] = 0.05735$$

Since the information gain for the cut point is higher than the threshold, this cut point is added and the bins become: $B_1 = [1, 8]$ and $B_2 = (8, 12]$.

7.1 Advantages of *MDLP*

- An advantage of the *MDLP* method is that it does take the class-variable into consideration.

7.2 Disadvantages of *MDLP*

- A disadvantage of the *MDLP* method as opposed to the *Equal Frequency* and *Equal Width* methods is that the algorithm is harder to compute.

We will investigate how *MDLP* works for several probability distributions.

Zero influence relation We defined a zero influence relation in Definition 13. In Chapter 6, we saw that *ChiMerge* puts all the values of X of a zero influence relation into one bin. Discretization in the context of a zero influence relation with *MDLP* gives the same result:

Theorem 8. *If the relation $X \rightarrow C$ is a zero influence relation, upon discretization of X MDLP will place all values of X in a single bin.*

To prove Theorem 8, we first need the following Lemma:

Lemma 3. *If $X \rightarrow C$ is a zero influence relation,* $\Pr(C = c_j | X \in B_i) = \Pr(C = c_j | X = x_1)$ *for each* $C_j \in C$ *and each possible bin* B_i.

Proof. Assume $X \rightarrow C$ is a zero influence relation. Then, $\Pr(C = c_j | X = x_1) = \Pr(C = c_j | X = x_2) = ... = \Pr(C = c_j | X = x_n) = a$ for all c_j and some constant $a \in [0,1]$. We will show that $\Pr(C = c_j | X \in B_i) = a$ for the specific subset $B_i = \{x_1, x_2\}$ of X:

$$
\begin{aligned}
\Pr(C = c_i | X = x_1 \vee X = x_2) &= \frac{\Pr(C = c_i \wedge (X = x_1 \vee X = x_2))}{\Pr(X = x_1 \vee X = x_2)} \\
&= \frac{\Pr(C = c_i \wedge X = x_1) + \Pr(C = c_i \wedge X = x_2)}{\Pr(X = x_1 \vee X = x_2)} \\
&= \frac{\Pr(C = c_i | X = x_1)\Pr(X = x_1) + \Pr(C = c_i | X = x_2)\Pr(X = x_2)}{\Pr(X = x_1 \vee X = x_2)} \\
&= \frac{\Pr(C = c_i | X = x_1)\Pr(X = x_1) + \Pr(C = c_i | X = x_1)\Pr(X = x_2)}{\Pr(X = x_1 \vee X = x_2)} \\
&= \frac{\Pr(C = c_i | X = x_1)(\Pr(X = x_1) + \Pr(X = x_2))}{\Pr(X = x_1 \vee X = x_2)} \\
&= \frac{\Pr(C = c_i | X = x_1)(\Pr(X = x_1 \vee X = x_2)}{\Pr(X = x_1 \vee X = x_2)} \\
&= \Pr(C = c_i | X = x_1) \\
&= a
\end{aligned}
\tag{9}
$$

In the same way it can be shown for every other subset of X.

□

Using Lemma 3, we can now prove Theorem 8:

Proof. Assume $X \rightarrow C$ is a zero influence relation. In step 1 of the *MDLP* method, all values of X are put together into one bin. In step 2, the information gain for each possible cut point p is calculated. From Lemma 3, it follows that the probabilities $\Pr(C = c_j | X \in B_i)$ for each possible bin B_i and each $c_j \in C$ are equal. Since the entropy is calculated solely from these probabilities, the entropy for each subset will be equal to the entropy of the original set. Therefore, the information

gain for each cut point will be equal to 0:

$$\begin{aligned} IG &= Ent(S) - \frac{|S_1|}{|S|}Ent(S_1) - \frac{|S_2|}{|S|}Ent(S_2) \\ &= Ent(S) - \frac{|S_1|}{|S|}Ent(S) - \frac{|S_2|}{|S|}Ent(S) \\ &= Ent(S) - \frac{|S_1|+|S_2|}{|S|}Ent(S) \\ &= Ent(S) - Ent(S) \\ &= 0 \end{aligned} \tag{10}$$

Since the information gain is zero for each cut point, no cut points will be added and thus all values of X will still be in one bin after discretization. □

As we already saw with *ChiMerge*, this is not a flaw of *MDLP*, but actually the desired result.

Deterministic relation We defined a deterministic relation and a deterministic outcome in Definition 14. For *ChiMerge*, we investigated how the method worked on a variable in a deterministic relation to the class variable. Since *ChiMerge* starts with creating bins such that each data point with a unique x-value is put into its own bin, when we first compute the χ^2-values for these bins, each pair of adjacent bins has either the same deterministic outcome or a different deterministic outcome. This made it possible to investigate what happens when we use *ChiMerge* for deterministic relation.
For *MDLP*, however, it is harder to investigate what happens with a deterministic relations. This is because *MDLP* starts by putting all the data points together into one single bin. Therefore, when we calculate the information gain for a possible cut point p, the values in a subsets that is created by the cut point p do not necessarily have the same deterministic outcome. Thus, we cannot use the deterministic characteristic to calculate the information gain. This makes the deterministic relation a less interesting special case to investigate for *MDLP*, even though it was an interesting special case for *ChiMerge*.

8 Experiments

In this chapter we will examine how *ChiMerge* and *MDLP* work for several probability distributions that differ in the extent of their monotonicity by conducting experiments. In order to investigate how *ChiMerge* and *MDLP* handle monotonicity we want to examine:

- the influence of the number of available data points. For this we define the parameter α_x as the number of times that $x \in X$ occurs in the data.
- the influence of the magnitude of the step size between probabilities, where step size is defined in Definition 10

- when the resulting bins for *ChiMerge* and *MDLP* are the same and when they are different

We will now explain why we choose to examine these three points.
When we examined the use of *MDLP* on a variable involved in a deterministic relation with the class variable, we found that two adjacent bins with a different deterministic outcome were only merged if the combined number of data points was less than the threshold. Thus, it seems that the number of data points has influence on the result of discretization. Therefore, the number of data points is one of the parameters we vary in our experiments. Also, the probability that the real relation is (non-)monotone is higher when we have more data points to base our findings on.
When we examined the *Equal Width* method we proposed a guideline such that consecutive values of X with a similar relative frequency of occurrence for C given X in the data are placed in the same bin. This happens if cut points are placed where the step size is the greatest. In all experiments we want to examine if *ChiMerge* and *MDLP* do this too.
Lastly, when the resulting bins for *ChiMerge* and *MDLP* differ, we want to compare the resulting bins in order to see if there is a reason to prefer one of these discretization methods over the other when we want do discretize a variable involved in a (non-)monotone relation.

8.1 Linear monotone relation

In this section we will examine the discretization of a variable involved in a linear monotone relation, using experiments.

Definition 15. A relation $X \to C$ is *linear monotone*, if the relation is monotone and for every c_k and x_i, $1 \leq i \leq n+1$ we have that

$$\Pr(C \leq c_k | X = x_i) - \Pr(C \leq c_k | X = x_{i+1}) = a$$

for some constant $a \in \mathbb{R}$.

Since the step size between each pair of probabilities given consecutive values is equal for a linear monotone relation, we can use this relation to examine the influence of the number of available data points on the results of discretization. To test how *ChiMerge* and *MDLP* work for a linear monotone relation we propose 7 experiments. We consider the discretization of variable X involved in the relation $X \to C$, with $X = \{1,2,3,4,5\}$ and $C = \{1,2\}$. Throughout the experiments we ensure that the probabilities $\Pr(C = c_i | X = x_j)$, $i = 1,2$, $j = 1,2,3,4,5$ are kept constant. We vary the absolute frequency of occurrence of the different value combinations for X and C upon which the probabilities are based.

8.1.1 Data & Methods

For each experiment the probability $\Pr(C = 1 | X = 1)$ is chosen to be 0.1. The step size is chosen to be 0.2 for each pair of consecutive values of X, ensuring that the

relation is linear monotone.
For our first experiment, Experiment 1.1, α_x is 10 for each $x \in X$. The frequencies and probabilities for the combinations of values for X and C for this Experiment are shown in Table 16.
For Experiment 1.2, α_x is 100 for each $x \in X$. The frequencies and probabilities for the combinations of values for X and C for this Experiment are given in Table 18.
For Experiment 1.3, the α_x for each $x \in X$ is chosen at random from the interval $[1,100]$. Then, the frequencies for the combinations of values for X and C are chosen according to the intended probability for this combination. Since frequencies are chosen to be integers the actual probabilities slightly differ from those intended. The frequencies and probabilities for Experiment 1.3 are shown in Table 20.
Experiment 1.4 until 1.7 are created in the same way as Experiment 1.3. These experiments can be found in Appendix A.

In each experiment we apply *ChiMerge* and *MDLP* to the described data. For the implementation we use R. We first use a significance level of 0.95 for *ChiMerge*. If the number of bins returned differs for *ChiMerge* and *MDLP*, we also consider another significance level for *ChiMerge* that gives the same number of bins as *MDLP*. This significance level is found by first trying a significance level of 0.99 and if this does not work, trying 0.999, then 0.9999 etc. until the significance level is found that gives the same number of bins as *MDLP*. When the results for *ChiMerge* and *MDLP* with the same number of bins are different, i.e. the bins cover different intervals of data points, we also give the Kullback-Leibler divergence between the original relation and the newly created relations, as defined in 7. If, in this case, the relation $X \rightarrow C$ is monotone, the values $GR(X,C)$ and $AV(X,C)$, capturing degrees of monotonicity, are provided, which are defined in respectively Definition 11 and Definition 12.

8.1.2 Results

We will now give the results of Experiment 1.1 until 1.3. The results for Experiment 1.4 until 1.7 can be found in Appendix A.

Table 16: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.1

	$C=1$	$C=2$
$X=1$	1	9
$X=2$	3	7
$X=3$	5	5
$X=4$	7	3
$X=5$	9	1

	$C=1$	$C=2$
$X=1$	0.1	0.9
$X=2$	0.3	0.7
$X=3$	0.5	0.5
$X=4$	0.7	0.3
$X=5$	0.9	0.1

Table 17: resulting bins for Experiment 1.1; subscript of ChiMerge indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$					
$ChiMerge_{0.9999}$					
$MDLP$					

Experiment 1.1 The resulting bins for Experiment 1.1 for the different discretization methods are shown in Table 17.
We use horizontal bars to indicated for which values of X datapoints are included in the same bin A gap thus indicates a cut point. Actual bins returned by the methods range from $infty$ rather than x_1 and to $+\infty$ rather than x_5. We see that, in Table 17 the resulting bins for the different discretization methods are shown. For example, $ChiMerge_{0.95}$ gives the two following bins: $B_1 = (-\infty, 4)$ and $B_2 = [4, \infty)$ with a cut point between 3 and 4. Both $ChiMerge_{0.9999}$ and $MDLP$ put all values in a single bin.

Table 18: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.2

	$C=1$	$C=2$
$X=1$	10	90
$X=2$	30	70
$X=3$	50	50
$X=4$	70	30
$X=5$	90	10

	$C=1$	$C=2$
$X=1$	0.1	0.9
$X=2$	0.3	0.7
$X=3$	0.5	0.5
$X=4$	0.7	0.3
$X=5$	0.9	0.1

Table 19: resulting bins for Experiment 1.2; subscript of ChiMerge indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$					
$ChiMerge_{0.9999}$					
$MDLP$					

Experiment 1.2 The resulting bins for Experiment 1.2 for the different discretization methods are shown in Table 19. We can see that $ChiMerge_{0.95}$ puts each value into its own bin. Both $ChiMerge_{0.9999}$ and $MDLP$ created three bins, but $ChiMerge_{0.9999}$ put $X = 2$ and $X = 3$, and $X = 4$ and $X = 5$ into one bin whereas $MDLP$ put $X = 1$ and $X = 2$, and $X = 3$ and $X = 4$ into one bin.

For Experiment 1.2, the difference in results between $ChiMerge_{0.9999}$ and *MDLP* can be explained, by the fact that the χ^2-value of the bins corresponding with the pair $X = 2$ and $X = 3$ is the same as the χ^2-value of the bins corresponding with the pair $X = 3$ and $X = 4$. For the result for $ChiMerge_{0.9999}$ in Table 19 the bins corresponding with $X = 2$ and $X = 3$ are chosen to be merged first, because they are considered first by the implementation of the algorithm used. If we would have merged the bins corresponding with the pair $X = 3$ and $X = 4$ first, the result would be the same as the result for *MDLP*.

Table 20: frequencies and probabilities for the combinations of values $\Pr(C|X)$ for X and C for Experiment 1.3

	$C=1$	$C=2$
$X=1$	10	85
$X=2$	8	17
$X=3$	34	34
$X=4$	46	20
$X=5$	17	2

	$C=1$	$C=2$
$X=1$	0.11	0.89
$X=2$	0.32	0.68
$X=3$	0.50	0.50
$X=4$	0.70	0.30
$X=5$	0.89	0.11

Table 21: resulting bins for Experiment 1.3; subscript of ChiMerge indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$	□	▭		▭	
$ChiMerge_{0.99999}$	□	▭			
MDLP	▭		▭		

Table 22: two degrees of monotonicity and the Kullback-Leibler divergence for Experiment 1.3

	$GR(X,C)$	$AV(X,C)$	$KL^{\Sigma}(\Pr',\Pr)$
before discretization	0.21	0.20	
$ChiMerge_{0.99999}$	0.48	0.48	0.49391
MDLP	0.48	0.48	0.36002

Experiment 1.3 The resulting bins for Experiment 1.3 are shown in Table 21. $ChiMerge_{0.95}$ created three bins whereas both $ChiMerge_{0.99999}$ and *MDLP* created two bins, albeit two different ones. To further study the difference between these latter two discretizations, we computed the degree of monotonicity and the Kullback-Leibler convergence, which are shown in Table 22. We can see that $GR(X,C)$ and $AV(X,C)$ are the same for $ChiMerge_{0.9999}$. Also, $GR(X,C)$ and

$AV(X,C)$ are the same for $ChiMerge_{0.9999}$ as they are for *MDLP*. The Kullback-Leibler convergence is greater for $ChiMerge_{0.99999}$ than it is for *MDLP*.

8.1.3 Analysis

When we compare the results from Experiment 1.1 and 1.2 we can see how the number of data points influences the discretization, since all frequencies are multiplied by 10. For Experiment 1.2, both *ChiMerge* and *MDLP* return more bins than for Experiment 1.1. It seems that a higher number of data points results in more cut points.
With Experiment 1.3, we can examine how different values of α_x influence the discretization. For Experiment 1.3, *ChiMerge* starts by merging the bins with the lowest combined number of data points and *MDLP* creates the bins such that the number of data points are as evenly distributed over the bins as possible. Analogous observations can be made for Experiment 1.4 until 1.7, which can be found in Appendix A.
In some cases *MDLP* gives different resulting bins as *ChiMerge* with a significance level that gives the same number of bins as *MDLP*. In Table 22 we can see that for Experiment 1.3 the two degrees of monotonicity are the same for $ChiMerge_{0.99999}$ and *MDLP*. The Kullback-Leibler divergence between the original distribution and the distribution resulting from *MDLP* is smaller than the Kullback-Leibler divergence between the original distribution and the distribution resulting from $ChiMerge_{0.99999}$. These results vary for Experiment 1.4 until 1.7. For Experiment 1.7, for example, the Kullback-Leibler divergence is smaller for *MDLP* as well. However, for Experiment 1.4 and 1.6, the Kullback-Leibler divergence is larger for *MDLP*. The degrees of monotonicity also vary for the different experiments. For experiment 1.4 and 1.6, both degrees of monotonicity are smaller for *MDLP*. For Experiment 1.7, both degrees of monotonicity are larger for *MDLP*. This means that we can not draw any conclusions about the difference in performance of *ChiMerge* and *MDLP* from these results, considering preserving the degree of monotonicity or the distance between the original distribution and the newly created distribution.

8.2 Monotone relation

In this section we will examine the discretization of a variable involved in a monotone relation, using experiments. Recall that a relation $X \to C$ is monotone if it is either isotone or antitone in distribution. An isotone relation is defined in Definition 8 and an antitone relation is defined in 9.

With the experiments for the linear monotone relation, we already saw that, generally, a larger number of data points results in more cut points. Since, for the monotone relation the step size between the probabilities is not necessarily equal, we can use this relation to investigate what the influence of the magnitude of the step size between the probabilities is on the discretization of X.
To test how *ChiMerge* and *MDLP* work for a monotone relation, we propose the

following experiments where we discretize a variable X in volved in the relation $X \rightarrow C$ where $X = \{1,2,3,4,5,6,7,8,9,10\}$ and $C = \{1,2\}$. The probabilities for the experiments are chosen such that there are some small steps and some big steps between the probabilities. We vary the frequency of occurrence of the different value combinations upon which the probabilities are based.

8.2.1 Data & Methods

For Experiment 2.1, α_x is 100 for each $x \in X$. The frequencies and probabilities for Experiment 2.1 are shown in Table 23. Experiment 2.5 is created in the same way as Experiment 2.1 and can be found in Appendix A.
Experiment 2.2 has the same probabilities as Experiment 2.1, but α_x for each $x \in X$ are chosen at random from the interval $[1,1000]$. The frequencies and probabilities for Experiment 2.2 are shown in Table 25.
Experiment 2.3 and 2.4 are created in the same way as Experiment 2.2 and can be found in Appendix A.
Likewise, Experiments 2.6, 2.7 and 2.8 have the same estimated probabilities as Experiment 2.5, but α_x for each $x \in X$ are chosen at random from the interval $[1,1000]$. Experiments 2.6 until 2.8 can be found in Appendix A.
We now apply *ChiMerge* and *MDLP* to the above data using the same approach as desribed in 8.1.1.

8.2.2 Results

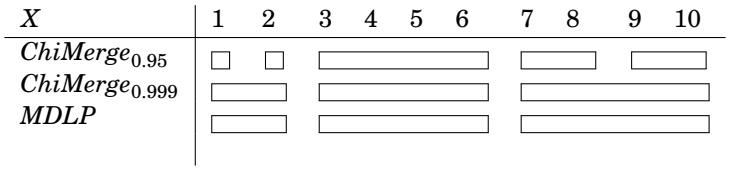
We will now give the results of Experiment 2.1 and 2.2. The results for Experiment 2.3 until 2.8 can be found in Appendix A.

Table 23: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.1

	$C=1$	$C=2$
$X=1$	5	95
$X=2$	15	85
$X=3$	40	60
$X=4$	42	58
$X=5$	45	55
$X=6$	50	50
$X=7$	80	20
$X=8$	85	15
$X=9$	90	10
$X=10$	95	5

	$C=1$	$C=2$
$X=1$	0.05	0.95
$X=2$	0.15	0.85
$X=3$	0.40	0.60
$X=4$	0.42	0.58
$X=5$	0.45	0.55
$X=6$	0.50	0.50
$X=7$	0.80	0.20
$X=8$	0.85	0.15
$X=9$	0.90	0.10
$X=10$	0.95	0.05

Table 24: resulting bins for Experiment 2.1; subscript of ChiMerge indicates significance level used



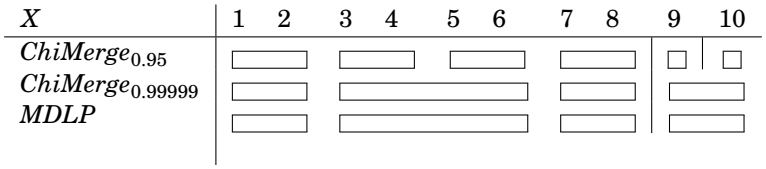
Experiment 2.1 The resulting bins for Experiment 2.1 are shown in Table 24. We can see that $ChiMerge_{0.999}$ and *MDLP* give the same resulting bins.

Table 25: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.2

	$C=1$	$C=2$
$X=1$	3	54
$X=2$	43	246
$X=3$	349	523
$X=4$	348	481
$X=5$	194	237
$X=6$	83	83
$X=7$	323	81
$X=8$	102	18
$X=9$	753	84
$X=10$	835	44

	$C=1$	$C=2$
$X=1$	0.05	0.95
$X=2$	0.15	0.85
$X=3$	0.40	0.60
$X=4$	0.42	0.58
$X=5$	0.45	0.55
$X=6$	0.50	0.50
$X=7$	0.80	0.20
$X=8$	0.85	0.15
$X=9$	0.90	0.10
$X=10$	0.95	0.05

Table 26: resulting bins for Experiment 2.2; subscript of ChiMerge indicates significance level used



Experiment 2.2 The resulting bins for Experiment 2.2 are shown in Table 26. We can see that $ChiMerge_{0.99999}$ and *MDLP* give the same resulting bins.

8.2.3 Analysis

For Experiment 2.1, the frequencies for each value of X are equal. Therefore, we can examine the influence of the step size with this experiment. We can see

in Table 23 that the two largest step sizes are between $X = 2$ and $X = 3$ and between $X = 6$ and $X = 7$. Both $ChiMerge_{0.999}$ and *MDLP* put cut points between these values. $ChiMerge_{0.95}$ puts cut points between $X = 1$ and $X = 2$, $X = 2$ and $X = 3$, $X = 6$ and $X = 7$ and between $X = 8$ and $X = 9$. These cut points include or coincide with the largest step sizes, as expected. It seems that, if the frequencies for each value of X are equal, both *ChiMerge* and *MDLP* put cut points where the step size is the largest. Analogous observations can be made for Experiment 2.5.

For Experiment 2.2, we can examine the influence of varying the frequency of occurrence for values of X. For the monotone relations, we expect that cut points will be placed between values where the step size is the biggest with both *ChiMerge* and *MDLP*. Whenever cut points are different from expected, we have indicated the expected cut points in the tables using vertical lines (see e.g. Table 26). In the cases where there is a cut point where we would not have expected it or no cut point where we would have expected it, looking solely at the step sizes between the probabilities, this can be explained by a large or a small number of data points.
For example, for Experiment 2.2, there is no cut point between $X = 1$ and $X = 2$, as we can see in Table 26. This can be explained by the relatively small number of data points for $X = 1$. It seems that, generally, both *ChiMerge* and *MDLP* put cut points where the step size is the largest. However, if two bins have a relatively small number of data points they are more likely to be merged together, and if two bins have a relatively large number of data points they are less likely to be merged together. Analogous observations can be made for Experiment 2.3, 2.4 and 2.6 until 2.8.

8.3 Non-monotonic relation

In this section we will examine the discretization of a variable involved in a non-monotone relation, using experiments.

To test how *ChiMerge* and *MDLP* work for a non-monotonic relation, we propose the following experiments where we discretize a variable X involved in the relation $X \rightarrow C$ where $C = \{1,2\}$ and $X = \{1,2,3,4,5,6,7,8,9,10\}$. For all experiments, the frequencies for the combinations of X and C are chosen at random. The probabilities are calculated from these frequencies.

8.3.1 Data & Methods

For Experiment 3.1, the frequencies for all combinations of X and C are chosen at random from the interval $[0,100]$. The frequencies and probabilities for Experiment 3.1 are shown in Table 27.
For Experiment 3.2, the frequencies are the frequencies of Experiment 3.1 multiplied by 10. The frequencies and probabilities are shown in Table 29.
Experiment 3.3 and 3.5 are created in the same way as Experiment 3.1 and can be found in Appendix A.

The frequencies for Experiment 3.4 and Experiment 3.6 are the frequencies of Experiment 3.3 and 3.5 respectively multiplied by 10. These Experiments can be found in Appendix A.
We now apply *ChiMerge* and *MDLP* to the above data using the same approach as desribed in 8.1.1.

8.3.2 Results

Table 27: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 3.1

	$C=1$	$C=2$
$X=1$	45	19
$X=2$	61	50
$X=3$	79	6
$X=4$	96	85
$X=5$	93	97
$X=6$	3	8
$X=7$	78	53
$X=8$	1	45
$X=9$	46	18
$X=10$	86	27

	$C=1$	$C=2$
$X=1$	0.70	0.30
$X=2$	0.55	0.45
$X=3$	0.93	0.07
$X=4$	0.53	0.47
$X=5$	0.49	0.51
$X=6$	0.27	0.73
$X=7$	0.60	0.40
$X=8$	0.02	0.98
$X=9$	0.72	0.28
$X=10$	0.76	0.24

Table 28: resulting bins for Experiment 3.1; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	□	□				□	□	
$ChiMerge_{0.99}$	□		□	□				□	□	
$MDLP$	□		□	□				□	□	

Experiment 3.1 The resulting bins for Experiment 3.1 are shown in Table 28. We can see that $ChiMerge_{0.99}$ and *MDLP* give the same resulting bins.

Table 29: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 3.2

	$C=1$	$C=2$
$X=1$	450	190
$X=2$	610	500
$X=3$	790	60
$X=4$	960	850
$X=5$	930	970
$X=6$	30	80
$X=7$	780	530
$X=8$	10	450
$X=9$	460	180
$X=10$	860	270

	$C=1$	$C=2$
$X=1$	0.70	0.30
$X=2$	0.55	0.45
$X=3$	0.93	0.07
$X=4$	0.53	0.47
$X=5$	0.49	0.51
$X=6$	0.27	0.73
$X=7$	0.60	0.40
$X=8$	0.02	0.98
$X=9$	0.72	0.28
$X=10$	0.76	0.24

Table 30: resulting bins for Experiment 3.2; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	□	□	□	□	□	□	□	□
$ChiMerge_{0.99}$	□	□	□	▭		□	□	□	▭	
$MDLP$	□	□	□	▭		□	□	□	▭	

Experiment 3.2 The resulting bins for Experiment 3.2 are shown in Table 30. We can see that $ChiMerge_{0.99}$ and *MDLP* give the same resulting bins. $ChiMerge_{0.95}$ puts each value of X into its own bin.

8.3.3 Analysis

For Experiment 3.1, $ChiMerge_{0.99}$ and *MDLP* put cut points where the step size is the largest. With $ChiMerge_{0.95}$, there is a cut point between $X=1$ and $X=2$, even though the step size would suggest that cut points between $X=5$ and $X=6$, and $X=6$ and $X=7$ would be chosen prior to a cut point between $X=1$ and $X=2$. This can be explained by the relatively small number of data points for $X=6$ (11). Analogous observations can be made for Experiment 3.3 and Experiment 3.5.

When we compare Experiment 3.2 with Experiment 3.1, we can examine the influence of the number of data points. The frequencies for Experiment 3.2 are the frequencies for Experiment 3.1 multiplied by 10. We notice that Experiment 3.2 results in a lot more bins for each discretization method than Experiment 3.1. It seems that, generally, more data points results in more bins. Analogous observations can be made when we compare Experiment 3.4 with Experiment 3.3 and

Experiment 3.6 with Experiment 3.5

8.4 Non-monotonic relation with obvious peaks

In this section we will examine the discretization of a variable involved in a non-monotone relation with obvious peaks, using experiments.

To test how *ChiMerge* and *MDLP* work for a non-monotonic relation with obvious peaks, we propose the following experiments in which we discretize a variable X involved in the relation $X \rightarrow C$ where $X = \{1,2,3,4,5,6,7,8,9,20\}$ and $C = \{1,2\}$. The probabilities are created such that there are some small step sizes and some large step sizes.

8.4.1 Data & Methods

For Experiment 4.1, the frequencies for each value of X are chosen at random from the interval $[1,100]$. The frequencies and probabilities for Experiment 4.1 are shown in Table 31.
For Experiment 4.2, the frequencies are the frequencies of Experiment 4.1 multiplied by 10. The frequencies and probabilities for Experiment 4.2 are shown in Table 33.
For Experiment 4.3 and 4.5, the frequencies for each value of X are chosen at random from the interval $[1,100]$. For Experiment 4.4 and 4.6, the frequencies are the frequencies of respectively Experiment 4.3 and 4.5 multiplied by 10. Experiments 4.3 until 4.6 can be found in Appendix A. We now apply *ChiMerge* and *MDLP* to the above data using the same approach as described in 8.1.1.

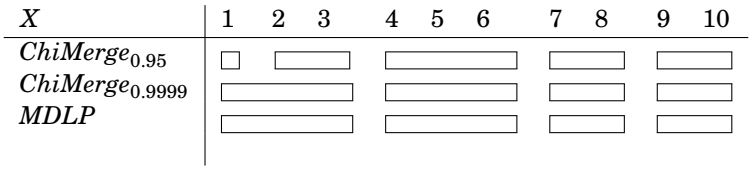
8.4.2 Results

Table 31: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.1

	$C=1$	$C=2$
$X=1$	9	78
$X=2$	5	10
$X=3$	17	66
$X=4$	7	1
$X=5$	56	3
$X=6$	9	2
$X=7$	4	16
$X=8$	4	33
$X=9$	21	20
$X=10$	26	17

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.33	0.67
$X=3$	0.20	0.80
$X=4$	0.88	0.12
$X=5$	0.95	0.05
$X=6$	0.81	0.19
$X=7$	0.20	0.80
$X=8$	0.11	0.89
$X=9$	0.51	0.49
$X=10$	0.60	0.40

Table 32: resulting bins for Experiment 4.1; subscript of *ChiMerge* indicates significance level used



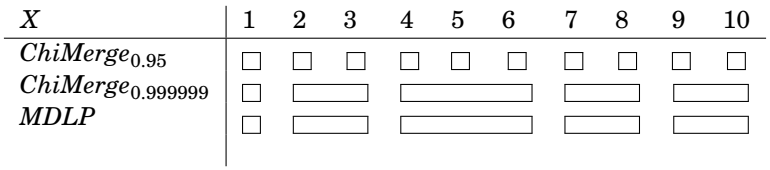
Experiment 4.1 The resulting bins for Experiment 4.1 are shown in Table 32. We can see that $ChiMerge_{0.9999}$ *and MDLP give the same resulting bins.*

Table 33: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.2

	$C=1$	$C=2$
$X=1$	90	780
$X=2$	50	100
$X=3$	170	660
$X=4$	70	10
$X=5$	560	30
$X=6$	90	20
$X=7$	40	160
$X=8$	40	330
$X=9$	210	200
$X=10$	260	170

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.33	0.67
$X=3$	0.20	0.80
$X=4$	0.88	0.12
$X=5$	0.95	0.05
$X=6$	0.81	0.19
$X=7$	0.20	0.80
$X=8$	0.11	0.89
$X=9$	0.51	0.49
$X=10$	0.60	0.40

Table 34: resulting bins for Experiment 4.2; subscript of *ChiMerge* indicates significance level used



Experiment 4.2 The resulting bins for Experiment 4.2 are shown in Table 32. We can see that $ChiMerge_{0.9999}$ and *MDLP* give the same resulting bins and that $ChiMerge_{0.95}$ puts each value of X into its own bin.

8.4.3 Analysis

For Experiment 4.1, $ChiMerge_{0.95}$, $ChiMerge_{0.9999}$ and *MDLP* all put cut points where the step size is the largest. Analogous observations can be made for Experiment 4.3. For Experiment 4.5, there is a cut point between $X = 5$ and $X = 6$, even though we would have expected a cut point between $X = 2$ and $X = 3$. This can not be explained by the number of data points. However, the step size between $X = 5$ and $X = 6$ is the second largest step size and only 0.006 smaller than the step size between $X = 2$ and $X = 3$.
When comparing Experiment 4.2 with Experiment 4.1, we can examine the influence of the number of data points, since the frequencies for Experiment 4.2 are the frequencies for Experiment 4.1 multiplied by 10. We notice that for all discretization methods, the number of bins is larger for Experiment 4.1 than for Experiment 4.1. Again, it seems that more data points results in more bins. Analogous observations can be made when comparing Experiment 4.4 with Experiment 4.3 and when comparing Experiment 4.6 with Experiment 4.5.

8.5 Analysis and discussion

With these experiments we wanted to examine how *ChiMerge* and *MDLP* handle (non-)monotone relations. In order to do this, we wanted to study:

- the influence of the number of available data points. For this we define the parameter α_x as the number of times that $x \in X$ occurs in the data.
- the influence of the magnitude of the step size between probabilities, where step size is defined in Definition 10
- when the resulting bins for *ChiMerge* and *MDLP* are the same and when they are different

To examine the influence of the number of available data points we compared experiments where each α_x for the first experiment was multiplied by 10 to create each α_x for the other experiment. In all these cases, we noticed that the experiments with a larger number of data points, for both discretization methods, always resulted in more bins.
We also examined the influence of the number of available data points by comparing experiments where α_x was constant for each $x \in X$ with experiments where we varied α_x for each $x \in X$. We noticed that adjacent bins with a relatively small combined number of data points were more likely to be merged together and that a cut point was more likely to be put between two adjacent bins with a relatively large combined number of data points.

To examine the influence of the magnitude of the step size between probabilities, we varied these step sizes. For example, with the monotone relation, we noticed that in general, both *ChiMerge* and *MDLP* put cut points where there is

a relatively large step size. When this did not happen, the unexpected cut points could be explained by the number of data points.

To examine the difference in results for *ChiMerge* and *MDLP*, we used both *ChiMerge* and *MDLP* for all experiments. To be able to make a good comparison, we used a significance level for *ChiMerge* such that the number of bins for the discretization with *ChiMerge* was the same as for the discretization with *MDLP*. For all experiments where the relation was monotone, non-monotonic or non-monotonic with obvious peaks, *ChiMerge* and *MDLP* resulted in the same bins. *ChiMerge* and *MDLP* only gave different results for the experiments where the relation was linear monotone. We could not conclude which method gave a better result in these cases. It seems that, in general, both *ChiMerge* and *MDLP* give the same results. The advantage that *ChiMerge* has over *MDLP* is that by varying the significance level you can control the number of bins. With *MDLP* the number of bins is completely determined by the algorithm.

To preserve (non-)monotonicity, it seems that there is no reason to prefer *ChiMerge* or *MDLP*. In most cases, the resulting bins were the same for both methods.

9 Cut Points

9.1 Proposals for determining cut points

Discretization methods, as we have seen previously, determine cut points. With the exception of *Equal Width*, they always use data points as the start and end point of each bin. With *Equal Frequency* for example, we have seen that the discretization of $X = \{1,2,3,4,5,6,7,8,9,10\}$ into 3 bins ($t = 3$) gave the bins $[1,5),[5,8),[8,10]$. All the start and end points of the bins (1,5,8 and 10) are points in the dataset. However, if we would have taken $[1,4.1)$, $[4.1,7.1)$, $[7.1,10]$, each corresponding bin would still contain the same number of data points. To maintain the same number of data points in each bin in this case the only requirement is that the first cut point lies between 4 and 5 and the second cut point lies between 7 and 8. It seems that the placement of this cut point is chosen arbitrarily.

Depending on the context there might be better ways to determine the cut points. In the previous example the first cut point should lie between 4 and 5, but we can choose how we want to divide the space between 4 and 5. A first proposal to divide this space between the two data points in a different way is to divide it evenly. In the example of X with *Equal Frequency* and 3 bins, this would give $[1,4.5)$, $[4.5,7.5)$, $[7.5,10]$. The cut points are placed exactly between 4 and 5 and between 7 and 8.

The first proposal might seem fair for the previous example, but this could be different when there is more variation in the differences between the data points.

Consider $Y = \{1, 5, 100, 200\}$. Here, the differences between the data points are respectively 4, 985 and 100, while in the previous example all the differences were 1. The discretization of Y with *Equal Frequency*, $t = 2$ and dividing the space between the data points evenly as per the previous proposal would give $[1, 49)$, $[49, 200]$, instead of $[1, 100)$, $[100, 200]$ as per the algorithm for *Equal Frequency*. In some applications it might seem unfair that the space between 5 and 100 is divided evenly between the bins even though the difference between 1 and 5 is much smaller than the difference between 100 and 200. Another proposal it so divide the space proportionate to the difference between the data points. As we want to achieve *Equal Frequency*, the first bin should contain the data points 1 and 5 and the second bin should contain 100 and 200. The space that has to be divided between the bins is 100-5=95. The difference between the first two data points is $5 - 1 = 4$ and the difference between the last two data points is $200 - 100 = 100$. If we divide the space to ratio of the difference between the data points, the cut points will be $5 + \frac{4}{104} \times 98 = 8.77$. The discretization will then become $[1, 8.77)$, $[8.77, 200]$.

The problem with these proposals is that they can not be justified, because there is no data in this open space. When deciding which proposal to use, the specific application should be taken into consideration. Domain knowledge can be very useful in this decision.

10 Conclusions & Further research

In this thesis we set out to examine the relation between discretization and monotonicity. For relations involving two variables, we found that monotonicity is preserved upon discretization, independent of which discretization method we use. We also found that discretization can induce monotonicity when the relation was non-monotone prior to discretization. When we examined relations involving more than two variables we found that we could not make many predictions about what would happen with monotonicity.
Monotonicity is a relation between two (or more) variables. Both *Equal Frequency* and *Equal Width* do not take any other variable into consideration than the variable under discretization. This means that the only statements about *Equal Frequency* and *Equal Width* in relation to monotonicity are the statements that we already found when we examined relations involving two variables. For both methods, the number of bins should be predetermined. We proposed guidelines on how to choose the number of bins.
Both *ChiMerge* and *MDLP* do take the another variable into consideration. We investigated some special cases for these methods to find out how they handle (non-)monotonicity for different probability distributions. From this, we found two parameters that seemed to influence the resulting bins for *ChiMerge* and *MDLP*. Namely, the number of data points and the step size between probabilities. Subsequently we performed experiments in which we varied these parameters to examine how *ChiMerge* and *MDLP* handle different probability distribu-

tions. We found that adjacent bins with a relatively small combined number of data points were more likely to be merged and that a cut point was more likely to be put between two adjacent bins with a relatively large combined number of data points. We also found that, in general, both *ChiMerge* and *MDLP* put cut points where there is a relatively large step size. These experiments were also used to compare the results for *ChiMerge* and *MDLP* and we found that in almost all cases *ChiMerge* and *MDLP* give the same result.

In this thesis we limited our research to artificial examples. For future research it will be interesting to look at real data. Moreover, in the experiments we only considered relations with binary output variables C. This made it easier to examine the influence of the step size. However, in real applications the output variable can have more than two values. It will be interesting to examine these relations to see if the results we found with the experiments also hold for relations where the output variable has more than two values. Finally, in this thesis we proposed two measures for a degree of monotonicity. We could test these degrees with artificial and real data. Preferably, the degrees would also work for non-monotone relations. If this is not the case for our degrees, we could try to find another degree that would work for both monotone and non-monote relations.

References

[1] Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1022–1027.

[2] M Julia Flores, José A Gámez, Ana M Martínez, and José M Puerta. Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter? *Applied Intelligence*, 34(3):372–385, 2011.

[3] Randy Kerber. Chimerge: Discretization of numeric attributes. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 123–128. AAAI Press, 1992.

[4] Sotiris Kotsiantis and Dimitris Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[5] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[6] Johan Kwisthout. The computational complexity of monotonicity in probabilistic networks. In E. Csuhaj-Varj and Z. Esik, editors, *Proceedings of the Sixteenth International Symposium on Fundamentals of Computation Theory*, volume 4639 of *LNCS*, pages 388–399. Springer-Verlag, 2007.

[7] Johan Kwisthout, Hans L. Bodlaender, and Gerard Tel. Local monotonicity in probabilistic networks. In K. Mellouli, editor, *Proceedings of the Ninth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 4724 of *LNCS*, pages 548–559. Springer-Verlag, 2007.

[8] Thomas D. Nielsen and Finn V. Jensen. *Bayesian Networks and Decision Graphs*. Springer Science & Business Media, 2009.

[9] Linda C. van der Gaag, Hans L. Bodlaender, and Ad Feelders. Monotonicity in Bayesian networks. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 569–576. AUAI Press, 2004.

[10] Linda C. van der Gaag, Hermina J.M. Tabachneck-Schijf, and Petra L Geenen. Verifying monotonicity of Bayesian networks with domain experts. *International Journal of Approximate Reasoning*, 50(3):429–436, 2009.

Appendices

A Experiments

Experiment 1.4

Table 35: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.4

	$C=1$	$C=2$
$X=1$	1	13
$X=2$	9	21
$X=3$	41	40
$X=4$	63	27
$X=5$	58	6

	$C=1$	$C=2$
$X=1$	0.07	0.93
$X=2$	0.30	0.70
$X=3$	0.51	0.49
$X=4$	0.70	0.30
$X=5$	0.91	0.09

Table 36: resulting bins for Experiment 1.4; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$					
$ChiMerge_{0.99999}$					
$MDLP$					

Table 37: two degrees of monotonicity and the Kullback-Leibler divergence for Experiment 1.4

	$GR(X,C)$	$AV(X,C)$	$KL^{\Sigma}(\Pr',\Pr)$
before discretization	0.22	0.21	
$ChiMerge_{0.99999}$	0.46	0.46	0.37746
MDLP	0.38	0.38	0.54206

Experiment 1.5

Table 38: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.5

	$C=1$	$C=2$
$X=1$	2	15
$X=2$	3	7
$X=3$	34	34
$X=4$	15	6
$X=5$	16	2

	$C=1$	$C=2$
$X=1$	0.12	0.88
$X=2$	0.30	0.70
$X=3$	0.50	0.50
$X=4$	0.71	0.29
$X=5$	0.89	0.11

Table 39: resulting bins for Experiment 1.5; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$					
$ChiMerge_{0.999}$					
MDLP					

Experiment 1.6

Table 40: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.6

	$C=1$	$C=2$
$X=1$	6	55
$X=2$	21	50
$X=3$	22	22
$X=4$	36	16
$X=5$	9	1

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.30	0.70
$X=3$	0.50	0.50
$X=4$	0.69	0.31
$X=5$	0.90	0.10

Table 41: resulting bins for Experiment 1.6; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$	□	□	□	▭	
$ChiMerge_{0.999}$	▭			▭	
$MDLP$	▭		▭		

Table 42: two degrees of monotonicity and the Kullback-Leibler divergence for Experiment 1.6

	$GR(X,C)$	$AV(X,C)$	$KL^{\Sigma}(\Pr',\Pr)$
before discretization	0.21	0.20	
$ChiMerge_{0.999}$	0.45	0.45	0.36952
$MDLP$	0.44	0.44	0.37073

Experiment 1.7

Table 43: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 1.7

	$C=1$	$C=2$
$X=1$	9	85
$X=2$	1	3
$X=3$	39	39
$X=4$	21	9
$X=5$	15	2

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.25	0.75
$X=3$	0.50	0.50
$X=4$	0.70	0.30
$X=5$	0.88	0.12

Table 44: resulting bins for Experiment 1.7; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5
$ChiMerge_{0.95}$	□	▭		▭	
$ChiMerge_{0.999}$	□	▭			
$MDLP$	▭		▭		

Table 45: two degrees of monotonicity and the Kullback-Leibler divergence for Experiment 1.7

	$GR(X,C)$	$AV(X,C)$	$KL^{\Sigma}(\Pr',\Pr)$
before discretization	0.25	0.20	
$ChiMerge_{0.999}$	0.49	0.49	0.57037
MDLP	0.50	0.50	0.36697

Experiment 2.3

Table 46: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.3

	$C=1$	$C=2$
$X=1$	39	741
$X=2$	52	295
$X=3$	236	354
$X=4$	420	579
$X=5$	437	535
$X=6$	43	42
$X=7$	714	179
$X=8$	66	12
$X=9$	346	38
$X=10$	411	22

	$C=1$	$C=2$
$X=1$	0.05	0.95
$X=2$	0.15	0.85
$X=3$	0.40	0.60
$X=4$	0.42	0.58
$X=5$	0.45	0.55
$X=6$	0.51	0.49
$X=7$	0.80	0.20
$X=8$	0.85	0.15
$X=9$	0.90	0.10
$X=10$	0.95	0.05

Table 47: resulting bins for Experiment 2.3; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	▭		▭		▭		□	□
$ChiMerge_{0.999}$	□	□	▭				▭		▭	
MDLP	□	□	▭				▭		▭	

Experiment 2.4

Table 48: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.4

	$C=1$	$C=2$
$X=1$	45	862
$X=2$	36	206
$X=3$	56	85
$X=4$	293	405
$X=5$	223	272
$X=6$	282	281
$X=7$	5	1
$X=8$	343	60
$X=9$	270	30
$X=10$	832	44

	$C=1$	$C=2$
$X=1$	0.05	0.95
$X=2$	0.15	0.85
$X=3$	0.40	0.60
$X=4$	0.42	0.58
$X=5$	0.45	0.55
$X=6$	0.51	0.49
$X=7$	0.83	0.17
$X=8$	0.85	0.15
$X=9$	0.90	0.10
$X=10$	0.95	0.05

Table 49: resulting bins for Experiment 2.4; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	▭	▭	▭	□	▭	▭	▭	□
$ChiMerge_{0.999}$	□	□	▭	▭	▭	▭	▭	▭	▭	□
MDLP	□	□	▭	▭	▭	▭	▭	▭	▭	□

Experiment 2.5

Table 50: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.5

	$C=1$	$C=2$
$X=1$	10	90
$X=2$	12	88
$X=3$	15	85
$X=4$	20	80
$X=5$	55	45
$X=6$	60	40
$X=7$	63	37
$X=8$	80	20
$X=9$	85	15
$X=10$	90	10

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.12	0.88
$X=3$	0.15	0.85
$X=4$	0.20	0.80
$X=5$	0.55	0.45
$X=6$	0.60	0.40
$X=7$	0.63	0.37
$X=8$	0.80	0.20
$X=9$	0.85	0.15
$X=10$	0.90	0.10

Table 51: resulting bins for Experiment 2.5; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$										
$MDLP$										

Experiment 2.6

Table 52: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.6

	$C=1$	$C=2$
$X=1$	82	740
$X=2$	117	856
$X=3$	75	422
$X=4$	118	474
$X=5$	21	18
$X=6$	405	270
$X=7$	293	87
$X=8$	484	121
$X=9$	247	43
$X=10$	837	93

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.12	0.88
$X=3$	0.15	0.85
$X=4$	0.20	0.80
$X=5$	0.54	0.46
$X=6$	0.60	0.40
$X=7$	0.63	0.37
$X=8$	0.80	0.20
$X=9$	0.85	0.15
$X=10$	0.90	0.10

Table 53: resulting bins for Experiment 2.6; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$										
$ChiMerge_{0.99}$										
$MDLP$										

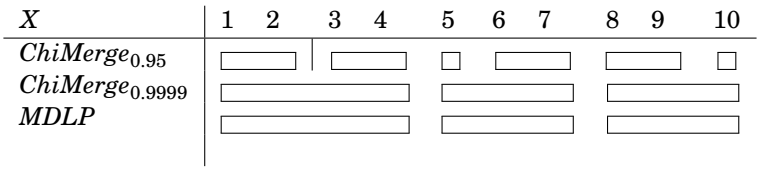
Experiment 2.7

Table 54: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.7

	$C=1$	$C=2$
$X=1$	51	459
$X=2$	90	662
$X=3$	100	567
$X=4$	57	228
$X=5$	515	422
$X=6$	95	64
$X=7$	325	192
$X=8$	218	54
$X=9$	97	17
$X=10$	338	37

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.12	0.88
$X=3$	0.15	0.85
$X=4$	0.20	0.80
$X=5$	0.55	0.45
$X=6$	0.60	0.40
$X=7$	0.63	0.37
$X=8$	0.80	0.20
$X=9$	0.85	0.15
$X=10$	0.90	0.10

Table 55: resulting bins for Experiment 2.7; subscript of *ChiMerge* indicates significance level used



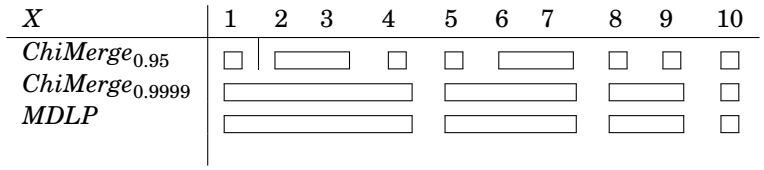
Experiment 2.8

Table 56: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 2.8

	$C=1$	$C=2$
$X=1$	94	846
$X=2$	91	665
$X=3$	44	252
$X=4$	50	199
$X=5$	489	400
$X=6$	395	264
$X=7$	72	43
$X=8$	793	198
$X=9$	715	126
$X=10$	429	48

	$C=1$	$C=2$
$X=1$	0.10	0.90
$X=2$	0.12	0.88
$X=3$	0.15	0.85
$X=4$	0.20	0.80
$X=5$	0.55	0.45
$X=6$	0.60	0.40
$X=7$	0.63	0.37
$X=8$	0.80	0.20
$X=9$	0.85	0.15
$X=10$	0.90	0.10

Table 57: resulting bins for Experiment 2.8; subscript of *ChiMerge* indicates significance level used



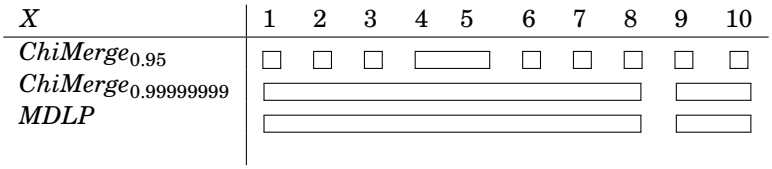
Experiment 3.3

Table 58: frequencies and probabilities Pr($C|X$) for the combinations of values for X and C for Experiment 3.3

	$C=1$	$C=2$
$X=1$	71	75
$X=2$	7	45
$X=3$	25	55
$X=4$	50	21
$X=5$	93	52
$X=6$	20	56
$X=7$	64	48
$X=8$	27	53
$X=9$	59	10
$X=10$	98	1

	$C=1$	$C=2$
$X=1$	0.49	0.51
$X=2$	0.13	0.87
$X=3$	0.31	0.69
$X=4$	0.70	0.30
$X=5$	0.64	0.36
$X=6$	0.26	0.74
$X=7$	0.57	0.43
$X=8$	0.34	0.66
$X=9$	0.86	0.14
$X=10$	0.99	0.01

Table 59: resulting bins for Experiment 3.3; subscript of *ChiMerge* indicates significance level used



Experiment 3.4

Table 60: frequencies and probabilities Pr($C|X$) for the combinations of values for X and C for Experiment 3.4

	$C=1$	$C=2$
$X=1$	710	750
$X=2$	70	450
$X=3$	250	550
$X=4$	500	210
$X=5$	930	520
$X=6$	200	560
$X=7$	640	480
$X=8$	270	530
$X=9$	590	100
$X=10$	980	10

	$C=1$	$C=2$
$X=1$	0.49	0.51
$X=2$	0.13	0.87
$X=3$	0.31	0.69
$X=4$	0.70	0.30
$X=5$	0.64	0.36
$X=6$	0.26	0.74
$X=7$	0.57	0.43
$X=8$	0.34	0.66
$X=9$	0.86	0.14
$X=10$	0.99	0.01

Table 61: resulting bins for Experiment 3.4; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	□	□	□	□	□	□	□	□
$ChiMerge_{0.999}$	□	□	□	▭		□	□	□	□	□
MDLP	□	□	□	▭		□	□	□	□	□

Experiment 3.5

Table 62: frequencies and probabilities Pr($C|X$) for the combinations of values for X and C for Experiment 3.5

	$C=1$	$C=2$
$X=1$	19	71
$X=2$	46	2
$X=3$	92	94
$X=4$	76	79
$X=5$	99	12
$X=6$	76	16
$X=7$	60	12
$X=8$	9	34
$X=9$	41	17
$X=10$	62	71

	$C=1$	$C=2$
$X=1$	0.21	0.79
$X=2$	0.96	0.04
$X=3$	0.49	0.51
$X=4$	0.49	0.51
$X=5$	0.89	0.11
$X=6$	0.83	0.17
$X=7$	0.83	0.17
$X=8$	0.21	0.79
$X=9$	0.71	0.29
$X=10$	0.47	0.53

Table 63: resulting bins for Experiment 3.5; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	▭		▭			□	□	□
$ChiMerge_{0.99999}$	□	□	▭		▭			▭		
$MDLP$	□	□	▭		▭			▭		

Experiment 3.6

Table 64: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 3.6

	$C=1$	$C=2$
$X=1$	190	710
$X=2$	460	20
$X=3$	920	940
$X=4$	760	790
$X=5$	990	120
$X=6$	760	160
$X=7$	600	120
$X=8$	90	340
$X=9$	410	170
$X=10$	620	710

	$C=1$	$C=2$
$X=1$	0.21	0.79
$X=2$	0.96	0.04
$X=3$	0.49	0.51
$X=4$	0.49	0.51
$X=5$	0.89	0.11
$X=6$	0.83	0.17
$X=7$	0.83	0.17
$X=8$	0.21	0.79
$X=9$	0.71	0.29
$X=10$	0.47	0.53

Table 65: resulting bins for Experiment 3.6; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	□	□	▭		□	▭		□	□	□
$MDLP$	□	□	▭		□	▭		□	□	□

Experiment 4.3

Table 66: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.3

	$C=1$	$C=2$
$X=1$	22	2
$X=2$	7	1
$X=3$	83	11
$X=4$	27	7
$X=5$	4	8
$X=6$	11	11
$X=7$	53	44
$X=8$	3	24
$X=9$	1	19
$X=10$	7	15

	$C=1$	$C=2$
$X=1$	0.92	0.08
$X=2$	0.88	0.12
$X=3$	0.88	0.12
$X=4$	0.80	0.20
$X=5$	0.33	0.67
$X=6$	0.50	0.50
$X=7$	0.55	0.45
$X=8$	0.11	0.89
$X=9$	0.05	0.95
$X=10$	0.32	0.68

Table 67: resulting bins for Experiment 4.3; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	[	–	–	]	[	–	]	[	]	[]
$ChiMerge_{0.99}$	[	–	–	]	[	–	]	[	–	]
MDLP	[	–	–	]	[	–	]	[	–	]

Experiment 4.4

Table 68: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.4

	$C=1$	$C=2$
$X=1$	220	20
$X=2$	70	10
$X=3$	830	110
$X=4$	270	70
$X=5$	40	80
$X=6$	110	110
$X=7$	530	440
$X=8$	30	240
$X=9$	10	190
$X=10$	70	150

	$C=1$	$C=2$
$X=1$	0.92	0.08
$X=2$	0.88	0.12
$X=3$	0.88	0.12
$X=4$	0.80	0.20
$X=5$	0.33	0.67
$X=6$	0.50	0.50
$X=7$	0.55	0.45
$X=8$	0.11	0.89
$X=9$	0.05	0.95
$X=10$	0.32	0.68

Table 69: resulting bins for Experiment 4.4; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	[	]		[ ]	[ ]	[	]	[ ]	[ ]	[ ]
$ChiMerge_{0.999999}$	[		]		[		]	[	]	[ ]
$MDLP$	[		]		[		]	[	]	[ ]

Experiment 4.5

Table 70: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.5

	$C=1$	$C=2$
$X=1$	18	41
$X=2$	12	34
$X=3$	34	40
$X=4$	32	37
$X=5$	5	5
$X=6$	27	12
$X=7$	4	1
$X=8$	18	4
$X=9$	32	13
$X=10$	52	27

	$C=1$	$C=2$
$X=1$	0.31	0.69
$X=2$	0.26	0.74
$X=3$	0.46	0.54
$X=4$	0.46	0.54
$X=5$	0.50	0.50
$X=6$	0.69	0.31
$X=7$	0.80	0.20
$X=8$	0.82	0.18
$X=9$	0.71	0.29
$X=10$	0.66	0.34

Table 71: resulting bins for Experiment 4.5; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	[	]		[	]	[				]
$ChiMerge_{0.999999}$	[				]	[				]
$MDLP$	[				]	[				]

Experiment 4.6

Table 72: frequencies and probabilities $\Pr(C|X)$ for the combinations of values for X and C for Experiment 4.6

	$C=1$	$C=2$
$X=1$	180	410
$X=2$	120	340
$X=3$	340	400
$X=4$	320	370
$X=5$	50	50
$X=6$	270	120
$X=7$	40	10
$X=8$	180	40
$X=9$	320	130
$X=10$	520	270

	$C=1$	$C=2$
$X=1$	0.31	0.69
$X=2$	0.26	0.74
$X=3$	0.46	0.54
$X=4$	0.46	0.54
$X=5$	0.50	0.50
$X=6$	0.69	0.31
$X=7$	0.80	0.20
$X=8$	0.82	0.18
$X=9$	0.71	0.29
$X=10$	0.66	0.34

Table 73: resulting bins for Experiment 4.6; subscript of *ChiMerge* indicates significance level used

X	1	2	3	4	5	6	7	8	9	10
$ChiMerge_{0.95}$	[	]	[		]	[	]	[	]	
$ChiMerge_{0.9999}$	[	]	[		]	[				]
$MDLP$	[	]	[		]	[				]