Utrecht University

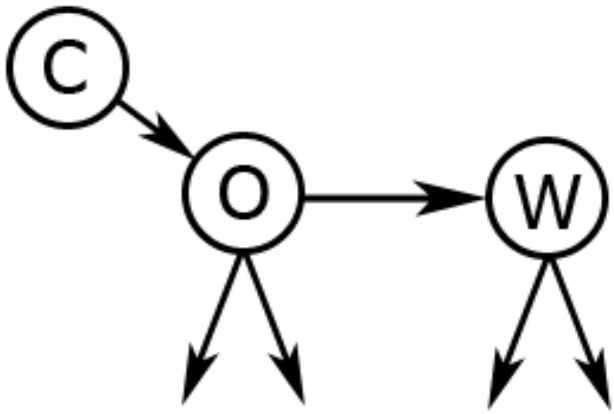
Department of Information and Computing Sciences

Discretization of continuous-valued feature variables in naive Bayesian networks

Author:
Roel Bertens

ICA-3037398

Supervisors:
prof. dr. ir. L.C. van der Gaag
and
dr. S. Renooij



June 2011

Abstract

In Bayesian networks all variables should be discrete. Therefore all continuous-valued domain variables should be discretized when building a Bayesian network. To this end, several methods with different input parameters exist, which can all have a considerable impact on the posterior distributions computed from a network.

The consequences for the posterior distribution of the class variable in a naive Bayesian network, of changing the discretization for a feature variable, are presented in this thesis. Using the concepts from sensitivity analyses we will determine some conditions that need to hold in order to change the behavior of a naive Bayesian classifier as a result of a different discretization for a feature variable. These conditions comprise bounds for the values of the prior class variable distribution and they constrain the relationship between the parameters from the conditional probability table of the variable that is to be discretized. It is useful to know for which variables discretization can change the classifier's behavior in order to identify possibilities to improve the model.

Contents

1 Introduction

A Bayesian network is a representation of a joint probability distribution on a set of stochastic variables. A more constrained Bayesian network consists of a set of feature variables and a single class variable, where all feature variables only relate directly to the class variable; this is called a naive Bayesian network. In Bayesian networks all variables should be discrete. Therefore all continuous-valued domain variables should be discretized when building a Bayesian network. To this end, several methods exist, each requiring as input the number of intervals into which to split the continuous domain, and possibly also the boundaries of these intervals. This discretization process is not evident, because there are many different choices and each might have a considerable impact on the posterior distributions computed from a network. In this thesis, we will investigate the consequences for the posterior distribution of the class variable in a naive Bayesian network, of changing the discretization for a feature variable.

When building a Bayesian network, we need to specify the network parameters, which consist of (conditional) probability distributions for all variables. There are two ways to obtain these parameters, either by estimation of domain experts or by learning them from collected data. We know that changing the discretization for a feature variable can influence the parameters in its conditional probability table. E.g. suppose that we have the (naive) Bayesian network from Figure 1 containing a feature variable E and a class variable C. We assume that in practice the class variable C is binary, with values $c \in \{1, 2\}$ and the feature variable E is continuous-valued, with values $e \in \mathbb{N}$. Further suppose, that we learn our parameters from the following instances:

c	1	1	2	1	1	2	1	2	2	2	2
e	1	3	4	9	13	14	17	19	20	21	23

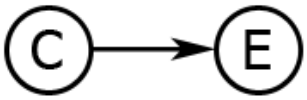


Figure 1: A (naive) Bayesian network with class variable C and feature variable E.

When we discretize our feature variable by splitting its domain into the two intervals; $e < 10$ and $e \geq 10$, we can compute the conditional probability table. E.g. $\Pr(e < 10 \mid c = 1) = 3/5$. The complete table looks as follows:

	1	2
< 10	$3/5$	$1/6$
≥ 10	$2/5$	$5/6$

When we move our threshold from 10 to 15, i.e. we split the domains of E into $e < 15$ and $e \geq 15$, we get different values:

	1	2
< 15	$4/5$	$2/6$
≥ 15	$1/5$	$4/6$

We conclude that a different discretization can lead to a different network and consequently to different values for posterior probabilities of interest upon inference.

To investigate the influence of different discretizations we make use of concepts from sensitivity analyses. Sensitivity analysis is a general technique to study the variation in the output of a mathematical model while varying the input parameters. For a Bayesian network, more specifically, a sensitivity analysis can be used to investigate the impact of changing network parameters on a posterior probability of interest, which is exactly what we are after. To study the influence of different discretizations in a realistic setting, we made use of real data from a dairy farm. Research has been done in order to predict clinical mastitis, which is an udder infection in cows, on farms with an automated milking system. An automated milking system can measure specific characteristics of the milk and decide to issue an alert after a cow has been milked. In [8] a prediction model, in the form of a naive Bayesian network, is built on information about the cows, which consists of variables such as parity, days in milk and somatic cell count; the model tries to predict whether an alert from the automated milking system is a true-positive alert or a false-positive one.

The results from our research will be helpful in understanding the effects and possibilities of applying different discretizations on continuous-valued variables in naive Bayesian networks. It will give insight into the conditions that need to hold in order for the behavior of the classifier to change as a result of a different discretization for a specific feature variable.

The remainder of this thesis is organized as follows. In Section 2, we present some preliminaries on naive Bayesian networks, discretization and sensitivity analysis. In Section 3, we will describe the data set that we used in our experiments and how we applied different discretizations to the continuous-valued feature variables in the domain. Section 4 contains the results and notions of a one-way sensitivity analysis on a specific continuous-valued domain variable. In Section 5 we will investigate the effects on the performance of a naive Bayesian classifier of shifting the threshold value for a continuous-valued feature variable, as well in our domain as in general, making use of two-way sensitivity functions. Our concluding observations will be stated in Section 6.

2 Preliminaries

We will start with reviewing some basic concepts from Bayesian networks. A Bayesian network essentially is a compact representation of a joint probability distribution Pr over a set of stochastic variables V [4]. The variables and their interrelationships are captured as nodes and arcs, respectively, in an acyclic directed graph G. Associated with each node in the graph is a set of parameter probabilities $\theta(V|\pi(V))$ that capture the strength of the relationship between a variable V and its parents $\pi(V)$; we will call this set of parameters the conditional probability table (CPT) of the variable corresponding with the node. From a Bayesian network, any prior or posterior probability of interest over its variables can be computed.

In this thesis, we focus more specifically on naive Bayesian networks (NBN), in which the digraph modeling the interrelationships between the variables has a restricted topology. An

NBN consists of a single class variable and one or more feature variables, which are all directly connected with the class variable in the graph, see Figure 2. In the remainder of this thesis

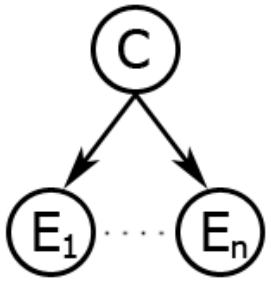


Figure 2: A naive Bayesian network with class variable C and feature variables $E_1, ..., E_n$, with $n \geq 1$.

we will only consider binary class variables C, for which we will use c and $\bar{c}$ to denote the values. The network parameters, i.e. the CPT, for a feature variable E_i, with $i \in \{1, ..., n\}$, consist of probabilities for each value of the variable E_i conditioned on each value for the class variable C. With binary class and feature variables the CPT for variable E_i will look as follows

	c	$\bar{c}$
e_i	p_1	p_3
$\bar{e_i}$	p_2	p_4

where p_j, $j \in \{1, ..., 4\}$, represent the probabilities $\Pr(E_i \mid C)$ for all combinations of values for the variables E_i and C. Note that in this table the probabilities in the same column need to sum to one. In an NBN any two feature variables are conditionally independent given the class variable. For more details, we refer to [6]. Although any probability can be computed from an NBN, the posterior probabilities of the various class values are of primary interest. The network further is associated with a classification function which, based upon these posterior probabilities, returns a single most likely class value, breaking ties at random.

The parameter probabilities of a Bayesian network are either estimated from data or assessed by domain experts, and inevitably include some inaccuracies. To investigate the effects of these inaccuracies on the computed posterior probabilities, a Bayesian network can be subjected to a sensitivity analysis. In such an analysis, one or more network parameters are varied systematically and the effects of this variation on an output probability of interest are studied. We can distinguish between a sensitivity analysis in which just one parameter is being varied; such an analysis is termed a one-way sensitivity analysis, and an n-way sensitivity analysis where $n \geq 2$ parameters are varied simultaneously. For a Bayesian network in general, the effects of the parameter variation are captured by a simple mathematical function, called a sensitivity function. Upon varying a single parameter probability x, the function $f_{\Pr(c|e)}(x)$ that expresses the output probability of interest $\Pr(c|e)$ in terms of x takes the form

$$f_{\Pr(c|e)}(x) = \frac{f_{\Pr(c,e)}(x)}{f_{\Pr(e)}(x)} = \frac{a_1 \cdot x \ + \ a_2}{b_1 \cdot x \ + \ b_2}$$

where the constants a_1, a_2, b_1, b_2 are built from the non-varied parameters[1] from the network under study. For more details and how the constants are derived and computed, see [1, 6]. In the sequel, instead of $f_{\Pr(c|e)}(x)$ we will often write $f(x)$ for short, as long as no confusion is possible. In our analyses, we further assume that parameters with an original assessment of 0 or 1 are not varied, since these parameters represent logical consequences or impossibilities and therefore do not include any inaccuracies. Note that although it is possible to generate a 0 or 1 for a parameter probability by discretization, it would not make much sense to choose such a discretization.

A one-way sensitivity function $f(x)$ for a posterior probability of interest can alternatively be viewed as a fragment of a rectangular hyperbola, which takes the general form

$$f(x) = \frac{r}{x - s_v} + s_h = \frac{t \cdot x + r - s_v \cdot s_h}{x - s_v}, \quad \text{with } s_v = -\frac{b_2}{b_1}, \quad s_h = \frac{a_1}{b_1}, \quad \text{and} \quad r = \frac{a_2}{b_1} + s_v \cdot s_h$$

with the constants a_1, a_2, b_1, b_2 as above. In the remainder of this thesis, we focus on this last representation and assume any sensitivity function to be hyperbolic unless explicitly stated otherwise. A rectangular hyperbola in general has two branches and two asymptotes defining its center (s_v, s_h); Figure 3 illustrates the locations of the possible branches relative to the asymptotes. We observe that a sensitivity function is defined by $0 \leq x, f(x) \leq 1$; the two-dimensional space of feasible points thus defined, is termed the unit window. Since a sensitivity function moreover should be continuous for $x \in [0,1]$, its vertical asymptote necessarily lies outside the unit window, that is, either $s_v < 0$ or $s_v > 1$. From these observations we conclude that a hyperbolic sensitivity function is a fragment of just one of the four possible branches shown in Figure 3.

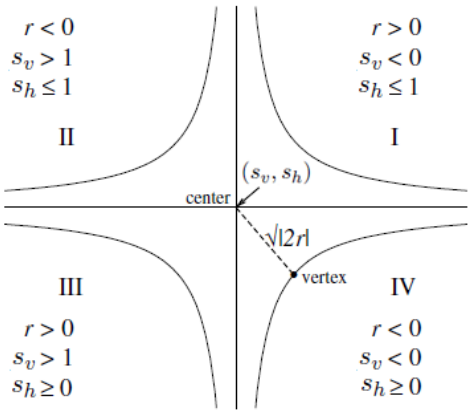


Figure 3: Two rectangular hyperbolas with branches in the Ist and IIIrd quadrants relative to the hyperbola's center, and in the IInd and IVth quadrants, respectively; the constraints on the constants s_v and s_h are specific for sensitivity functions.

In our research we focus on the discretization of continuous-valued feature variables in an NBN. For an NBN, the form of the sensitivity function is simplified [6]. For any sensitivity

[1]When a parameter θ varies as x, its complement $\bar{\theta} = 1 - \theta$ from the same distribution varies as $1 - x$. If θ concerns an k-valued variable, $k > 2$, then the $k - 1$ complementing parameters are covaried proportionally.

function that describes an output probability of a naive Bayesian network in terms of a single feature parameter $x = \theta(e'|c')$, where e' denotes the observed value of the feature variable E and c' is a value of the class variable, we end up with the following constrained sensitivity function:

$$f_{\Pr(c|e)}(x) = \begin{cases} \dfrac{x}{x - s_v} & \text{if } c = c' \text{ and } e = e' \\ \dfrac{x - 1}{x - s_v} & \text{if } c = c' \text{ and } e \neq e' \\ p_0 \cdot \dfrac{x_0 - s_v}{x - s_v} & \text{otherwise} \end{cases}$$

where x_0 is the original value for x and $\Pr(c|e)$ is the output probability of interest with the original value p_0, and p'_0 is the original value of $\Pr(c'|e)$. The value s_v defining the function's vertical asymptote, equals

$$s_v = \begin{cases} x_0 - \dfrac{x_0}{p'_0} & \text{if } e = e' \\ x_0 + \dfrac{1 - x_0}{p'_0} & \text{otherwise} \end{cases}$$

The value s_h defining the horizontal asymptote of the sensitivity function, is

$$s_h = \begin{cases} 1 & \text{if } c = c' \\ 0 & \text{otherwise} \end{cases}$$

Further on in this thesis, it will become clear that we need to look at the joint influence of multiple feature parameters in a BN; a more complex n-way sensitivity function can be used for this goal. Such a function is formed in a similar way as a one-way sensitivity function. For example, a two-way sensitivity function that expresses the posterior probability of interest $\Pr(c|e)$ in terms of the two parameter probabilities x and y is constructed as follows

$$f_{\Pr(c|e)}(x, y) = \frac{f_{\Pr(c,e)}(x, y)}{f_{\Pr(e)}(x, y)} = \frac{a_1 \cdot xy \ + \ a_2 \cdot x \ + \ a_3 \cdot y \ + \ a_4}{b_1 \cdot xy \ + \ b_2 \cdot x \ + \ b_3 \cdot y \ + \ b_4}$$

where $a_i, b_i, i \in \{1, ..., 4\}$, are built from the non-varied parameters in the network under study. In this thesis, we will choose parameters x and y from the same CPT but with a different conditioning context[2]. Since these parameters are independent, we can simplify our sensitivity function: the constants a_1 and b_1 will be equal to zero since there will be no interaction terms [5]:

$$f_{\Pr(c|e)}(x, y) = \frac{a_2 \cdot x \ + \ a_3 \cdot y \ + \ a_4}{b_2 \cdot x \ + \ b_3 \cdot y \ + \ b_4}$$

A naive Bayesian network can in essence only handle discrete variables. The domains of continuous-valued variables need to be discretized before we can use them in an NBN. There exist various methods for this purpose. The easiest being the binary discretization, which just splits the domain of the variable into two intervals. That is, a threshold value is chosen and each value of the variable either falls below the threshold or above it. This

[2]Specific value(s) for the variable(s) on which another variable is conditioned.

defines a new binary variable. We have already demonstrated this method in the example about the effect of discretization on the feature parameters in an NBN, in Section 1. Note that this method is only suitable for a variable whose distribution given a specific class value is monotonically increasing or decreasing. Other discretization techniques comprise equal-width and equal-frequency discretization. These and many other techniques are also used for multi-interval discretization, where the resulting variable consists of more than two intervals. For an (incomplete) overview of discretization methods, see [2].

3 The example domain

We will start with a description of our example domain, followed by a study about the discretization of continuous-valued variables in our domain.

3.1 Description of the example domain

All over the world dairy farms are growing rapidly and start making use of more technology. When using an automated milking system (AMS), a farmer will have less direct contact with his animals and thus it becomes more difficult to monitor their health manually. One of the biggest problems in the industry is the loss of milk due to clinical mastitis (CM), which is an infection in a cow's udder making the milk unsuitable for consumption. Before the modernization, CM was identified relatively easily, since during the milking process the farmer was physically there and able to study the udder health more closely. In [8], a model is proposed that tries to predict CM in cows, making use of data collected by farmers and by the AMS. This data consists of information about the cows themselves, such as parity[3](PAR), days in milk (DIM) and somatic cell count[4] (SCC), and about their milk, such as electrical conductivity (EC) and color. For more information about CM, how the data was collected and the variables itself, we direct the reader to [8].

In this thesis we use the raw data from [8], which comprises information about milkings for which the AMS issued an alert, indicating that it suspects the milked cow to have CM. Because alerted cows were investigated manually by the personnel on the farms, we can distinguish between true-positive (tp) and false-positive (fp) alerts. This means that we can distinguish between alerts that are sent out correctly and alerts that incorrectly indicate an infected udder. The AMS alert information additionally consists of the following variables: the electrical conductivity (based on the highest EC of the four quarters[5] of the cow), the origin of the alert and whether it was based on an increased EC, a deviated color measurement or on both. Two variables describe whether or not a color alert for mastitic and abnormal milk was given. Also the deviation from the expected milk yield is recorded and the number of alerts for that cow in the preceding 12-96 hours. The data set consists of information about 11,156 alerts, with prior probabilities $\Pr(\text{tp}) = 0.014$ and $\Pr(\text{fp}) = 0.986$. Furthermore, cow information not originating from the AMS was added to each alert. This information includes the parity of a cow, the days in milk, the season of the year, the SCC in the previous 30 days

[3]The number of times a cow has given birth.

[4]The number of white blood cells quantified as cells per mL, providing information about the quality of the milk.

[5]One of the four sections in a cow's udder.

(SCC1), the SCC in the 30 days before the previous 30 days (SCC2), for multiparous[6] cows the geometric mean SCC from all available test-day records from the previous lactation, the number of CM cases of the cow in the previous 30 days and the accumulated number of CM cases of the cow in the days before the previous 30 days. The variables which are continuous-valued and thus interesting for our research are: EC, decreased milk production, number of alerts for that cow in the preceding 12-96 hours, PAR, DIM, SCC1, SCC2 and mean SCC.

3.2 Discretizations for the example domain

In order to study the influence that discretizations can have on the performance of a naive Bayesian classifier, we will investigate continuous-valued variables from our domain. In Figure 4 we show the model (NBN) based on the non-AMS information, used to classify an alerted milking as tp or fp; in addition to the four feature variables displayed, the model includes another three features. Based on domain knowledge, we have reason to believe that the

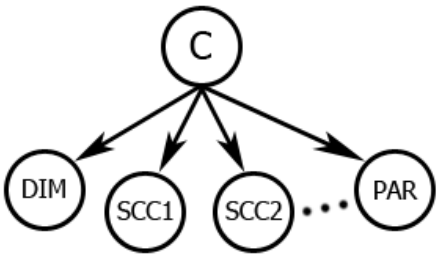


Figure 4: A representation of the NBN used to classify alerts as tp or fp, where C is the class variable and DIM, $SCC1$, $SCC2$ and PAR are (an incomplete set of) the feature variables.

discretization of the $SCC1$ variable from the non-AMS information, which varies from 2,000 to 9,999,000 cells/mL, might have some impact on the classifier's output. Domain experts made the following observation: an SCC $<$ 100,000 cells/mL indicates most likely a healthy cow, an SCC of 100,000-200,000 cells/mL marks a grey area, and an SCC $>$ 200,000 cells/mL should ring the alarm bells. In the data there are many exceptions to this observation, therefore an optimal discretization is not obvious. We have to question ourselves whether we need more than two intervals, when an increasing SCC only gives rise to more suspicion. Because of the nature of the data, we believe that two intervals will fit the data best. More intervals will only blur the results and make the distinction between a normal and an alarming SCC less clear. We thus chose a two-interval discretization for the $SCC1$ variable. To establish an optimal threshold, we shifted the threshold value t with steps of 50,000 cells/mL in a range from 50,000 to 950,000 cells/mL. That is, we started with $t = 50{,}000$, resulting in the intervals [2,000; t) and [t; 9,999,000], and we resumed by increasing t with 50,000 in each step. Values higher than 950,000 cells/mL are very abnormal, as well in the data set as in practice, and therefore it does not make sense to split the range from 950,000 to 9,999,000 cells/mL. For each investigated threshold value we divided our data over the corresponding intervals and established the probabilities of finding $SCC1$ values in those intervals given tp and fp alerts, respectively.

[6] A cow that has given birth two or more times.

In Figure 5 we find the results of our discretizations. Our goal is to identify the threshold value t, for which the difference between $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$ is at its maximum. With such a threshold value we can distinguish best between tp and fp alerts given the value for the $SCC1$ variable. Because of the skewed distribution for the class variable, the difference between $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$ will hardly vary for different threshold values and thus will not give us any information. Therefore we used the difference between the parameters $\Pr(SCC1 < t \mid \text{tp})$ and $\Pr(SCC1 < t \mid \text{fp})$, from the CPT of the $SCC1$ variable, as a heuristic to determine the maximal difference between $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$. To justify this heuristic we can rewrite the probabilities $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$ using Bayes' theorem:

$$\Pr(\text{tp} \,|\, SCC1 < t) = \frac{\Pr(SCC1 < t \,|\, \text{tp}) \cdot \Pr(\text{tp})}{\Pr(SCC1 < t)}$$

$$\Pr(\text{fp} \,|\, SCC1 < t) = \frac{\Pr(SCC1 < t \,|\, \text{fp}) \cdot \Pr(\text{fp})}{\Pr(SCC1 < t)}$$

First, we observe that these two equations have identical denominators. Second, we note that when the value for $\Pr(\text{tp}) = \Pr(\text{fp}) = 0.5$, the difference between $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$ will be proportional to the difference between $\Pr(SCC1 < t \mid \text{tp})$ and $\Pr(SCC1 < t \mid \text{fp})$. As a result we conclude that for prior class probabilities close to 0.5 we can expect our heuristic to perform quite well. However, when we have a very skewed distribution for our class variable, this might not be the case.

In spite of the skewed distribution for our class variable, we will now use our heuristic to identify the best threshold value for the $SCC1$ variable, because the CPT values are the only information we have available. Therefore we will look at the maximal difference between two consecutive bars in Figure 5. Unfortunately each pair of consecutive bars only shows small differences and as a consequence it is hard to distinguish between tp and fp alerts based on just the $SCC1$ variable, regardless of the threshold value t. We do observe that, for a very small threshold value, a switch of the largest bar between each two consecutive bars takes place. That is, in Figure 5 the probability $\Pr(SCC1 < t \mid \text{tp})$ is larger than $\Pr(SCC1 < t \mid \text{fp})$ when the threshold value $t = 50$ and bigger when $t = 100$. However, there are very few true positive alerts in our data set, especially with a very small $SCC1$ value, and as a result we can not draw conclusions based on this observation. Note that for an $SCC1$ above t given a tp or fp alert (not shown) similar results are obtained. We performed a similar study on the $SCC2$ variable from the non-AMS information, which varies from 8,000 to 6,901,000 cells/mL in the data. We again used the range from 50,000 to 950,000 cells/mL for the values for our threshold, but now with steps of 100,000 cells/mL. The results are displayed in Figure 6 and are comparable to Figure 5, which is not a surprise because the definitions of the $SCC1$ and $SCC2$ variables are very much related.

Another variable that was pointed out as potentially interesting by our domain experts, contains values that represent the number of days after a cow has given birth, also called the days in milk (DIM). In the available data the values for this variable ranged from 4 to 798 days. Firstly, we tried a two-interval discretization where we ranged our threshold value from 30 to 330 days with steps of 30 days and from 330 to 690 days, which is less likely, with steps of 60 days. From Figure 7 we can conclude that there are some thresholds for which

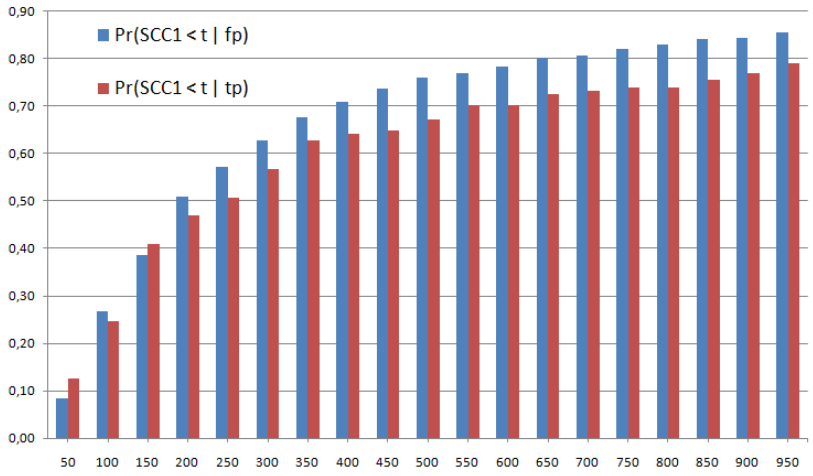


Figure 5: Possible discretizations of the $SCC1$ variable with different thresholds (x 1,000) on the x-axis, and $\Pr(SCC1 < t \mid \text{fp})$ and $\Pr(SCC1 < t \mid \text{tp})$ respectively on the y-axes.

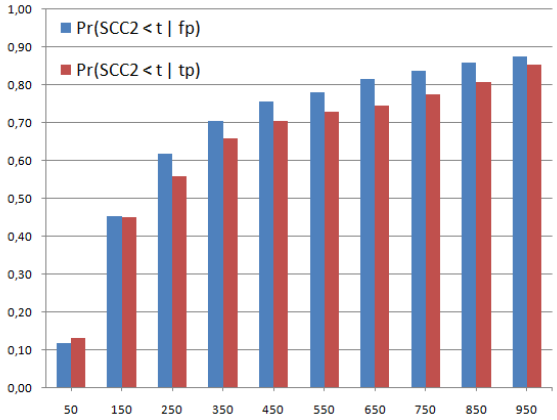


Figure 6: Possible discretizations of the $SCC2$ variable with different thresholds (x 1,000) on the x-axis, and $\Pr(SCC2 < t \mid \text{fp})$ and $\Pr(SCC2 < t \mid \text{tp})$ respectively on the y-axes.

the DIM variable shows some distinction between tp and fp alerts, e.g. with a threshold of 240 days we find $\Pr(DIM < 240 \mid \text{tp}) = 0.748$ and $\Pr(DIM < 240 \mid \text{fp}) = 0.486$. Note that this is the biggest difference between these two probabilities that we found, regarding the threshold values from our experiments. There might be a higher value for this difference with a threshold value between 210 and 270 days, but it has to be in that range; we will return to this observation in more detail in Section 5.4. Thus looking at a two-interval discretization a threshold value of 240 days would distinguish best between tp and fp alerts.

Domain experts did not only point out the DIM variable as interesting, in addition they suggested a split into three intervals. Their advice was to choose a narrow first interval, a medium-sized second interval and a very wide third interval, in order to distinguish between cows with and without CM. We investigated their suggestion of splitting into three intervals,

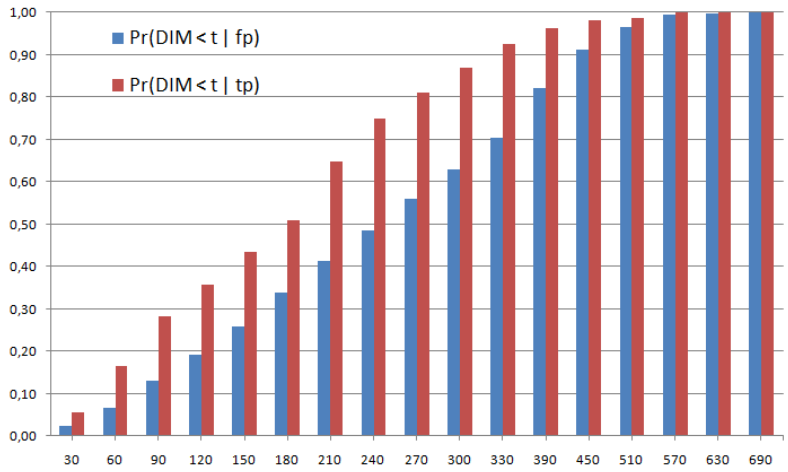


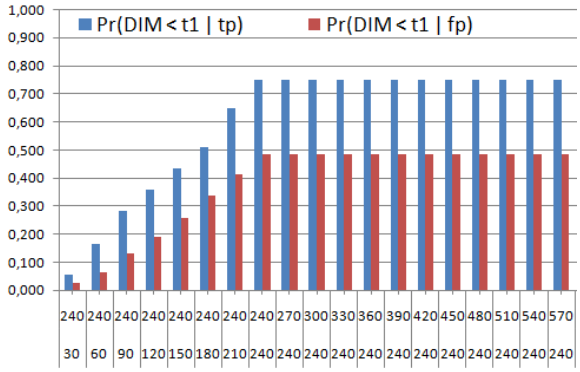
Figure 7: Possible discretizations of the DIM variable with different thresholds on the x-axis, and $\Pr(DIM < t \mid \text{fp})$ and $\Pr(DIM < t \mid \text{tp})$ respectively on the y-axes.

using many combinations and sizes for the intervals. To divide the range for the DIM variable into three intervals we thus need to specify two threshold values: t_1 and t_2. With similar reasoning as before, we again used a heuristic to distinguish between tp and fp alerts given the different intervals for the DIM variable. We looked at the combined absolute difference between the probability of a tp and an fp instance within each interval. That is, we looked at the sum of the following three expressions:

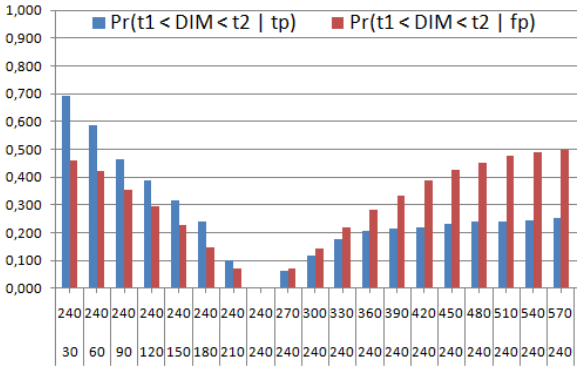
- $|\ \Pr(DIM < t_1 \mid \text{tp}) \ - \ \Pr(DIM < t_1 \mid \text{fp})\ |$
- $|\ \Pr(t_1 < DIM < t_2 \mid \text{tp}) \ - \ \Pr(t_1 < DIM < t_2 \mid \text{fp})\ |$
- $|\ \Pr(DIM \geq t_2 \mid \text{tp}) \ - \ \Pr(DIM \geq t_2 \mid \text{fp})\ |$

With this heuristic, a discretization where one of the thresholds had a value of 240 days always scored the highest. More specifically, for a discretization with one threshold set to 240 days, the value for this combined absolute difference always was the same, regardless of the value for the other threshold. Figure 8 gives an overview of different discretizations where one of the threshold values was 240 days. Because these results are indifferent for the value of the other threshold, they indicate that adding an extra interval to our discretization does not make sense. Hereby we conclude that it is best to choose a two-interval discretization with a threshold value of 240 days, because an extra interval will only complicate our model without actually adding information. We already claimed that a binary discretization is only suitable for a variable whose distribution given a specific class value is monotonically increasing or decreasing. Our findings here suggest to strengthen this claim by stating that when the distribution for a variable conditioned on its parents is either monotonically increasing or decreasing for all conditioning contexts, a binary discretization for that variable is not only suitable, but even preferred above a higher-order discretization.

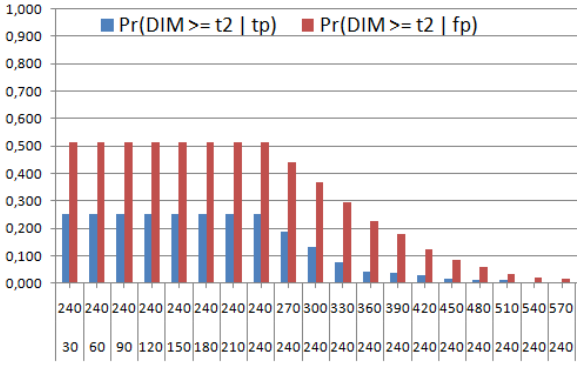
In this section we used heuristics to identify the best threshold values for the discretization of some feature variables from our domain. Because of the highly skewed class variable



(a)



(b)



(c)

Figure 8: Possible discretizations of the DIM variable with different thresholds t_1 and t_2 on the x-axis, and the conditional probabilities for the different intervals given a tp or fp alert on the y-axes.

distribution, we want to verify our results. In the next section we will introduce the concept of sensitivity analyses as a tool for this purpose.

4 One-way sensitivity analysis

In this section we want to use sensitivity analyses to study the effects of changing the discretization for a feature variable on the posterior class variable distribution. In Section 3.2 we have seen that a different discretization for a feature variable leads to other values for its CPT in the NBN. An important observation here is that we can not vary a single value in this CPT, but we always vary them all at the same time. E.g. let's look at the CPT for the $SCC1$ variable with t = 500,000 cells/mL

	tp alert	fp alert
$\theta(SCC1 < t \mid \text{alert})$	0.67	0.76
$\theta(SCC1 \geq t \mid \text{alert})$	0.33	0.24

and assume that we decrease our threshold value t. As a result the values for $\theta(SCC1 < t \mid \text{tp})$ and $\theta(SCC1 < t \mid \text{fp})$ can never increase, but more importantly they will very likely decrease. Since variables with an identical conditioning context need to sum to one, the values for $\theta(SCC1 \geq t \mid \text{tp})$ and $\theta(SCC1 \geq t \mid \text{fp})$ will automatically covary and thus will change too. A one-way sensitivity analysis only captures the variation in one parameter, thereby taking into account the covarying parameter. In general we can conclude that we can never vary just one pair of covarying parameters in the NBN when we change the discretization for a feature variable. However, we will start our investigation with a one-way sensitivity analysis to get some insight into the possible effects on the posterior class variable distribution. In Section 5 we will continue our investigation using a higher-order sensitivity analysis.

As discussed above, we will perform a one-way sensitivity analysis on our domain, more specifically on the separate parameters from the CPT of the $SCC1$ variable. We used the tool called Dazzle [7] to build a classifier on the raw data. First we chose a threshold value t = 500,000 cells/mL and we consider $p = \Pr(\text{tp} \mid SCC1 \geq t)$ and $\bar{p} = \Pr(\text{fp} \mid SCC1 \geq t)$ as probabilities of interest and

- $x = \theta(SCC1 < t \mid \text{tp})$
- $y = \theta(SCC1 < t \mid \text{fp})$
- $u = \theta(SCC1 \geq t \mid \text{tp})$
- $v = \theta(SCC1 \geq t \mid \text{fp})$

for the parameters under study. With Dazzle we determine the original parameter and class probabilities: $x_0 = 0.67$, $y_0 = 0.76$, $u_0 = 0.33$, $v_0 = 0.24$, $p_0 = 0.019$ and $\bar{p}_0 = 0.981$. In Section 2 we have already seen how the sensitivity functions are constructed; the functions corresponding to the four parameters are displayed in Figures 9a, 9b, 9c and 9d respectively. We note that with our sensitivity analysis we also considered $\Pr(\text{tp} \mid SCC1 < t)$ and $\Pr(\text{fp} \mid SCC1 < t)$ as probabilities of interest, however we omitted those results because of similarity.

$$f_{\Pr(\text{fp} \mid SCC1 \geq t)}(x) = \frac{-17.04}{x - 18.04} \qquad f_{\Pr(\text{tp} \mid SCC1 \geq t)}(x) = \frac{x - 1}{x - 18.04}$$

$$f_{\Pr(\text{fp} \mid SCC1 \geq t)}(y) = \frac{y - 1}{y - 1.005} \qquad f_{\Pr(\text{tp} \mid SCC1 \geq t)}(y) = \frac{-0.005}{y - 1.005}$$

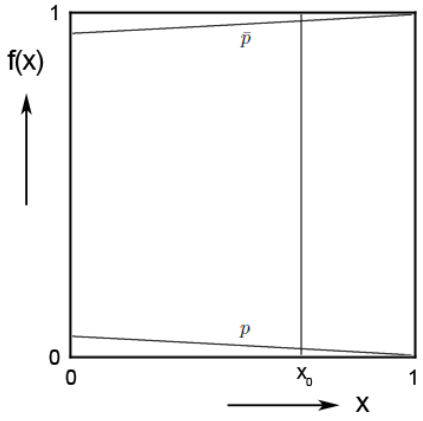
$$f_{\Pr(\text{fp} \mid SCC1 \geq t)}(u) = \frac{17.04}{u + 17.04} \qquad f_{\Pr(\text{tp} \mid SCC1 \geq t)}(u) = \frac{u}{u + 17.04}$$

$$f_{\Pr(\text{fp} \mid SCC1 \geq t)}(v) = \frac{v}{v + 0.005} \qquad f_{\Pr(\text{tp} \mid SCC1 \geq t)}(v) = \frac{0.005}{v + 0.005}$$

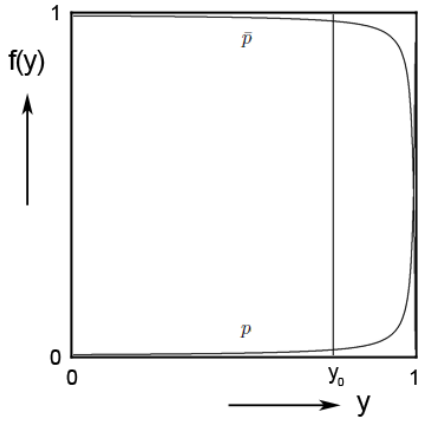
Figures 9a and 9c show that the pairs of sensitivity functions for $\Pr(\text{tp} \mid SCC1 \geq t)$ and $\Pr(\text{fp} \mid SCC1 \geq t)$ do not intersect within the unit window. This means that the classifier will always indicate an fp alert when our instance has an $SCC1 \geq t$, no matter how we change our parameters x and u, which will always covary, because they have to sum to 1. That is, when we choose a new threshold value in order to change *only* the parameter values for x and u, this can not lead to another most likely class value and consequently change the behavior of our classifier; see [9] for more details. Note that the almost horizontal linear behavior of these functions within the unit window is caused by the very small original value p_0 of $\Pr(\text{tp} \mid SCC1 \geq t)$ which dominates the value for the vertical asymptote s_v. For example, let's look at $\Pr(\text{tp} \mid SCC1 \geq t)$ as probability of interest with original value $p_0 = 0.019$ and parameter probability $u = \theta(SCC1 \geq t \mid \text{tp})$ with original value $u_0 = 0.33$. The value for the vertical asymptote can then be computed as follows: $s_v = u_0 - {}^{u_0}/_{p'_0} = 0.33 - {}^{0.33}/_{0.019} = -17.04$, which is dominated by $p'_0 = p_0$, because $c = c'$. Review Section 2 for the formula of the vertical asymptote.

In contrast with Figures 9a and 9c, Figures 9b and 9d do show an observable curve within the unit window. The sensitivity functions for the probabilities of interest p and $\bar{p}$ in fact intersect; note that the value for the sensitivity function always is 0.5 at this intersection point, because we consider a binary class variable. This indicates that there are values for the parameters y and v for which the most likely class value will change, therefore these parameters are interesting for further research. We will now try to establish values for the feature parameters y and v which will change the classifier's behavior, by applying a new discretization for the associated $SCC1$ variable. That is, in Figures 9b and 9d we will try to shift the vertical lines, representing the original values for the parameter probabilities y and v, to the intersection points. In order to obtain such a value for y_0 we need to increase the threshold value t. As a result more fp cases will have an $SCC1$ under the threshold and y_0 will become larger. A similar argument holds for the parameter v: we have to increase our threshold value t for v_0 to move to the intersection point within the graph.

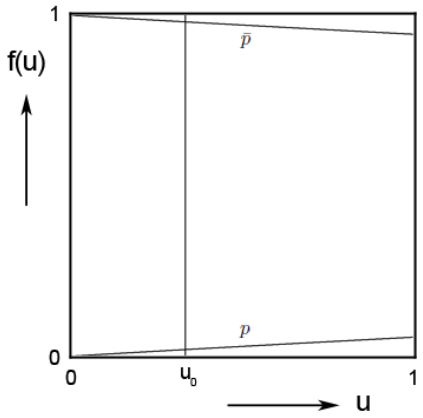
In line with our observations we changed the discretization for the $SCC1$ variable by shifting the threshold value from 500,000 cells/mL to 900,000 cells/mL. Focusing on the sensitivity function from Figure 9b, we observed that by changing the discretization for the $SCC1$ variable, not only the CPT changed, but the sensitivity function itself changed too. With the



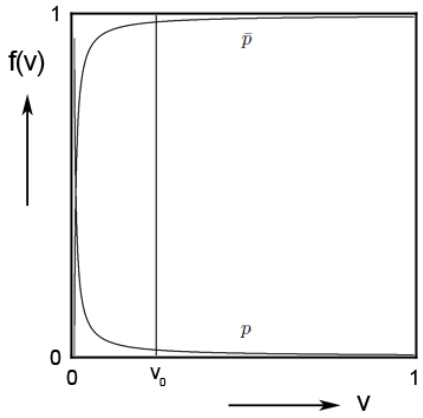
(a) $x = \theta(SCC1 < t \mid \text{tp})$ with $x_0 = 0.67$.



(b) $y = \theta(SCC1 < t \mid \text{fp})$ with $y_0 = 0.76$.



(c) $u = \theta(SCC1 \geq t \mid \text{tp})$ with $u_0 = 0.33$.



(d) $v = \theta(SCC1 \geq t \mid \text{fp})$ with $v_0 = 0.24$.

Figure 9: One-way sensitivity functions with $p = \Pr(\text{tp} \mid SCC1 \geq t)$ and $\bar{p} = \Pr(\text{fp} \mid SCC1 \geq t)$ as probabilities of interest, for different $SCC1$ parameters and with threshold value $t = 500{,}000$ cells/mL.

new threshold value of $t = 900{,}000$ cells/mL, we found the following new sensitivity function, which is represented in Figure 10.

$$f_{\Pr(\text{fp} \mid SCC1 \geq t)}(y) = \frac{y - 1}{y - 1.002}$$

We notice that by moving y_0 closer to the point where $f(y) = 0.5$, represented by the shift from Figure 9b to Figure 10, the curve itself gets pushed away from y_0. That is, the value for y for which $f(y) = 0.5$ will move in the same direction as y_0. As a consequence, changing the value for the parameter y will never have a significant impact on the probability of interest, no matter how high we set the threshold value t for the $SCC1$ variable.

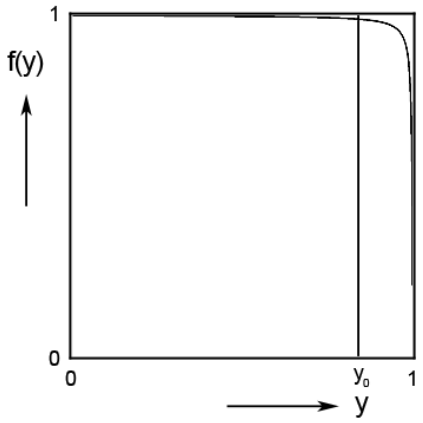


Figure 10: One-way sensitivity function $f_{\Pr(\text{fp} \mid SCC1 \geq t)}(y)$ with $y = \theta(SCC1 < t \mid \text{fp})$, $y_0 = 0.84$ and $t = 900{,}000$.

An observant reader might put some question marks by the fact that our sensitivity function changed when we shifted the value for our parameter y, while the goal of the sensitivity function itself was to show the influence of changing the parameter on the probability of interest. We note however, that we did not just change the value of the parameter y, but changed the way it was computed, by shifting the threshold value t for the $SCC1$ variable. By changing y from $y = \theta(SCC1 < 500{,}000 \mid \text{fp})$ to $y = \theta(SCC1 < 900{,}000 \mid \text{fp})$, the parameter x also changed to $x = \theta(SCC1 < 900{,}000 \mid \text{tp})$ with a different x_0. This effect was not captured in our one-way sensitivity function. Note that this is a different type of covariation than the covariation between x and u, and y and v, which was described earlier and which *is* captured by a one-way sensitivity function. We conclude that a one-way sensitivity analysis does not fully describe the effect of alternative discretizations, which will be further explained in Section 5.

5 Shifting the threshold value

In this section we will describe what the influence will be on the posterior class distribution if we shift the threshold value for the discretization of a feature variable in an NBN, as described in Section 3.2. This section builds on the conclusions formed in Section 4 and consequently we switch from one-way to two-way sensitivity analyses. We first establish some theoretical properties of two-way sensitivity functions for NBNs, which simplify our experiments later on in this section. Thereafter we explain when one-way sensitivity functions are not suitable as a tool for investigation. We continue with a two-way sensitivity analysis on our example domain and we end to explain how we can use two-way sensitivity functions to identify situations in which discretization can have impact on the most likely class value.

5.1 Two-way sensitivity functions for NBNs

Building on Section 4, we will continue to investigate the effect on the posterior class variable distribution of changing the discretization for the $SCC1$ variable. Because our feature

variable $SCC1$ is binary we know that we will always vary all four feature parameters when we change the discretization for this variable and therefore we know that we will need a two-way sensitivity function to fully capture the effect of a different discretization. When a feature variable consists of more than two intervals an even higher-order sensitivity function will be necessary to capture all parameters that vary as a result of a new discretization. Before we start with the actual analysis we begin by deriving the general form of a two-way sensitivity function in an NBN, on two parameters from the same CPT. Note that where a one-way sensitivity function describes a line in two dimensions, a two-way sensitivity function describes a plane in three dimensions.

Proposition 1. *Consider an NBN with the class variable C, with values c and $\bar{c}$, and a single feature variable E. Let e be an arbitrary value of E, and let $x = \theta(e|c)$ and $y = \theta(e|\bar{c})$ be parameter probabilities for E. The sensitivity function for the probability of interest $\Pr(c|e)$ then has the following form:*

$$f_{\Pr(c|e)}(x, y) = \frac{\Pr(c) \cdot x}{\Pr(c) \cdot x \ + \ \Pr(\bar{c}) \cdot y} \tag{1}$$

Proof. We consider the probability of interest $\Pr(c|e)$. Using Bayes' theorem we can express the probability of interest as follows:

$$\Pr(c|e) = \frac{\Pr(e|c) \cdot \Pr(c)}{\Pr(e|c) \cdot \Pr(c) + \Pr(e|\bar{c}) \cdot \Pr(\bar{c})}$$

Note that all terms involved correspond to parameters in the NBN; in fact, with $x = \theta(e|c)$ and $y = \theta(e|\bar{c})$ the result follows. □

A two-way sensitivity function in an NBN will look slightly different from the above form in the cases described next:

- We have $\Pr(\bar{c}|e)$ as probability of interest instead of $\Pr(c|e)$, with parameter probabilities $x = \theta(e|c)$ and $y = \theta(e|\bar{c})$ as before:

$$f_{\Pr(\bar{c}|e)}(x, y) = \frac{\Pr(\bar{c}) \cdot y}{\Pr(c) \cdot x \ + \ \Pr(\bar{c}) \cdot y}$$

- The probability of interest $\Pr(c|\bar{e})$ has a different observation for the (binary) feature variable than the parameter probabilities $x = \theta(e|c)$ and $y = \theta(e|\bar{c})$ considered:

$$f_{\Pr(c|\bar{e})}(x, y) = \frac{\Pr(c) \cdot (1 - x)}{\Pr(c) \cdot (1 - x) \ + \ \Pr(\bar{c}) \cdot (1 - y)}$$

- The probability of interest $\Pr(c|e_i, e_j)$ is conditioned on multiple feature variables, with the parameter probabilities $x = \theta(e_j|c)$ and $y = \theta(e_j|\bar{c})$:

$$f_{\Pr(c|e_i,e_j)}(x, y) = \frac{\Pr(c) \cdot \Pr(e_i|c) \cdot x}{\Pr(c) \cdot \Pr(e_i|c) \cdot x \ + \ \Pr(\bar{c}) \cdot \Pr(e_i|\bar{c}) \cdot y}$$

- Any combination of the three cases above.

In Section 2 we already stated that a two-way sensitivity function where x and y are from the same CPT, but with a different conditioning context, has the following general form

$$f_{\Pr(c|e)}(x,y) = \frac{a_2 \cdot x \ + \ a_3 \cdot y \ + \ a_4}{b_2 \cdot x \ + \ b_3 \cdot y \ + \ b_4}$$

When we combine this form with the property stated in Proposition 1, we find that $a_3 = a_4 = b_4 = 0$, $a_2 = b_2 = \Pr(c)$ and $b_3 = \Pr(\bar{c})$. In the case where the probability of interest is conditioned on multiple feature variables the constants a_2, b_2 and b_3 need to be multiplied by some further constants, which are built from the CPT values of the extra variable(s) on which the probability of interest is conditioned. Note that these two-way sensitivity functions do not depend on the original values for the probability of interest and the parameters under study. Furthermore, we note that this proposition does not only hold in NBNs, but also in Bayesian networks with the variables C and $E = \{E_1, ..., E_n\}$, $n \geq 1$, considering the probability of interest $\Pr(c \mid e_1, ..., e_n)$ and the following restrictions:

- when there is a connection $C \longrightarrow E_i$, with $i \in \{1, ..., n\}$, it forms a bridge in the network, i.e. removing the arrow from the variable C to E_i will result in two unconnected sub graphs, see Figure 11
- the variables E with a connection to C have no other parents than the variable C in the network

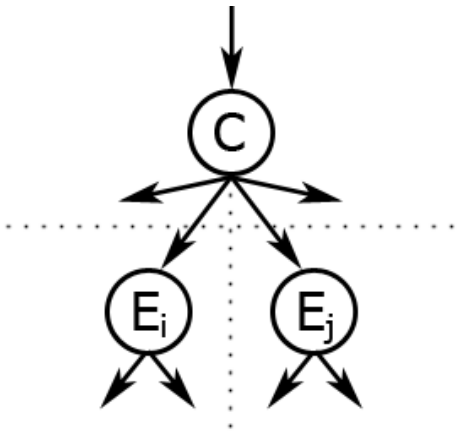


Figure 11: The arrows from variable C to variables E_i and E_j form bridges in this network.

5.2 Unsuitability of a one-way sensitivity function

Here we will state a proposition about the effect that we have seen in Section 4, where we used discretization to push a parameter's original value to the point where the most likely class value changes. We will use the form of a two-way sensitivity function to show why a one-way sensitivity function does not provide enough insights when discretization is used to change a parameter's original value.

Proposition 2. *Consider a Bayesian network and two parameters x and y from the same CPT, but for different conditioning contexts. For r, the vertical asymptote s_v and the horizontal asymptote s_h, describing a sensitivity function $f(x)$ that is a fragment of a rectangular hyperbola, we have that*

$$\begin{aligned} s_v &= d_1 \cdot y + d_2 \\ s_h &= d_3 \\ r &= d_4 \cdot y + d_5 + s_v \cdot s_h \end{aligned}$$

where $d_i, i \in \{1, ..., 5\}$, are constants with respect to x and y.

Proof. In Section 5.1 we saw that we can establish the following associated two-way sensitivity function:

$$f(x, y) = \frac{a_2 \cdot x + a_3 \cdot y + a_4}{b_2 \cdot x + b_3 \cdot y + b_4}$$

When we look at this function as a one-way sensitivity function in x, where y is a constant, we can determine r, s_v and s_h as follows:

$$\begin{aligned} s_v &= -\frac{b_3 \cdot y + b_4}{b_2} = -\frac{b_3}{b_2} \cdot y - \frac{b_4}{b_2} \\ s_h &= \frac{a_2}{b_2} \\ r &= \frac{a_3 \cdot y + a_4}{b_2} + s_v \cdot s_h = \frac{a_3}{b_2} \cdot y + \frac{a_4}{b_2} + s_v \cdot s_h \end{aligned}$$

the result now follows immediately, with $d_1 = -b_3/b_2$, $d_2 = -b_4/b_2$, $d_3 = a_2/b_2$, $d_4 = a_3/b_2$ and $d_5 = a_4/b_2$. □

We can now conclude that the vertical asymptote of the one-way sensitivity function in x has a linear relation with respect to the parameter y. Because the value for y changes when we vary the value for x by means of a new discretization, our one-way sensitivity function will vary too and thus will not give an accurate and complete view. The horizontal asymptote s_h does not depend on the value for parameter y and will stay constant, furthermore the value for r will, similar to the vertical asymptote, have a linear relationship with parameter y. The change in the vertical asymptote s_v for $f(x)$ upon varying parameter x, due to the covarying of the parameter y, will depend on the relation between the parameters x and y. For example, when parameters x and y exhibit some linear relationship, $y = a \cdot x + b$, parameter x will also have a linear relationship with s_v:

$$\begin{aligned} s_v &= d_1 \cdot y + d_2 \\ &= d_1 \cdot g_1 \cdot x + (d_1 \cdot g_2 + d_2) \\ &= g_3 \cdot x + g_4 \end{aligned}$$

We will now apply the property stated in Proposition 2 to the constrained two-way sensitivity function from Proposition 1.

Corollary 1. *Consider an NBN with variables C and E and parameters x en y as in Proposition 1. Then r, the vertical asymptote s_v and the horizontal asymptote s_h, for the sensitivity function $f_{\Pr(c|e)}(x)$ for probability of interest $\Pr(c|e)$ will have the following form:*

$$\begin{aligned} s_v &= -\frac{\Pr(\bar{c})}{\Pr(c)} \cdot y \\ s_h &= 1 \\ r &= s_v \cdot s_h = s_v \end{aligned}$$

Proof. Consider parameter y as a constant in the two-way sensitivity function $f_{\Pr(c|e)}(x, y)$ of Proposition 1. The result then follows from the proof of Proposition 2. □

Here we note that apart from the type of the relationship between the parameters x and y, the function for the vertical asymptote s_v will never change sign as a result of a change in the value for x. Since, when we vary the value for parameter x, the only variable who's value can change in the function for s_v is y, which represents a probability and thus will always be positive. Moreover, in the NBN under consideration, the coefficient for y in the expression for s_v is always negative, resulting in a vertical asymptote to the left of the unit window. The variable r will display exactly the same behavior as our vertical asymptote, because the value for our horizontal asymptote will always be $s_h = 1$.

The above observations explain the phenomenon described in Section 4 (Figures 9b and 10), where our one-way sensitivity function shifted when we varied the value for the parameter under study by means of a new discretization.

5.3 Two-way sensitivity analysis in our example domain

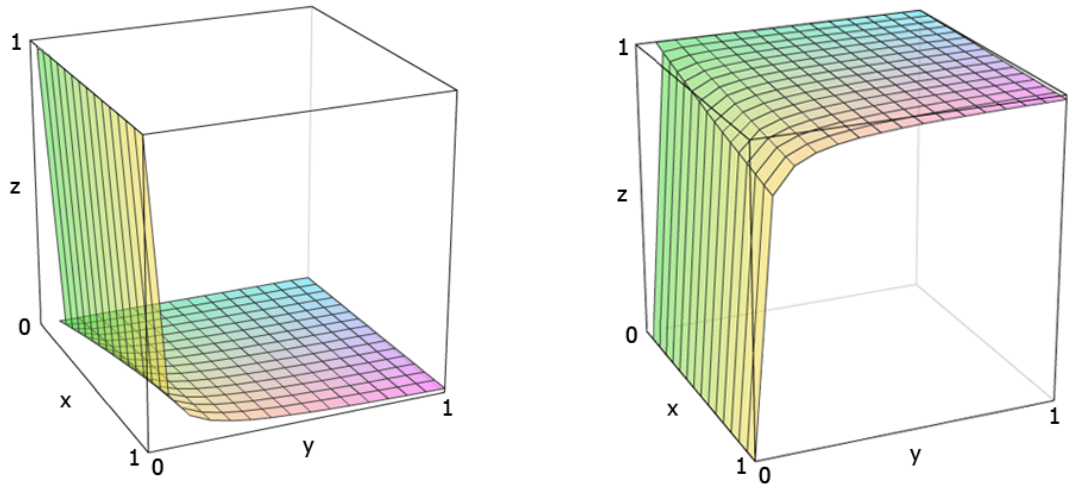
In this section we will perform a two-way sensitivity analysis on the $SCC1$ feature variable from [8] and we will look at the influence of the parameters $x = \theta(SCC1 < t \mid \text{tp})$ and $y = \theta(SCC1 < t \mid \text{fp})$ on the posterior class probabilities $\Pr(\text{tp} \mid SCC1 < \text{t})$ and its complement $\Pr(\text{fp} \mid SCC1 < \text{t})$. From the data we know that $\Pr(\text{tp}) = 0.014$ and $\Pr(\text{fp}) = 0.986$, and we construct the following two-way sensitivity functions, using Proposition 1:

$$f_{\Pr(\text{tp} \mid SCC1 < t)}(x, y) = \frac{0.014 \cdot x}{0.014 \cdot x \ + \ 0.986 \cdot y} \tag{1}$$

$$f_{\Pr(\text{fp} \mid SCC1 < t)}(x, y) = \frac{0.986 \cdot y}{0.986 \cdot y \ + \ 0.014 \cdot x} \tag{2}$$

These functions are displayed in Figures 12a and 12b. In these figures we can observe exactly what we have stated in Section 5.2. That is, when we only look at the influence of parameter x in Figure 12a, we see that for every y this sensitivity function will look slightly different; for every y a 2D slice of the cube represents our one-way sensitivity function in parameter x only.

In Section 5.1 we saw that the parameters x and y exhibit some sort of covariation, as a result of the discretization. Therefore not all combinations of x and y values are possible, which means that the sensitivity functions (1) and (2) can not take on all the values displayed



(a) $f_{\Pr(\text{tp} \mid SCC1 < t)}(x, y) = \frac{0.014 \cdot x}{0.014 \cdot x \ + \ 0.986 \cdot y}$ (b) $f_{\Pr(\text{fp} \mid SCC1 < t)}(x, y) = \frac{0.986 \cdot y}{0.986 \cdot y \ + \ 0.014 \cdot x}$

Figure 12: Two-way sensitivity functions with the parameters $x = \theta(SCC1 < t \mid \text{tp})$ and $y = \theta(SCC1 < t \mid \text{fp})$.

respectively in Figures 12a and 12b. Note that the actual variation we achieve in x and y is by using a different threshold value t for the $SCC1$ variable. To limit our sensitivity functions to the values that are actually reachable in our domain, we first have to formalize the relationship between x and y. Thereafter we can restrict our sensitivity functions by rewriting x in terms of y or vice versa, which will result in two-dimensional sensitivity functions again. These functions will give a better view of the actual values for our sensitivity function, because they do not contain unreachable points. In our example the relation between x and y seems to resemble a linear function. We find the following relation using linear regression on the x and y values from Figure 5, described in Section 3.2:

$$\begin{aligned} x &= 0.845 \cdot y + 0.0448, \text{ or} \\ y &= 1.17 \cdot x - 0.0478 \end{aligned}$$

In Figure 13 we see a step function representing the values used for the regression and the resulting linear function itself, both corresponding to the parameter x expressed in terms of the parameter y. We observe that our linear approximation matches the true relation between parameter x and y very well and therefore we continued to use this approximation in this thesis.

We can now use these equations to express x in y and vice versa in our sensitivity functions (1) and (2), to end up with four new functions that only have one unknown. For each two-way function we get two new functions: one with x as parameter and one with y as parameter. Note that both pairs of functions are mere approximations, because the linear relationship between the parameters x and y is an approximation of the true relationship. Our resulting one-way sensitivity functions express the effect of variation in parameter x (or y), as a result of discretization, on the output probability of interest. These one-way sensitivity functions have the following form:

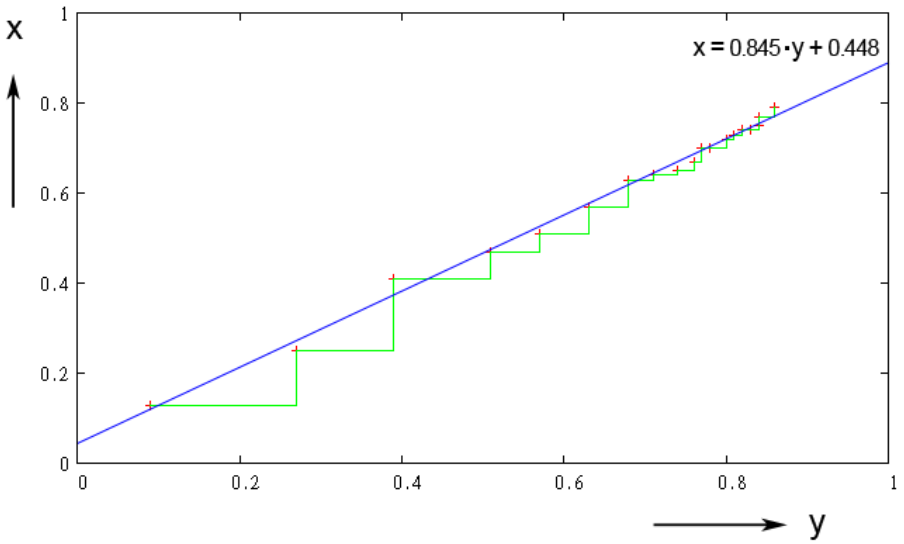


Figure 13: The true (step function) and approximate (linear function) relation between parameters x and y.

$$f_{\Pr(\text{tp} \mid SCC1 < t)}(x) = \frac{0.014 \cdot x}{1.17 \cdot x - 0.0471}$$

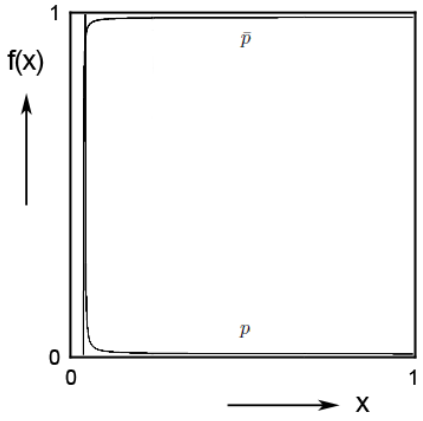
$$f_{\Pr(\text{tp} \mid SCC1 < t)}(y) = \frac{0.012 \cdot y + 0.0006}{0.998 \cdot y + 0.0006}$$

and

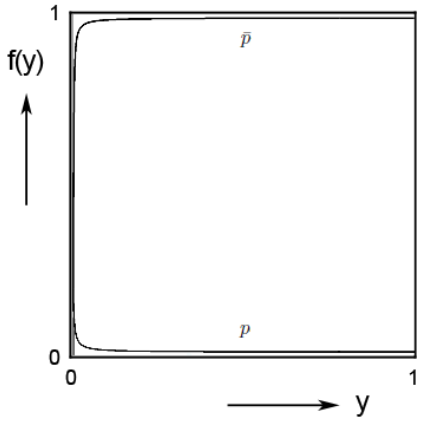
$$f_{\Pr(\text{fp} \mid SCC1 < t)}(x) = \frac{1.15 \cdot x - 0.0471}{1.16 \cdot x - 0.0471}$$

$$f_{\Pr(\text{fp} \mid SCC1 < t)}(y) = \frac{0.986 \cdot y}{0.998 \cdot y + 0.0006}$$

Looking at our results in Figure 14 we can conclude that it is very unlikely that the $SCC1$ parameters will ever have any influence on the posterior class distribution of the NBN used to distinguish between tp and fp alerts. We have to make a very unrealistic discretization for the $SCC1$ variable to obtain the extremely low original parameter values needed to change the most likely class value. That is, we would need to set the threshold value for our $SCC1$ variable below 50,000 cells/mL, which does not seem very logical from domain expert point of view. In the process we did find a way to approximate the one-way sensitivity functions in the two parameters from a two-way function, by formalizing the relationship between these two parameters. This limits the possible values for our (two-way) sensitivity function drastically by eliminating unreachable points. However, this is only possible when the two parameters



(a) Sensitivity functions $p = f_{\Pr(\text{tp} \mid SCC1 < t)}(x)$ and $\bar{p} = f_{\Pr(\text{fp} \mid SCC1 < t)}(x)$, with $x = \theta(SCC1 < t \mid \text{tp})$.



(b) Sensitivity functions $p = f_{\Pr(\text{tp} \mid SCC1 < t)}(y)$ and $\bar{p} = f_{\Pr(\text{fp} \mid SCC1 < t)}(y)$, with $y = \theta(SCC1 < t \mid \text{fp})$.

Figure 14: Approximations of one-way sensitivity functions.

show some relation in the corresponding domain. Furthermore, we repeated our experiments for the $SCC2$ and the DIM variable, but we omitted the results because of similarity.

5.4 When can discretization have impact?

With respect to an NBN, it is interesting to know in which situations a different discretization for a feature variable can have impact on the behavior of the corresponding classifier. Here we will try to describe the conditions that need to be met in order to change the most likely class value for a particular probability of interest, by changing the (binary) discretization for a feature variable. Consider an NBN with binary class (C) and feature (E) variables, $\Pr(c|e)$ and $\Pr(\bar{c}|e)$ as probabilities of interest and two network parameters $x = \theta(e|c)$ and $y = \theta(e|\bar{c})$ that represent probabilities in the same CPT, but with a different conditioning context. We have already seen that the influence of changing the threshold value for the variable E on the NBN can be studied using a two-way sensitivity function in the parameters x and y. Our goal is to determine in which situations we can change the (binary) discretization for the variable E in such a way that it will lead to values for x and y for which the most likely class value will change. That is, we want to identify for which values of $\Pr(c)$, $\Pr(\bar{c})$, x and y the most likely class value for a particular probability of interest can change. To find such values for $\Pr(c)$, $\Pr(\bar{c})$, x and y, we will first need to identify the intersection between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$. Note that the intersection is a line. Next, we check if there is a point on that intersection line with reachable x and y values. That is, because of the relationship between the parameters x and y, not all combinations of x and y values can occur. Therefore we have to check whether the previously mentioned intersection line will intersect, with the function describing the relation between x and y, within the unit window. If the latter is the case, there will very likely be pairs of x and y values leading to both values for the class variable and thus making a change in the most likely class value possible.

From Section 5.3 we know that in our domain example the relation between x and y is almost linear. In the remainder of this section we will assume that this relationship has the form $y = a \cdot x + b$. Furthermore, we know that $a \geq 0$, because when we change the threshold value in our discretization the values for x and y can only change in the same direction. I.e. it can never happen that x becomes smaller and y becomes larger or vice versa, when we increase or decrease our threshold value. Note that even when the relationship between x and y is not linear, it will *always* be monotonic. We will use this property later on in this section, but we will illustrate it here. For example, again let's look at Figure 1 and the two corresponding CPTs from Section 1. When we look at the shift of the threshold value t from 10 to 15 we can observe that $\Pr(e < t \mid c = 1)$ and $\Pr(e < t \mid c = 2)$ both become larger and $\Pr(e \geq t \mid c = 1)$ and $\Pr(e \geq t \mid c = 2)$ both become smaller, which is exactly the behavior that we described. In general, increasing the threshold value t, for a binary (feature) variable E, can *never* lead to lower CPT values for $\Pr(e < t \mid c)$ and higher values for $\Pr(e \geq t \mid c)$, where c is a specific value for the class variable. This should be clear, because increasing the threshold value for a variable will lead to a larger interval below the threshold for that variable and as a result the number of cases in the data with a value under the threshold for that variable can never decrease, which means that $\Pr(e < t \mid c)$ can never decrease when we increase the threshold value. With similar reasoning the same argument holds for an interval above the threshold value or when we decrease the threshold value. Now that we have established the relationship between the parameters x and y, we continue to introduce the variable $q = \Pr(c)$, and thus $q - 1 = \Pr(\bar{c})$, in order to rewrite our two-way sensitivity functions to find the equation for their intersection line:

$$
\begin{aligned}
f_{\Pr(c|e)}(x, y) &= f_{\Pr(\bar{c}|e)}(x, y) \qquad \Longleftrightarrow \\
\frac{q \cdot x}{q \cdot x \ + \ (1-q) \cdot y} &= \frac{(1-q) \cdot y}{q \cdot x \ + \ (1-q) \cdot y} \qquad \Longleftrightarrow \\
y &= \frac{q}{1-q} \cdot x
\end{aligned}
$$

In the function for $z = f_{\Pr(c|e)}(x, y)$ we can substitute y with an equation in terms of x, as described above, to determine the corresponding z values for our intersection line:

$$
\begin{aligned}
z &= \frac{q \cdot x}{q \cdot x \ + \ (1-q) \cdot y} \\
&= \frac{q \cdot x}{q \cdot x \ + \ (1-q) \cdot \frac{q}{1-q} \cdot x} \\
&= 0.5
\end{aligned}
$$

We can conclude that the z value for our intersection line will always be 0.5, independent of the values for x and y. Note that this is similar to the case where we had a one-way sensitivity function, because we are still dealing with a binary class variable.

We can now express the values for the parameters x and y that identify the point where the most likely class value will change, in terms of a, b and q, where q represents the prior (binary) class variable distribution and a and b describe the linear relationship between the

parameters x and y. Using the equation for the intersection line between $f_{\Pr(c|e)}(x,y)$ and $f_{\Pr(\bar{c}|e)}(x,y)$, and the (assumed) linear relationship between x and y, we can compute the following expressions for x and y:

$$\begin{aligned} y &= a \cdot x + b &&= \frac{q}{1-q} \cdot x && \iff \\ b &= (\frac{q}{1-q} - a) \cdot x && && \iff \\ x &= \frac{b}{\frac{q}{1-q} - a} &&= \frac{b \cdot (1-q)}{q - a \cdot (1-q)} && \iff \\ y &= \frac{q}{1-q} \cdot x &&= \frac{b \cdot q}{q - a \cdot (1-q)} \end{aligned}$$

Because we are considering two-way sensitivity functions, we are dealing with a unit cube rather than a unit window, however we can limit the expressions for x and y above to two dimensions, because the change in the most likely class value will always take place when the value for $z = 0.5$. Further the relationship between the parameters x and y pertains to all values for z, also $z = 0.5$. In order for the intersection line between $f_{\Pr(c|e)}(x,y)$ and $f_{\Pr(\bar{c}|e)}(x,y)$ and the relationship between x and y to intersect within the unit cube, we need to state some conditions for the values of the parameters a, b and q. Consider Figure 15, where the gradient of (1), the intersection line $y = {}^{q}/_{(1-q)} \cdot x$ between $f_{\Pr(c|e)}(x,y)$ and $f_{\Pr(\bar{c}|e)}(x,y)$, is less than or equal to one, i.e. $q \leq {}^{1}/_{2}$. Here we can distinguish two different cases for line (1) to intersect with the line that describes the relation between x and y: the value for b is either positive (2) or negative (3).

• Case 1a) The value for b is positive and the value for a, the gradient of line (2), must be smaller than ${}^{q}/_{(1-q)}$. Further we already noted that it must hold that $a \geq 0$, therefore we can also compute the maximal value for b, resulting in the following possible values for a, b and q:

$$\begin{aligned} q &\leq \tfrac{1}{2} \\ 0 \leq a &< \frac{q}{1-q} \\ 0 \leq b &\leq \frac{q}{1-q} - a \end{aligned}$$

• Case 1b) The value for b is negative and the value for a, the gradient of line (3) must be larger than ${}^{q}/_{(1-q)}$, resulting in the following values for a, b and q:

$$\begin{aligned} q &\leq \tfrac{1}{2} \\ a &> \frac{q}{1-q} \\ \frac{q}{1-q} - a \leq b &\leq 0 \end{aligned}$$

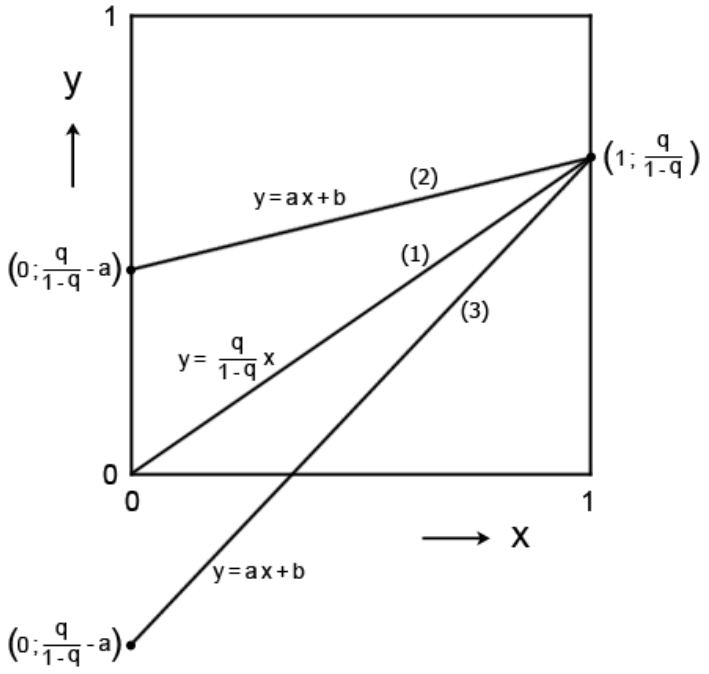


Figure 15: Graphical representation of the cases where $q = \Pr(c) \leq 1/2$. Line (2) has the highest possible gradient and line (3) the lowest possible gradient in order to intersect with line (1) within the unit window, taking into account the corresponding b values.

When we consider Figure 16, where the gradient of line (1) is greater than one and the corresponding value for $q > 1/2$, we can again similarly distinguish two different cases.

• Case 2a) The value for b is positive and the value for a, the gradient of line (2) must be smaller than $q/(1-q)$, resulting in the following values for a, b and q:

$$\begin{aligned} q &> \tfrac{1}{2} \\ 0 \leq a &< \frac{q}{1-q} \\ 0 \leq b &\leq 1 - \frac{1-q}{q} \cdot a \end{aligned}$$

• Case 2b) The value for b is negative and the value for a, the gradient of line (3) must be larger than $q/(1-q)$, resulting in the following values for a, b and q:

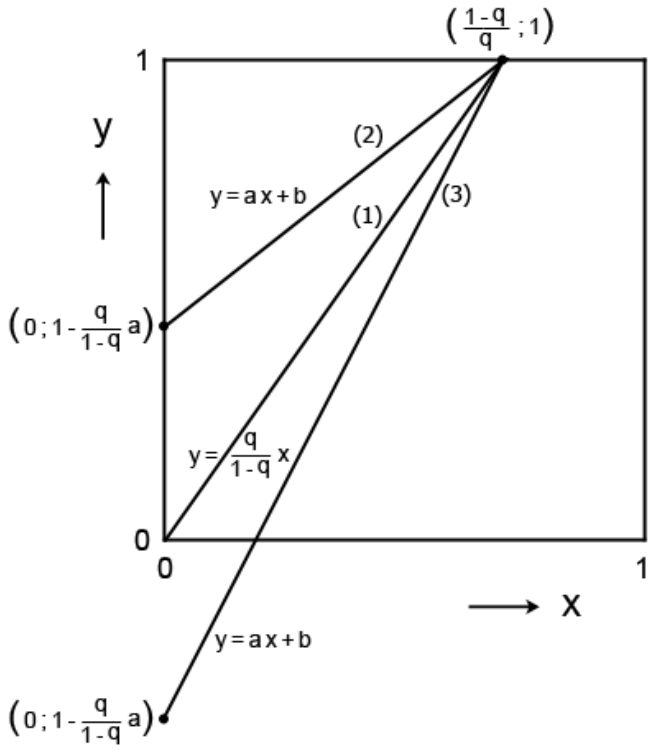


Figure 16: Graphical representation of the cases where $q = \Pr(c) > 1/2$. Line (2) has the highest possible gradient and line (3) the lowest possible gradient in order to intersect with line (1) within the unit window, taking into account the corresponding b values.

$$\begin{array}{rcl} q & > & \frac{1}{2} \\ a & > & \dfrac{q}{1-q} \\ 1 - \dfrac{1-q}{q} \cdot a & \leq & b \quad \leq \quad 0 \end{array}$$

When the values for a, b and q in your NBN correspond to one of these four cases, we know that a change in the most likely class value is possible for a particular probability of interest. However, when the values do not correspond to one of the cases above, we are certain that a different discretization can not lead to a change in the most likely class value for a particular probability of interest. As an example we will now use the above conditions to check whether a different discretization for the $SCC1$ variable of our domain example from Section 5.3 can have impact on the classifier's performance. In Section 5.3 we investigated the influence of the parameter probabilities $x = \theta(SCC1 < t \mid \text{tp})$ and $y = \theta(SCC1 < t \mid \text{fp})$ on the posterior class probabilities $\Pr(\text{tp} \mid SCC1 < \text{t})$ and its complement $\Pr(\text{fp} \mid SCC1 < \text{t})$. We will first identify the variables a, b and q for this example. We know that $\Pr(\text{tp}) = 0.014$ and $\Pr(\text{fp}) = 0.986$ together constitute the prior class variable distribution, i.e. $q = 0.014$. Furthermore, in Section 5.3 we found the relationship between the parameters x and y to be

$y = 1.17 \cdot x - 0.0478$, i.e. $a = 1.17$ and $b = -0.0478$. We now find that Case 1b corresponds to this situation, which means that a different discretization of the $SCC1$ variable can have impact on the NBN that tries to distinguish between tp and fp alerts. However, looking at the more detailed investigation from Section 5.3 we find that it is very unlikely that another discretization of the $SCC1$ variable will have a significant effect on the performance of the corresponding NBN. Therefore we can conclude that our findings in this section provide a mere tool to identify situations where discretization can have impact, but more importantly exclude situations in which discretization can not lead to another most likely class value for a particular probability of interest. This small drawback is caused by the gap between the theoretical possibilities and the limited reality. For example, in our domain a very low threshold value for the $SCC1$ variable can in theory lead to another most likely class value, however in practice these very low SCC values are not realistic.

When we can not approximate the relationship between the parameters x and y using linear regression, because their relation has a different nature, we can still determine whether there are values for x and y for which the most likely class value will change. That is, as long as we can formalize the relationship between x and y, we can just plot it together with the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$ and check if the two intersect within the unit window. We have to note that the quality of the approximation for the relation between x and y is very important for the quality of the plot. Small errors in this approximation can lead to an intersection in the plot that in practice may not exist. Alternatively to an approximation of the relation between parameters x and y, we can use only the set of computed x and y values, each belonging to a different threshold value, to determine whether the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$ intersects with the relation between x and y. That is, when we find a set of x and y values on both sides of the intersection line, we know there will be an intersection point, because the relationship between x and y will always be monotonic. Yet, it is not evident how to select the sets of x and y values to consider. A good heuristic might be to choose two (x, y) points, one with the smallest and one with largest value for x, and check whether those two (x, y) points lay on a different side of the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$. However, it can be the case that all considered (x, y) points lay on the same side of the intersection line, but in practice a change in most likely class value is possible. For example, consider some non-linear monotone relation between x and y and the following (x, y) points corresponding to different threshold values:

x	y
0.05	0.09
0.1	0.1
0.2	0.11
0.4	0.15
0.6	0.25
0.8	0.50
0.9	0.85

Furthermore, we have a binary class variable C with prior probability $\Pr(c) = 0.4$. We can now compute the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$ to be $y = 2/3 \cdot x$. The relation between the parameters x and y and the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$ are both displayed in Figure 17. When we use the function $y = 2/3 \cdot x$ to

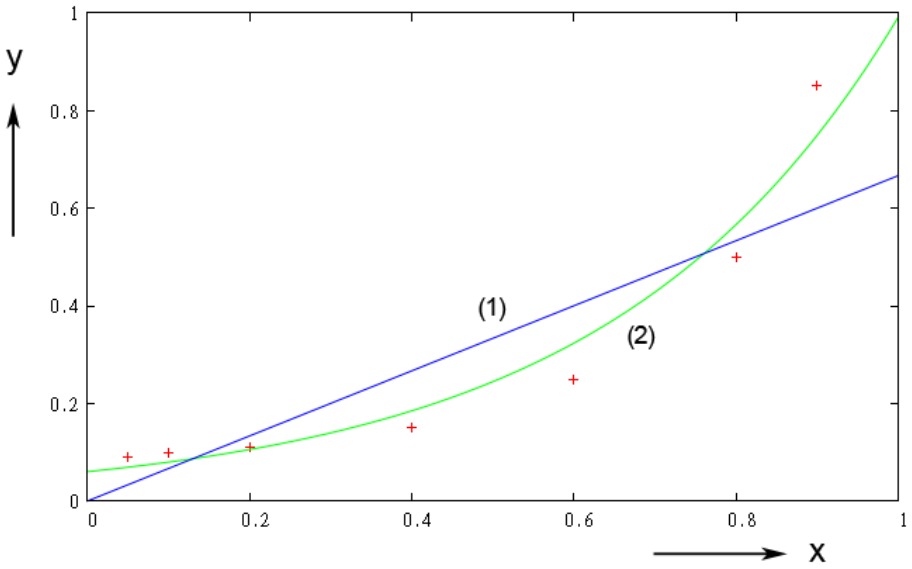


Figure 17: The intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$ (1) and the true (points) and approximate (2) relation between parameters x and y.

determine the y values corresponding to the smallest and largest x values from the table above, we find the values $1/30$ and $3/5$ respectively. These values are both lower than the y values from the table and thus we can conclude that our two (x, y) points from the table both lie above our intersection line. However, when we would have considered the (x, y) point where $x = 0.4$, the computed y value would have been $4/15$, which is larger than the corresponding y value from the table, resulting in a point (x, y) below the intersection line. We can conclude that, when we find two (x, y) points on different sides of the intersection line between $f_{\Pr(c|e)}(x, y)$ and $f_{\Pr(\bar{c}|e)}(x, y)$, the most likely class value can change by means of a different discretization for the feature variable E. When we can not find such two points, however, we can not claim that the most likely class value can not change.

Summarizing, we demonstrated how to use two-way sensitivity functions to determine whether discretization can be used in an NBN in order to change the most likely class value for a specific probability of interest. We defined some conditions that must be met for this change in most likely class value to take place. These conditions constrain the values for the variables a and b, describing the linear relationship between the parameters x and y, for different prior class probabilities. When the relation between x and y is not linear we proposed some heuristics to determine whether a change in most likely class value can be achieved.

6 Concluding observations

We have shown that when parameters in a naive Bayesian network change due to a change in discretization, a one-way sensitivity analysis of the affected parameters does not give an accurate and complete view anymore. We demonstrated the cause of this phenomenon to be a form of covariation between parameters specified for the variable which is newly discretized;

these covarying parameters come from the same CPT, but have a different conditioning context. We derived a condensed form of a two-way sensitivity function, which helps us to overcome this problem. When we can identify a relation between the two covarying parameters, such as we did for our domain example, we are able to simplify such a two-way sensitivity function into two one-way sensitivity functions. We have to note that the one-way sensitivity functions are mere approximations when the relationship between the parameters can only be estimated, e.g. by means of regression. Furthermore, we showed how we can use a two-way sensitivity analysis to determine whether a different discretization for a feature variable can lead to another most likely class value for a specific probability of interest. When in a two-way sensitivity analysis the two parameters x and y display a linear relationship $y = a \cdot x + b$, we can derive bounds for the values a, b and the binary class variable distribution; when all values are within these bounds we know that a different discretization for the corresponding feature variable can lead to another most likely class value. Additionally, we showed some heuristics to determine whether discretization can have a significant impact on the classifier's performance, which can also be used in situations where the relation between the two parameters is not linear.

We used our new condensed two-way sensitivity functions to study the effect of different binary discretizations for the $SCC1$ variable in our example domain. We were able to approximate the relation between the parameters $x = \theta(SCC1 < t \mid \text{tp})$ and $y = \theta(SCC1 < t \mid \text{fp})$ with a linear function, using linear regression, and we simplified our two-way functions into one-way functions. The goal of our analysis was to find a better discretization for the $SCC1$ variable, which could improve the naive Bayesian classifier that tries to distinguish between tp and fp alerts. However, we demonstrated that different discretizations of the $SCC1$ variable will have a negligible influence on the class variable distribution of the corresponding NBN. It looks like there is not enough information in the $SCC1$ variable to distinguish between tp and fp alerts, regardless of the discretization. Furthermore, we found that the same holds for the $SCC2$ and the DIM variable from our domain.

This thesis leaves opportunities for future research, because our findings mainly apply to binary variables and naive Bayesian networks. We would like to investigate whether we can generalize our results for NBNs containing class and feature variables with more than two intervals. Furthermore, it would be interesting to perform a similar analysis with Bayesian networks in general.

References

[1] V. M. H. Coupé and L. C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36:323–356, December 2002.

[2] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, 1995.

[3] K. B. Laskey. Sensitivity analysis for probability assessments in bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:901–909, 1995.

[4] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[5] S. Renooij. Bayesian network sensitivity to arc-removal. In P. Myllymaki, T. Roos, and T. Jaakkola, editors, *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pages 233 – 240, Helsinki, Finland, 2010. HIIT Publications.

[6] S. Renooij and L. C. van der Gaag. Evidence and scenario sensitivities in naive Bayesian classifiers. *International Journal of Approximate Reasoning*, 49:398–416, October 2008.

[7] M. M. Schrage, A. van IJzendoorn, and L. C. van der Gaag. Haskell ready to Dazzle the real world. In *Haskell '05: Proceedings of the 2005 ACM SIGPLAN workshop on Haskell*, pages 17–26. ACM Press, September 2005.

[8] W. Steeneveld, L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science*, 93(6):2559–68, 2010.

[9] L. C. van der Gaag, S. Renooij, and V. M. H. Coupé. Sensitivity analysis of probabilistic networks. In P. Lucas, J. Gámez, and A. Salmerón, editors, *Advances in Probabilistic Graphical Models*, volume 214 of *Studies in Fuzziness and Soft Computing*, pages 103–124. Springer Berlin / Heidelberg, 2007.