Evidence and Scenario Sensitivities in Naive Bayesian Classifiers

Silja Renooij

Linda C. van der Gaag

Technical Report UU-CS-2008-040
November 2008

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Evidence and Scenario Sensitivities in Naive Bayesian Classifiers

Silja Renooij Linda C. van der Gaag

Abstract

Empirical evidence shows that naive Bayesian classifiers perform quite well compared to more sophisticated classifiers, even in view of inaccuracies in their parameters. In this paper, we study the effects of such parameter inaccuracies by investigating the sensitivity functions of a naive Bayesian network. We show that, as a consequence of the network's independence properties, these sensitivity functions are highly constrained. We further investigate whether the patterns of sensitivity that follow from these functions support the observed robustness of naive Bayesian classifiers. In addition to standard sensitivities given available evidence, we also study the effect of parameter inaccuracies in view of scenarios of additional evidence. We show that standard sensitivity functions suffice to describe such scenario sensitivities.

Keywords: naive Bayesian classifiers, sensitivity, robustness

1 Introduction

Bayesian networks are often employed for classification purposes where an input instance described in terms of observable feature variables, is to be assigned to one of a number of possible output classes. The Bayesian network for this purpose computes, given the instance, the posterior probability distribution over the variable modelling these classes. The actual classifier then is a function that assigns a single class to the instance, based on the computed posterior distribution. Such classifiers are often built upon a naive Bayesian network, consisting of a single class variable and a number of feature variables, which are modelled as being mutually independent given the class variable. The parameter probabilities for such a network are generally estimated from data and inevitably are inaccurate.

Experiments have shown, time and again, that classifiers built on naive Bayesian networks are competitive with other, often more sophisticated classification models, regardless of the size and quality of the data set from which they are learned, e.g. [5, 9, 12]. Various aspects of the naive Bayesian classifier have been studied in attempts to explain these findings. For binary class variables, for example, it was shown that the commonly used winner-takes-all rule, which assigns an instance to a class with highest posterior probability according to the underlying Bayesian network, contributes to the naive classifier's success [5]. The observed robustness was also attributed to the independence properties of the classifier. It was shown, for example, that naive Bayesian classifiers perform well not only for completely independent feature variables, but also for functionally dependent ones [5, 12]. We note that most of the favourable experimental reports on naive Bayesian classifiers are based on the assumption of a binary class variable with a rather uniform prior probability distribution over its values.

In this paper, we follow up on the observation that, apparently, inaccuracies in the parameter probabilities of the underlying naive Bayesian network do not significantly affect

the performance of the classifier. We employ sensitivity-analysis techniques to study the effects of parameter variation on the posterior probability distributions computed from a naive Bayesian network, and thereby contribute further corroboration for the observed robustness of this type of classifier. We would like to note that sensitivity analysis has been applied before in the context of naive Bayesian classifiers, as a means of providing bounds on the amount of parameter variation that is allowed without changing, for any of the possible input instances, the most likely value of a binary class variable [3]. These bounds are useful for establishing which parameter upon variation can never change the classifier's output, regardless of the entered evidence. The computation of these bounds, however, requires either explicit enumeration over all possible instances or a conversion of the naive Bayesian classifier into an ordered decision diagram. We extend on these earlier results by studying the mathematical functions that describe the sensitivity of a posterior probability of interest computed from the naive Bayesian network, to variation of a parameter's value; these functions are termed sensitivity functions [1, 4]. We show that the independence assumptions underlying a naive Bayesian network constrain these sensitivity functions to such an extent that they can be established exactly from very limited information from the network at hand. In addition, we study the sensitivity properties that follow from the constrained functions and argue how these properties support the observed robustness of naive Bayesian classifiers.

In this paper, we also introduce the novel notion of scenario sensitivity, which we use for further studying a classifier's robustness. For classification problems, it is often assumed that evidence is available for every single feature variable. In numerous application domains, however, this assumption may not be realistic, especially not for domains in which evidence is gathered selectively in a stepwise manner. The question then arises how much impact further evidence could have on the computed posterior probability distributions and how sensitive this impact is to inaccuracies in the network's parameters. We introduce the notion of scenario sensitivity to capture the latter type of sensitivity and show that the effects of parameter variation in view of scenarios of additional evidence can be established efficiently for naive Bayesian networks.

The paper is organised as follows. In Section 2, we present some preliminaries on sensitivity functions and their associated sensitivity properties. In Section 3, we establish the functional form of the sensitivity functions for a naive Bayesian network in terms of variation of the parameter probabilities of the class variable, and address the ensuing sensitivity properties. Section 4 addresses the sensitivity functions and associated properties that result from variation of the parameter probabilities of the feature variables. In Section 5 we introduce the notion of scenario sensitivity and show that it can be established from standard sensitivity functions. The paper ends with our concluding observations in Section 6.

2 Preliminaries and Notation

A Bayesian network essentially is a compact representation of a joint probability distribution Pr over a set of stochastic variables V [10]. 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 $\theta(V \mid \pi(V))$ that capture the strength of the relationship between a variable V and its parents $\pi(V)$. From a Bayesian network, any prior or posterior probability of interest over its variables can be computed. In this paper, we focus more specifically on *naive Bayesian networks*, in which the digraph modelling the

interrelationships between the variables has a restricted topology. The digraph of such a network is composed of nodes $\{C\} \cup E$ with $E = \{E_1, \ldots, E_n\}$, $n \geq 2$, and arcs (C, E_i) for all $i = 1, \ldots, n$; the variable C is called the *class variable* and the variables E_i are termed *feature variables*. The restricted topology of the digraph of a naive Bayesian network captures conditional independence of any two feature variables given the class variable. Although any probability can be computed from a naive Bayesian network, 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. In this paper, we focus primarily on sensitivity analyses in which just one parameter is being varied; such an analysis is termed a one-way sensitivity analysis. The effects of the parameter variation are captured by a simple mathematical function, called a *sensitivity function*. Before reviewing the functional form of such a sensitivity function, we observe that upon varying a particular parameter probability, the parameters pertaining to the same conditional distribution should be co-varied to ensure that their sum remains one. The well-known scheme of proportional co-variation is often used for this purpose[1] as it has been shown to result in the smallest change in the output distribution [2]. Under this scheme, any one-way sensitivity function is a quotient of two linear functions in the parameter under study [1, 4]. More formally, upon varying a single parameter probability x, the function $f_{\Pr(c|e)}(x)$ that expresses the output probability of interest $\Pr(c \mid e)$ in terms of x takes the form

$$f_{\Pr(c|e)}(x) \;=\; \frac{f_{\Pr(c,\,e)}(x)}{f_{\Pr(e)}(x)} \;=\; \frac{a \cdot x + b}{g \cdot x + h}$$

where the constants a, b, g, h are built from the non-varied parameters from the network under study. The four constants are derived analytically in [4]; feasible algorithms for their computation are available from [4, 7]. In the sequel, instead of $f_{\Pr(c|e)}(x)$ we will often write $f_c(x)$ or $f(x)$ 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.

A one-way sensitivity function $f(x)$ can take one of three general forms. The function is *linear* if the probability of interest is a prior probability rather than a posterior probability, or if the probability of the entered evidence is unaffected by the parameter variation; note that in the latter case we have that $\Pr(e)$ is a constant and hence $g = 0$. If the probability $\Pr(e)$ of the evidence equals 0 whenever $x = 0$, in which case we have that $h = 0$, then it is readily shown that the same must hold for the marginal probability $\Pr(c,\, e)$, that is, we must have that $b = 0$; the sensitivity function then reduces to a *constant*. In all other cases the sensitivity function is a fragment of a *rectangular hyperbola*, which takes the general form

$$f(x) \;=\; \frac{r}{x - s} + t \;=\; \frac{t \cdot x + r - s \cdot t}{x - s}, \quad \text{with} \quad s = -\frac{h}{g},\; t = \frac{a}{g}, \quad \text{and} \quad r = \frac{b}{g} + s \cdot t$$

[1] When a parameter probability θ is varied from the value θ_{old} to the value θ_{new}, each parameter $\theta' \neq \theta$ from the same distribution is varied from θ'_{old} to θ'_{new}, where $\theta'_{\text{new}} = \theta'_{\text{old}} \cdot \dfrac{1 - \theta_{\text{new}}}{1 - \theta_{\text{old}}}$.

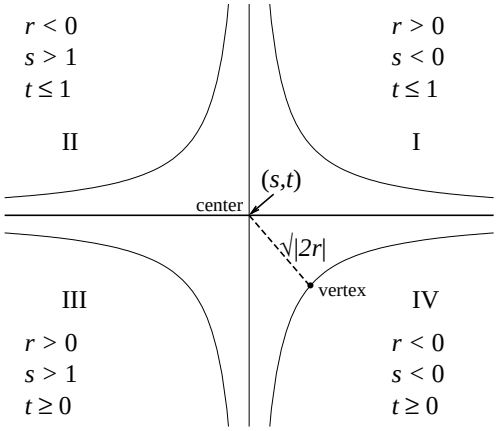

Figure 1: Two rectangular hyperbolas with branches in the Ist and IIIrd quadrants relative to the hyperbola's center, and in the IInd and IVth quadrants, respectively; the constraints on the constants r, s and t are specific for sensitivity functions.

with the constants a, b, g, h as above. In the remainder of the paper, we focus on this last type of function 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, t); Figure 1 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 $s < 0$ or $s > 1$. From these observations we conclude that a hyperbolic sensitivity function is a fragment of just one of the four possible branches shown in Figure 1.

From a sensitivity function, various properties can be computed that serve to summarise the effects of parameter variation. Here we briefly review the properties of *sensitivity value* [8] and of *admissible deviation* [13]. The *sensitivity value* for a parameter x is the absolute value $|\partial f/\partial x\,(x_0)|$ of the first derivative of the sensitivity function $f(x)$ at the parameter's original value x_0. It describes the effect of an infinitesimally small shift in the parameter on the output probability of interest. In essence, the larger the sensitivity value for a parameter is, the less robust the output probability of interest will be to inaccuracies in the parameter. We would like to note that the impact of a larger shift in a parameter's value is strongly dependent upon the location of the *vertex* of the sensitivity function, that is, of the point where $|\partial f/\partial x\,(x)| = 1$. The vertex can lie within the unit window, or to its left or right. A vertex that lies within the unit window basically marks the transition from parameter values with a large sensitivity value to parameter values with a small sensitivity value, or vice versa. A parameter with a small sensitivity value can thus have larger effects than its sensitivity value suggests, if it lies in the proximity of the vertex, that is, if its original value x_0 is close to the vertex' x-value. If the posterior probabilities computed from a Bayesian network are used for establishing the most likely value of an output variable, it is the effect of parameter variation on the output value that is of interest. For a parameter with an original value of x_0, the *admissible deviation* is a pair (α, β), where α is the amount of variation allowed to

values smaller than x_0 without changing the most likely output value and β is the amount of variation allowed to larger values; the symbols $\leftarrow$ and $\rightarrow$ are used to indicate that variation is allowed to the boundaries of the unit window. The larger the admissible deviation for a parameter is, therefore, the more robust the output value will be to inaccuracies in this parameter.

3 Sensitivity to Class Parameters

Upon being subjected to a sensitivity analysis, the independence properties of a naive Bayesian network strongly constrain the general form of the resulting sensitivity functions. In fact, given just limited information from the network, the exact functions can be readily established for each class value and each parameter probability. In this section we derive the sensitivity functions that describe an output probability of interest as a function of a parameter for the class variable. We detail the sensitivity properties that follow from these functions and discuss their possible effects on the robustness of naive Bayesian classifiers. In Section 4, we will address the sensitivity functions for parameters for the feature variables in a similar fashion.

3.1 Functional forms

The following proposition states the functional form of any (hyperbolic) sensitivity function that describes an output probability of a naive Bayesian network in terms of a single parameter $x = \theta(c')$, associated with a specific value c' of the class variable C. The proposition more specifically shows that such a function is highly constrained and can in fact take only one of two forms. Proofs of this and subsequent propositions are presented in Appendix A.

Proposition 1. *Let $x = \theta(c')$ be a parameter probability pertaining to the value c' of the class variable C, and let x_0 be its original value. Let $\Pr(c \mid e)$ be an output probability of interest with the original value p_0, and let p_0' be the original value of $\Pr(c' \mid e)$. Then, the sensitivity function $f_{\Pr(c|e)}(x)$ has the following form:*

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

in which the value s, defining the function's vertical asymptote, equals

$$s = \frac{(1-p_0')\cdot x_0}{x_0 - p_0'}$$

The value t, defining the horizontal asymptote of the sensitivity function, equals

$$t = \begin{cases} 1-s & \text{if } c = c' \\ \dfrac{p_0}{1-p_0'} \cdot s & \text{otherwise} \end{cases}$$

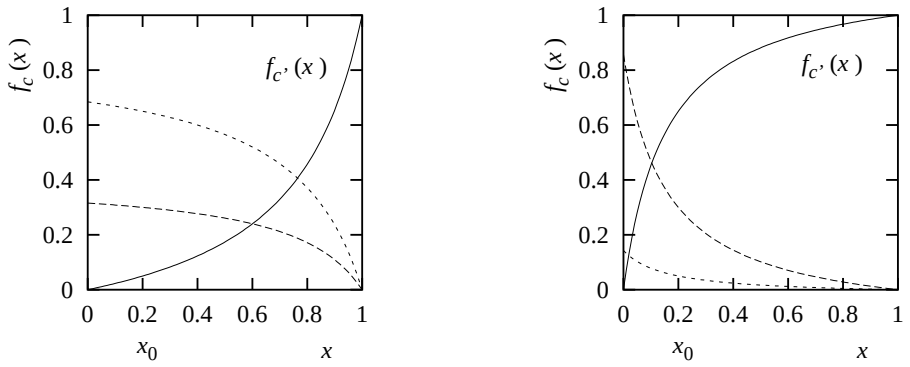

Figure 2: Example sensitivity functions for a class parameter $x = \theta(c')$ with an original value $x_0 = 0.2$; the original posterior of interest is $p_0' = 0.05$ (*left*) and $p_0' = 0.65$ (*right*), respectively.

We note that the above proposition pertains to a single output probability of interest $\Pr(c \mid e)$. Since the choice of class value c, however, is arbitrary, the proposition holds for any value of C. The original value p_0 of the output probability of interest obviously depends on the value of c under consideration and, as a result, so does the actual value of the horizontal asymptote t.

From the above proposition, we observe that the sensitivity function $f_{\Pr(c'|e)}(x)$ pertaining to the class value whose parameter probability is being varied, includes the points $f_{c'}(0) = 0$ and $f_{c'}(1) = 1$ from the unit window; the function thus is a fragment of either a IInd-quadrant or a IVth-quadrant hyperbola branch. The sensitivity functions $f_{\Pr(c|e)}(x)$, $c \neq c'$, pertaining to the other values of the class variable then are fragments of either Ist- or IIIrd-quadrant hyperbola branches. Figure 2 illustrates the two possible situations. The sensitivity function $f_{\Pr(c'|e)}(x)$ being a IInd-quadrant function corresponds with a vertical asymptote to the right of the unit window, that is, with $s > 1$ and hence with $x_0 > p_0'$; the sensitivity functions pertaining to the other values of C then are IIIrd-quadrant functions. The IVth- and Ist-quadrant combination of functions corresponds with $x_0 < p_0'$. To intuitively explain why the function $f_{\Pr(c'|e)}(x)$ pertaining to the class value c' has a different shape from the other functions, we observe that since this function must include the two points $f_{c'}(0) = 0$ and $f_{c'}(1) = 1$ from the unit window, it necessarily is an increasing function. Furthermore, because the function $f_{\Pr(c'|e)}(x)$ includes the point $f_{c'}(1) = 1$, all functions $f_{\Pr(c|e)}(x)$ with $c \neq c'$ must include $f_c(1) = 0$. Given their highly constrained functional form, these functions moreover have essentially the same shape. In addition, the function values of the functions have to sum to one for $x = 0$, from which we have that they are necessarily decreasing functions.

We illustrate the functional form of the sensitivity functions derived above with an example. The example demonstrates specifically that as a result of their constrained form, any sensitivity function can be established from very limited information.

Example 1. We consider a naive Bayesian network with a class variable S modelling the possible stages I, IIA, IIB, III, IVA and IVB of cancer of the oesophagus. The parameter probabilities for this class variable are:

S	I	IIA	IIB	III	IVA	IVB
$\theta(S)$	0.04	0.31	0.04	0.23	**0.10**	0.28

The feature variables of the network capture the results from diagnostic tests. For a particular patient, the available findings are summarised in the input instance e, giving rise to the following posterior probability distribution over the class variable:

S	I	IIA	IIB	III	IVA	IVB
$\Pr(S \mid e)$	0.01	0.19	0.01	0.07	**0.61**	0.11

Suppose that we are interested in the effect of inaccuracies in the parameter probability $x = \theta(S = \text{IVA})$ on the posterior probabilities $\Pr(S \mid e)$ computed for our patient. The effect is captured by six functions with the same vertical asymptote, whose value s is readily established: since the original parameter value is $x_0 = 0.10$, and for this patient the original posterior for stage IVA is 0.61, we find that $s = (1 - 0.61) \cdot 0.10/(0.10 - 0.61) = -0.08$. The sensitivity function $f_{\text{IVA}}(x)$ therefore is a IVth-quadrant hyperbola branch; the functions for the other stages are Ist-quadrant branches. Without performing any further computations, we establish that

$$f_{\text{IVA}}(x) = \frac{1.08 \cdot x}{x + 0.08} \quad \text{and} \quad f_S(x) = \frac{\Pr(S \mid e)}{1 - 0.61} \cdot \frac{-0.08 \cdot (x - 1)}{x + 0.08} \quad \text{for any } S \neq \text{IVA}$$ □

From the above considerations, we have that the sensitivity functions resulting from a one-way sensitivity analysis for a class parameter, are highly constrained. In fact, from Proposition 1 we observe that the functions are fully determined by just the original value x_0 for the class parameter being varied and the original posterior probability distribution over the output variable of interest. The exact functions as a consequence are readily computed, requiring just a single network propagation to establish the posterior class distribution.

3.2 Sensitivity properties

From the sensitivity functions derived above, any sensitivity property pertaining to a class parameter can be computed. We study the properties of sensitivity value and admissible deviation.

Sensitivity value From the sensitivity function $f_{\Pr(c|e)}(x)$ expressing the output probability $\Pr(c \mid e)$ in terms of a class parameter $x = \theta(c')$, the associated sensitivity value is readily established:

$$\left| \frac{\partial f_{\Pr(c|e)}}{\partial x}(x_0) \right| = \begin{cases} (1 - p_0') \cdot \dfrac{p_0'}{(1 - x_0) \cdot x_0} & \text{if } c = c' \\ p_0 \cdot \dfrac{p_0'}{(1 - x_0) \cdot x_0} & \text{otherwise} \end{cases}$$

where p_0 again is the original value of $\Pr(c \mid e)$ and p_0' is the original value of $\Pr(c' \mid e)$; x_0 is the parameter's original value as before. From $1 - p_0' \geq p_0$ for any value c of the class variable, we observe that the highest sensitivity value for the parameter x is obtained when the output probability of interest pertains to the class value c' whose parameter is being varied. The sensitivity value found then in fact matches the upper bound on sensitivity values for

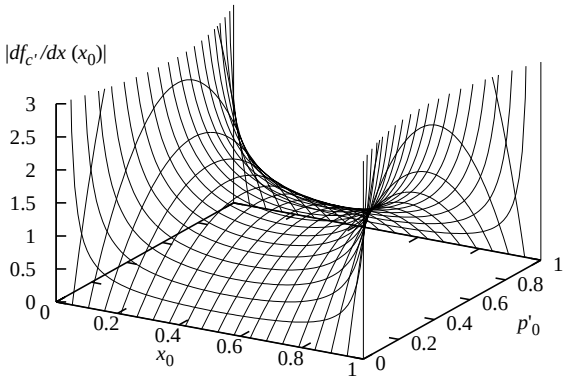

Figure 3: The sensitivity value for a class parameter $x = \theta(c')$ and sensitivity function $f_{\Pr(c'|e)}(x)$, as a function of the original parameter value x_0 and the original posterior p'_0.

sensitivity functions in general [11]. Note that for a binary class variable, we would have that $1 - p'_0 = p_0$ and the sensitivity values for the two class values would be the same. The sensitivity values computed from the function $f_{\Pr(c'|e)}(x)$ for different combinations of values for x_0 and p'_0 are depicted in Figure 3. The figure shows that large sensitivity values can be found only for rather extreme parameter values in combination with less extreme output probabilities. We further recall that, despite their small sensitivity values, also parameters with an original value close to the x-value of the sensitivity function's vertex may show significant effects upon variation. For class parameters, the sensitivity function $f_{\Pr(c'|e)}(x)$ always has its vertex within the unit window; the vertex in fact lies on the line $x = 1 - f(x)$. If the vertical asymptote of the sensitivity function lies quite close to the unit window and the vertex in addition is not too far from the asymptote, then a parameter with an original value in the proximity of the vertex will show considerable impact on the output probability if it is varied further to the nearby extreme.

From the above considerations, we have that in naive Bayesian networks the output probability of interest will be sensitive to variation in a class parameter only if the class value under consideration either occurs rather seldomly or quite frequently in the domain of application and the instance at hand does not support the (un)likelihood of this class value. As long as the class parameters of a naive Bayesian network are not highly unbalanced, therefore, will the probability with which an instance is predicted to belong to a particular class be quite insensitive to parameter variation. We note that it is not uncommon to find class variables with such distributions in the domains in which naive Bayesian classifiers are being applied. Yet, also the output probabilities computed from a naive Bayesian network in which one or more class values have rather small prior probabilities, will be quite robust as long as the posterior probabilities computed for these classes are quite extreme for all possible instances.

Admissible deviation In view of establishing the most likely class value for an instance, the property of admissible deviation is of interest. We recall that the admissible deviation for a parameter gives the amount of variation that is allowed in its original value before the most likely class value changes. The following proposition gives the admissible deviation for a class parameter in a naive Bayesian network. The proposition more specifically shows that

the most likely class value changes *exactly once* upon varying such a parameter.

Proposition 2. *Let $x = \theta(c')$ be a parameter probability pertaining to the value c' of the class variable C, and let x_0 be its original value. Let p_0' be the original value of $\Pr(c' \mid e)$, let $p_0^\intercal = \text{argmax}_{c \neq c'}\{\Pr(c \mid e)\}$, and let $c^\intercal$ be a value of C for which $\Pr(c^\intercal \mid e) = p_0^\intercal$. Then,*

$$f_{\Pr(c|e)}(x) \leq f_{\Pr(c^\intercal|e)}(x) \quad \textit{for all } c \neq c'$$

Furthermore, the admissible deviation for x is:

$$(\alpha, \beta) = \begin{cases} (x_0 - x_m, \rightarrow) & \textit{if } \; p_0^\intercal < p_0' \\ (\leftarrow, x_m - x_0) & \textit{if } \; p_0^\intercal > p_0' \\ (0, \rightarrow) \textit{ or } (\leftarrow, 0) & \textit{otherwise} \end{cases}$$

$$\textit{where } \; x_m = \frac{p_0^\intercal \cdot x_0}{(1 - x_0) \cdot p_0' + p_0^\intercal \cdot x_0}$$

From the above proposition, we observe that any class parameter can be varied to the boundary of the unit window in one of the two possible directions without changing the most likely class value, as indicated by the $\leftarrow$ and $\rightarrow$ symbols. Upon varying any such parameter in the other direction, the most likely class value changes exactly once. Note that Figure 2 supports these observations. For the case where $p_0^\intercal = p_0'$, we note that, given the original value x_0 of the parameter under study, the two class values c' and $c^\intercal$ are equally likely. The boundary to which the parameter can be varied without inducing a change then depends on which of the two class values is designated the unique most likely class by the network's associated classification function. We further observe that the smaller the difference is between the two posterior probabilities $p_0^\intercal$ and p_0', the smaller the admissible deviation for the parameter is and the less robust the class value returned by the classifier will be.

As a special case of the above proposition, we consider a uniformly distributed binary class variable, that is, we consider the case where $x_0 = 0.5$ and $p_0^\intercal = 1 - p_0'$; note that in this case we find for the value x_m at which the two sensitivity functions intersect, that $x_m = 1 - p_0'$. The admissible deviation for the class parameter $x = \theta(c')$ then equals $(p_0' - 0.5, \rightarrow)$ if $p_0^\intercal < p_0'$, that is, if the instance at hand supports the class value c' whose parameter probability is being varied. The admissible deviation equals $(\leftarrow, 0.5 - p_0')$ if $p_0^\intercal > p_0'$, that is, if the instance points to the other value of the class variable. We note that the more extreme the original value p_0' of the posterior probability of interest is, the larger the admissible deviation will be and the less impact variation of a class parameter can have on the most likely class value. Further note that for non-binary class variables, the value of x_m in the admissible deviation may be less extreme, which may result in smaller admissible deviations and a less robust output class.

Example 2. We consider again the naive Bayesian network and the associated patient information from Example 1. Suppose that we are interested in the effects of inaccuracies in the parameter probability $x = \theta(S = \text{IVA})$, with an original value of $x_0 = 0.10$, on the most likely class value established for our patient. We recall that, with the parameter's original value, stage IVA is the most likely stage for the patient, with a probability of 0.61; the second most likely stage is stage IIA, with a probability of 0.19. Using these probabilities, we find for the value x_m from Proposition 2 that $x_m = 0.19 \cdot 0.10/(0.90 \cdot 0.61 - 0.19 \cdot 0.10) = 0.03$. The admissible deviation for the parameter under study thus is $(0.07, \rightarrow)$. This admissible

deviation indicates that the parameter can be varied from 0.10 to 1.00 without inducing a change in the most likely stage for the patient. The parameter can also be varied to smaller values, but the most likely stage will change from IVA to IIA if the parameter adopts a value smaller than 0.03. Note that the most likely stage cannot change into any other value upon varying the parameter. Further note that, although in absolute terms only a small shift is allowed to smaller parameter values, the admissible deviation is quite large relative to the parameter's original value.

Now suppose that we were to use a two-valued rather than a six-valued variable class variable. We define for this purpose the new variable *Operable*, of which the value 'yes' coincides with stages I, IIA and IIB, and the value 'no' captures the stages III, IVA and IVB. Given the patient's available information, the posteriors $\Pr(\mathit{Operable} = \text{yes} \mid e) = 0.21$ and $\Pr(\mathit{Operable} = \text{no} \mid e) = 0.79$ are computed; the value 'no' thus is the most likely class value for the patient. Suppose that we are interested in the effects of inaccuracies in the parameter $x = \theta(\mathit{Operable} = \text{no})$, with an original value of $x_0 = 0.61$, on the most likely value of the output variable. We find that the sensitivity functions associated with the two values of the class variable intersect at $x_m = 0.29$. The admissible deviation for parameter x thus equals $(0.32, \rightarrow)$. This admissible deviation indicates that the parameter can be varied from 0.61 to 1.00 without inducing a change in the most likely class value. The parameter can also be varied to values smaller than 0.61, but the most likely value will change from 'no' to 'yes' if the parameter adopts a value smaller than 0.29. Note that the parameter can thus be varied to approximately half its original value. □

From the above considerations, we conclude that the most likely class value established from a naive Bayesian network will be quite sensitive to inaccuracies in the network's class parameters if the output probabilities for the class variable are more or less uniformly distributed. More specifically, the most likely class value will not be very robust to variation in the class parameters if it has approximately the same posterior probability as the runner-up value. The classification performance of a naive Bayesian classifier will thus be quite robust if the majority of presented instances result in a single rather likely class value. In fact, it will be robust as long as the majority of instances belong to the a priori most likely class, which we would expect if the classifier is sufficiently tailored to the domain of application.

4 Sensitivity to Feature Parameters

In this section we derive for a naive Bayesian network the sensitivity functions that describe an output probability of interest as a function of a parameter for a feature variable. We show that also for feature parameters the exact sensitivity functions can be readily established given just limited information from the network. We further detail the sensitivity properties following from the functions and discuss their possible effects on the robustness of naive Bayesian classifiers.

4.1 Functional forms

The following proposition states the functional form of any (hyperbolic) sensitivity function that describes an output probability of a naive Bayesian network in terms of a single feature parameter $x = \theta(e'_v \mid c')$, where e'_v denotes a value of the feature variable E_v and c' is a value of the class variable. The proposition shows that the function again is highly constrained; in

fact, for any class value and any feature parameter, only one of four functional forms can result.

Proposition 3. *Let E_v be a feature variable and let e_v be its value in the instance e. Let $x = \theta(e'_v \mid c')$ be a parameter probability pertaining to the value e'_v of E_v and the class value c', and let x_0 be its original value. Let $\Pr(c \mid e)$ be an output probability of interest with the original value p_0, and let p'_0 be the original value of $\Pr(c' \mid e)$. Then, the sensitivity function $f_{\Pr(c|e)}(x)$ has one of the following forms:*

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

in which the value s, defining the function's vertical asymptote, equals

$$s = \begin{cases} x_0 - \dfrac{x_0}{p'_0} & \text{if } e_v = e'_v \\ x_0 + \dfrac{(1-x_0)}{p'_0} & \text{otherwise} \end{cases}$$

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

$$t = \begin{cases} 1 & \text{if } c = c' \\ 0 & \text{otherwise} \end{cases}$$

We consider again the possible locations of the sensitivity functions for a feature parameter under study. For the case where $e_v = e'_v$, we find for the value $s = x_0 - x_0/p'_0$ of the vertical asymptote that $s < 0$. The asymptote thus lies to the left of the unit window. Since the sensitivity function $f_{\Pr(c'|e)}(x)$ pertaining to the class value c' further includes the point $f_{c'}(0) = 0$ from the unit window, we conclude that it is a fragment of a IVth-quadrant hyperbola branch. The function $f_{\Pr(c|e)}(x)$ for any class value $c \neq c'$ is a fragment of a Ist-quadrant branch. For the case where $e_v \neq e'_v$, we find that $s > 1$; we then find IIIrd- and IInd-quadrant branches, respectively. Figure 4 illustrates the two possible situations. To intuitively explain why the function $f_{\Pr(c'|e)}(x)$ again has a different shape from the other functions, we observe that varying a feature parameter $\theta(e_i \mid c')$ given a particular value c' of the class variable has a direct effect on the posterior probability $\Pr(c' \mid e)$ of this class value c' only; the probabilities $\Pr(c \mid e)$, $c \neq c'$, for the other values of the class variable are affected only indirectly to ensure that the distribution over the class variable sums to one. From the highly constrained form of the functions, moreover, we have that all functions $f_{\Pr(c|e)}(x)$, $c \neq c'$, have the same shape. The shape of the function $f_{\Pr(c'|e)}(x)$ therefore must be deviant.

We again illustrate the functional form of the sensitivity functions derived above with an example. The example once more demonstrates that as a result of their constrained form, any sensitivity function can be established from very limited information.

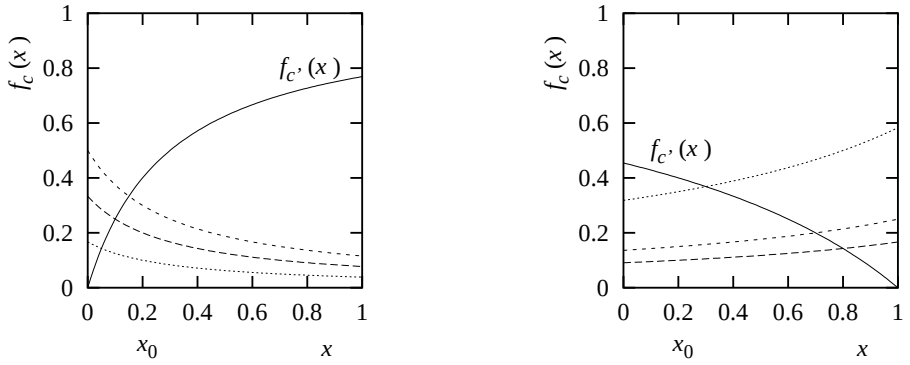

Figure 4: Example sensitivity functions for a feature parameter $x = \theta(e'_v \mid c')$ with original value $x_0 = 0.2$; the original posterior of interest is $p'_0 = 0.4$, with $e_v = e'_v$ (*left*) and $e_v \neq e'_v$ (*right*), respectively.

Example 3. We again consider the naive Bayesian network and the patient information from Example 1. We further consider the feature variable *CT-loco*, modelling the presence or absence of loco-regional metastases as suggested by a CT scan of the patient's thorax. The network includes the following parameter probabilities for this variable:

$\theta(CT\text{-}loco \mid S)$	I	IIA	IIB	III	IVA	IVB
$CT\text{-}loco$ = yes	0.02	0.02	0.48	0.48	0.48	0.27
no	0.98	0.98	0.52	0.52	**0.52**	0.73

The posterior probability distribution $\Pr(S \mid e)$ computed over the class variable given the available findings for our patient, who shows no signs of loco-regional metastases, are found in Example 1. Now suppose that we are interested in the effect of inaccuracies in the parameter probability $x = \theta(CT\text{-}loco = \text{no} \mid S = \text{IVA})$ on these posterior probabilities. The effect is captured by six sensitivity functions with the same vertical asymptote, whose value s is readily established: since the original value of the parameter equals 0.52 and the original posterior probability of stage IVA for the patient is 0.61, we find that $s = 0.52 - 0.52/0.61 = -0.33$. The sensitivity function $f_{\text{IVA}}(x)$ therefore is a IVth-quadrant hyperbola branch; the functions for the other stages are Ist-quadrant branches. Note that for the complement of the parameter x, the six sensitivity functions would all have their vertical asymptote at $s = 1.33$. Without performing any further computations, we establish that

$$f_{\text{IVA}}(x) = \frac{x}{x + 0.33} \quad \text{and} \quad f_S(x) = \Pr(S \mid e) \cdot \frac{0.85}{x + 0.33} \quad \text{for any } S \neq \text{IVA}$$ □

From the above considerations, we have that the sensitivity functions resulting from a one-way analysis for a feature parameter, are highly constrained. Just as the sensitivity functions for the class parameters, we find that the functions for the feature parameters are exactly determined by the original values for these parameters and the original posterior probability distribution for the output variable of interest. Computing the exact functions as a consequence again requires just a single network propagation to establish the posterior class distribution.

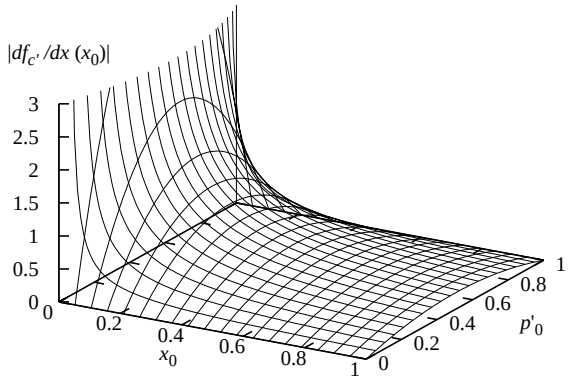

Figure 5: The sensitivity value for a feature parameter $x = \theta(e'_v \mid c')$ and sensitivity function $f_{\Pr(c'|e)}(x)$, where the instance e includes $e_v = e'_v$, as a function of the original parameter value x_0 and the original posterior p'_0.

4.2 Sensitivity properties

From the sensitivity functions derived above, any sensitivity property pertaining to a network's feature parameters can be computed. We again study the properties of sensitivity value and admissible deviation.

Sensitivity value From the sensitivity function $f_{\Pr(c|e)}(x)$ expressing the output probability $\Pr(c \mid e)$ in terms of a feature parameter $x = \theta(e'_v \mid c')$, the associated sensitivity value is readily established: if the observed instance e includes the value $e_v = e'_v$, we find that

$$\left|\frac{\partial f_{\Pr(c|e)}}{\partial x}(x_0)\right| = \begin{cases} (1 - p'_0) \cdot \dfrac{p'_0}{x_0} & \text{if } c = c' \\ p_0 \cdot \dfrac{p'_0}{x_0} & \text{otherwise} \end{cases}$$

where p_0 again is the original value of $\Pr(c \mid e)$ and p'_0 is the original value of $\Pr(c' \mid e)$; x_0 is the parameter's original value as before. If the instance e includes another observed value $e_v \neq e'_v$, then a similar result is found by replacing x_0 by $1 - x_0$. From $1 - p'_0 \geq p_0$ for any value c of the class variable, we find that the highest sensitivity value for the parameter x again is obtained when the output probability of interest pertains to the same class value c' as the feature parameter being varied, regardless of the value that is observed for the feature variable E_v. For the case where the instance e includes the value $e_v = e'_v$, Figure 5 depicts the sensitivity value of $f_{\Pr(c'|e)}(x)$ for different combinations of values for x_0 and p'_0. The figure shows that large sensitivity values can only be found if the original value x_0 for the parameter under study is quite small and the original posterior probability p'_0 is less extreme; more specifically, we have that $\left|\partial f_{\Pr(c'|e)}/\partial x\,(x_0)\right| > 1$ if and only if $x_0 < p'_0 \cdot (1 - p'_0) \leq 0.25$. For the case where $e_v \neq e'_v$, large sensitivity values are found only if x_0 is larger than 0.75 and p'_0 is non-extreme.

From the above considerations we have that large sensitivity values can only be found for feature variables that, given a particular class, have a rather unlikely value and whose unlikely value is found in an instance that does not strongly support this class; this property

was already noticed before for binary observable variables in Bayesian networks in general [14]. As long as the feature parameters given each class in a naive Bayesian network are not highly unbalanced, therefore, will the probability with which an instance is predicted to belong to a particular class be quite insensitive to parameter variation. Yet, also the output probabilities computed from a naive Bayesian network in which one or more features given a particular class have rather small prior probabilities, will be quite robust as long as the posterior probabilities computed for this class are rather extreme for all possible instances.

Admissible deviation The following proposition states the admissible deviation for a feature parameter in a naive Bayesian network. This admissible deviation gives the amount of variation that is allowed in the parameter's value before the most likely class changes. The proposition more specifically shows that in a naive Bayesian network, the most likely class value can change *at most once* upon varying a feature parameter.

Proposition 4. *Let E_v be a feature variable and let e_v be its value in the instance e. Let $x = \theta(e'_v \mid c')$ be a parameter probability pertaining to the value e'_v of E_v and the class value c', and let x_0 be its original value. Let p'_0 be the original value of $\Pr(c' \mid e)$; in addition, let $p^{\top}_0 = \text{argmax}_{c \neq c'}\{\Pr(c \mid e)\}$ and let $c^{\top}$ be a value of C for which $\Pr(c^{\top} \mid e) = p^{\top}_0$. Then,*

$$f_{\Pr(c|e)}(x) \leq f_{\Pr(c^{\top}|e)}(x) \quad \textit{for all } c \neq c'$$

Furthermore, the admissible deviation for x is:

$$(\alpha, \beta) = \begin{cases} (0, \rightarrow) \text{ or } (\leftarrow, 0) & \text{if } p^{\top}_0 = p'_0 \\ (x_0 - x_m, \rightarrow) & \text{if } e_v = e'_v \text{ and } p^{\top}_0 < p'_0, \\ & \quad \text{or if } e_v \neq e'_v,\ p^{\top}_0 > p'_0 \text{ and } 1 - x_0 < p'_0/p^{\top}_0 \\ (\leftarrow, x_m - x_0) & \text{if } e_v = e'_v,\ p^{\top}_0 > p'_0 \text{ and } x_0 < p'_0/p^{\top}_0, \\ & \quad \text{or if } e_v \neq e'_v \text{ and } p^{\top}_0 < p'_0 \\ (\leftarrow, \rightarrow) & \text{otherwise} \end{cases}$$

where

$$x_m = \begin{cases} p^{\top}_0 \cdot \dfrac{x_0}{p'_0} & \text{if } e_v = e'_v \\ 1 - p^{\top}_0 \cdot \dfrac{1 - x_0}{p'_0} & \text{if } e_v \neq e'_v \end{cases}$$

From the above proposition, we observe that any feature parameter can be varied to the boundary of the unit window in at least one direction without changing the most likely class value. Upon varying any such parameter in the other direction, the most likely class value can change at most once. Figures 4 and 6 support these observations. For the case where $p^{\top}_0 = p'_0$, we again note that the boundary to which the parameter can be varied without inducing a change depends on which of the two class values c' and $c^{\top}$ is designated the unique most likely class by the network's classification function. We further observe that the smaller the difference between the posterior probabilities $p^{\top}_0$ and p'_0 is, the smaller the admissible deviation for the parameter is and the less robust the returned class value will be.

We consider a feature parameter $x = \theta(e'_v \mid c')$ and an instance e in which $e_v = e'_v$ has been observed; similar arguments hold for $e_v \neq e'_v$. We suppose that the class value c' is not the

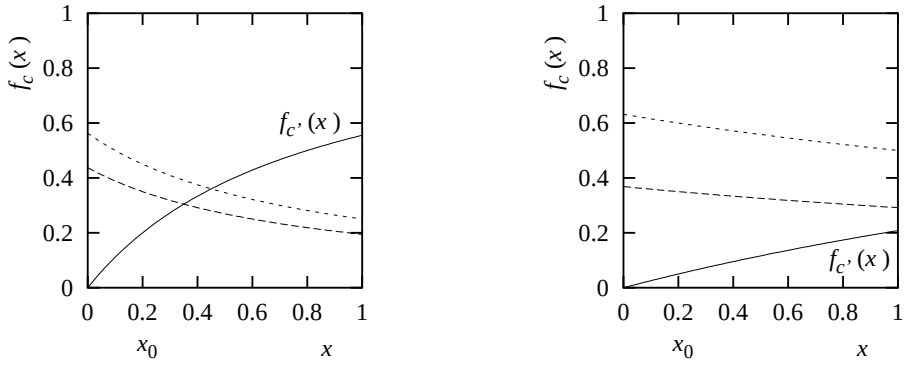

Figure 6: Example sensitivity functions for a feature parameter $x = \theta(e'_v \mid c')$ with original value $x_0 = 0.2$; c' is not the most likely class value, and variation of the parameter either makes c' the most likely value (*left*) or does not change the most likely value (*right*).

most likely class value given the available evidence. If the original value x_0 of the parameter is rather small, we would not expect to find the feature value e'_v with class c'. When actually observed, therefore, the feature value does not support c'. If the other observations in the instance also do not support c', we expect that $p_0^\intercal \gg p'_0$ and a large admissible deviation can be found for the parameter. If the other observations from the instance do support the class value c', however, we would expect to find a larger p'_0 and hence a smaller admissible deviation. Figure 6 illustrates these observations. Now if, on the other hand, the original parameter value x_0 is relatively large, we would indeed expect to find the feature value e'_v with class c'. The actual observation of e'_v then supports c' and we expect to find a somewhat larger posterior probability p'_0 and a relatively small admissible deviation. A reverse argumentation holds for the case where c' *is* the most likely class value.

Example 4. We consider again the naive Bayesian network and the patient information from Examples 1 and 3. Suppose that we are once more interested in the effect of inaccuracies in the parameter probability $x = \theta(\mathit{CT\text{-}loco} = \text{no} \mid S = \text{IVA})$, with an original value of $x_0 = 0.52$, on the most likely class value established for our patient. We recall that, with the parameter's original value, stage IVA is the most likely stage for the patient, with a probability of 0.61; the second most likely stage is stage IIA, with a probability of 0.19. We further recall that for the patient no signs of loco-regional metastases were found on the CT scan. Using the above probabilities, we now find for the value x_m from Proposition 3 that $x_m = 0.19 \cdot 0.52/0.61 = 0.16$. The admissible deviation for the parameter under study thus is $(0.36, \rightarrow)$. This admissible deviation indicates that the parameter can be varied from 0.52 to 1.00 without inducing a change in the most likely stage of the patient's cancer. The parameter can also be varied to smaller values, but the most likely stage will change from IVA to IIA if the parameter adopts a value smaller than 0.16. Note that the parameter can thus be varied to approximately one-third of its original value. Further note that the most likely stage cannot change into any other value upon varying the parameter under study. □

From the above considerations, we conclude that the most likely class value established from a naive Bayesian network will be quite sensitive to inaccuracies in the network's feature parameters if the output probabilities for the class variable are more or less uniformly distributed. More specifically, the most likely class value will not be very robust to variation in the feature

parameters if it has approximately the same posterior probability as the runner-up value. The classification performance of a naive Bayesian classifier will thus be quite robust if the majority of presented instances result in a single rather likely class value. Yet, the output class value established from a naive Bayesian network in which one or more features given a particular class have rather small prior probabilities, will also be quite robust as long as the posterior probabilities computed for this class are rather extreme for all possible instances.

5 Scenario Sensitivity

For classification problems, it is generally assumed that evidence is available for every single feature variable. In the previous section, in fact, we also adopted this assumption. In practical applications, however, this assumption may not always be realistic. In the medical domain, for example, a patient is to be classified into one of a number of diseases without being subjected to every possible diagnostic test. The question then arises how much impact additional evidence could have on the probability distribution over the class variable and how sensitive this impact is to inaccuracies in the network's parameters. The former issue is closely related to the notion of value of (perfect) information and can be studied as part of a sensitivity-to-evidence (SE) analysis [6]. The latter issue involves a notion of sensitivity that differs from the standard notion used in the previous sections in that it pertains not to actually available evidence but to scenarios with possibly additional evidence. We refer to this notion of sensitivity as *scenario sensitivity* and use the term *evidence sensitivity* to refer to the more standard notion. Although it is applicable to Bayesian networks in general, we restrict our discussion of the notion of scenario sensitivity here to the context of naive Bayesian networks.

Before elaborating on the effects of inaccuracies in a network's feature parameters on the impact of additional evidence, we begin by reviewing the impact of the new evidence itself on an output probability of interest. For this purpose, we consider, for a specific class value, the ratio of the two posterior probabilities given the available instance prior to and after receiving the new evidence, respectively. Let E^O and E^N be sets of feature variables with $\varnothing \subseteq E^O \subset E^N \subseteq E$ and $E^N - E^O = \{E_1, \ldots, E_l\}$, $1 \leq l \leq n$. Let e^O and e^N be consistent instances of E^O and E^N, respectively, such that e^N extends the available instance e^O with the newly obtained evidence for the variables $E_1, \ldots, E_l$. Then, for each class value c, we have that

$$\frac{\Pr(c \mid e^N)}{\Pr(c \mid e^O)} = \frac{\prod_{i=1}^{l} \Pr(e_i \mid c)}{\sum_{c_j} \prod_{i=1}^{l} \Pr(e_i \mid c_j) \cdot \Pr(c_j \mid e^O)}$$

where e_i is the value of E_i in E^N. Note that the above property allows us to compute the new posterior probability distribution over the class variable from the previous one without performing any additional network propagations.

Example 5. We consider again the naive Bayesian network and the associated patient information from the previous examples. Suppose that, in addition to the diagnostic tests to which the patient has already been subjected, a CT scan of the upper abdomen can be performed to establish the presence of metastases in the liver. The network includes the following parameter probabilities for the feature variable *CT-liver*:

$\theta(CT\text{-}liver \mid S)$		I	IIA	IIB	III	IVA	IVB
CT-liver	= yes	0.05	0.05	0.05	0.05	0.05	**0.69**
	no	0.95	0.95	0.95	0.95	0.95	0.31

For deciding whether or not to perform the scan, we would like to know the possible impact of the test result on the posterior probability distribution over the various stages computed for our patient, that is, we are interested in the posterior distributions given an additional positive result and given an additional negative result from the scan. From the parameter probabilities mentioned above and the original posterior probability distribution $\Pr(S \mid e)$ from Example 1, we compute the probability of a positive test result to be $\sum_S \Pr(\textit{CT-liver} = \text{yes} \mid S) \cdot \Pr(S \mid e) = 0.12$. We recall that for our patient, the original probability of stage IVB was computed to be 0.11. Given an additional positive result from the scan, the new probability of stage IVB would be

$$\Pr(\text{IVB} \mid e^N) = \frac{0.69}{0.12} \cdot 0.11 = 0.63$$

Note that a positive test result from the CT scan of the liver would, for this patient, change the most likely stage from IVA to IVB. Further note that the new posterior probability distribution over the various stages can be established without requiring any additional computations from the network. Similar observations hold for computing the impact of a negative test result on the posterior probability distribution. □

So far, we considered the impact of additional evidence on the posterior probability distribution over the class variable computed from a naive Bayesian network. We now turn to the sensitivity of this impact to inaccuracies in the network's parameters and write the above probability ratio as a function of a parameter x:

$$h(x) = \left(\frac{\Pr(c \mid e^N)}{\Pr(c \mid e^O)}\right)(x) = \frac{\Pr(c \mid e^N)(x)}{\Pr(c \mid e^O)(x)}$$

The impact of inaccuracies in parameters of already observed feature variables on the posterior distribution over the class variable can be studied with the sensitivity functions given in the previous sections. These functions, however, do not provide for establishing the effect of inaccuracies in parameters of yet unobserved features. Focusing on the latter, we now observe that, if the parameter x pertains to a variable from the set $E^N - E^O$ of newly observed feature variables, then the denominator in the above formula is a constant with respect to x. The function $h(x)$ then just *scales* the sensitivity function $f_{\Pr(c|e^N)}(x)$ describing the output probability given *all* available evidence. Given the posterior probability distribution $\Pr(C \mid e^O)$ over the class variable prior to obtaining the new information, we can therefore immediately determine the sensitivity of the impact of the additional evidence to parameter variation from the sensitivity function $f_{\Pr(c|e^N)}(x)$. Note that for a naive Bayesian network the latter function is readily established for each feature parameter x once the posterior probability distribution $\Pr(C \mid e^N)$ is available.

Example 6. We consider again the previous example. Suppose that we now are interested in the effects of inaccuracies in the parameter probability $x = \theta(\textit{CT-liver} = \text{yes} \mid \text{IVB})$ of the feature variable *CT-liver* on the ratio of the posterior probabilities of stage IVB. We recall that the new posterior probability of this stage given the additional evidence of a positive

test result would be 0.63; the probability of IVB given just the available evidence was 0.11. We now establish the sensitivity function $f_{\Pr(\text{IVB}|e^N)}(x) = x/(x + 0.41)$ and find that

$$h_{\text{IVB}}(x) = \frac{1}{0.11} \cdot \frac{x}{x + 0.41}$$

From Example 3 we had that the probability of the class value IVB increased from 0.11 to 0.63 upon a positive liver scan, thereby becoming 5.7 times as likely. We can now in addition conclude that if the parameter x is varied, the class value IVB can become at most 6.4 times as likely as without the additional evidence. □

6 Concluding observations

Numerous experiments have shown that classifiers built on naive Bayesian networks perform quite well, even if their parameter probabilities are known to include considerable inaccuracies. In this paper, we used sensitivity-analysis techniques to study the effects of these parameter inaccuracies on the posterior probability distributions computed from a naive Bayesian network. We showed that the independence properties of such a network serve to highly constrain the functional form of the associated sensitivity functions. These functions, in fact, are determined solely by the original value of the parameter under study and the original posterior probability distribution over the class variable, and can thus be efficiently computed, requiring a single network propagation only. The properties that we derived from the sensitivity functions further provided some fundamental corroboration for the empirically observed robustness of naive Bayesian classifiers in practice.

In our future research, we would like to further underpin the observed robustness of naive Bayesian classifiers by studying properties of sensitivity. In this paper, for example, we have studied the effects of varying a single parameter probability only on an output probability of interest. Especially when discussing the possible effects of inaccuracies in a network's feature parameters, it seems only natural to consider the effects of simultaneous variation of two or more parameters. In general, the n-way sensitivity functions that describe such effects are fractions of two multi-linear functions in the parameters varied. Establishing and analysing these functions quickly becomes infeasible. In naive Bayesian networks, however, these functions may again turn out to have rather constrained forms. For example, only feature parameters pertaining to the same value of the class variable can interact in their effect on the computed posterior probabilities. Without such interaction effects, we expect that our observations concerning the robustness of a classifier's performance to parameter variation can be generalised. Studying the interaction effects in detail, moreover, may result in additional insights in the apparent lack of sensitivity to parameter inaccuracies.

In this paper, we further introduced the novel notion of scenario sensitivity, which describes the effects of parameter inaccuracies in view of scenarios of additional evidence. We showed that for naive Bayesian networks such scenario sensitivities can be readily expressed in terms of the more standard sensitivity functions. More specifically, for parameters of newly observed feature variables, the scenario sensitivity functions just scale with the standard ones. In the near future, we would like to study the properties of scenario sensitivity functions for all classifier parameters, and study the notion of scenario sensitivity in Bayesian networks in general.

A Appendix

A.1 The proof of Proposition 1

Proof. Let $x = \theta(c')$ be a parameter with original value x_0, pertaining to the value c' of class variable C. Let $\Pr(c \mid e)$ be an output probability of interest with original value p_0, and let p_0' be the original value of $\Pr(c' \mid e)$.

We begin by writing the marginal probability $\Pr(c, e)$ as a function of the parameter $x = \theta(c')$ of the network's class variable. Using the definition of conditional probability, we have that

$$\Pr(c, e) \;=\; \Pr(e \mid c) \cdot \theta(c)$$

If c is the class value whose parameter probability is being varied, that is, if $c = c'$, then $\Pr(c, e)$ relates directly to the parameter $x = \theta(c')$ in the sense that $\Pr(c, e)(x) \;=\; \Pr(e \mid c) \cdot x$. If $c \neq c'$, on the other hand, we have that the conditional probability in the expression above co-varies with the parameter x. We then find that the marginal probability $\Pr(c, e)$ relates to x as

$$\Pr(c, e)(x) = \Pr(e \mid c) \cdot \theta(c) \cdot \frac{1 - x}{1 - x_0}$$

We thus find that the marginal probability $\Pr(c, e)$ can be expressed in terms of the parameter $x = \theta(c')$ as

$$\Pr(c, e)(x) = \begin{cases} \Pr(e \mid c') \cdot x & \text{if } \; c = c' \\ -\dfrac{\Pr(e \mid c) \cdot \theta(c)}{1 - x_0} \cdot x + \dfrac{\Pr(e \mid c) \cdot \theta(c)}{1 - x_0} & \text{otherwise} \end{cases}$$

Similarly, using the definition of marginalisation, the probability $\Pr(e)$ can be written as

$$\Pr(e)(x) = \Pr(c', e)(x) + \sum_{c \neq c'} \Pr(c, e)(x)$$

where the marginal probabilities $\Pr(c', e)$ and $\Pr(c, e)$ with $c \neq c'$, relate to the parameter x as indicated above.

We define the following three constants:

$$\begin{aligned} a &= \Pr(e \mid c') = \frac{\Pr(c', e)}{x_0} \\ b &= \frac{\Pr(c, e)}{1 - x_0} \\ h &= \sum_{c \neq c'} \frac{\Pr(c, e)}{1 - x_0} = \frac{\Pr(e) - \Pr(c', e)}{1 - x_0} \end{aligned}$$

We now find that the probabilities $\Pr(c, e)$ and $\Pr(e)$ can be expressed in terms of the parameter $x = \theta(c')$ and the above constants as

$$\Pr(c, e)(x) = \begin{cases} a \cdot x & \text{if } \; c = c' \\ -b \cdot x + b & \text{otherwise} \end{cases}$$

and

$$\Pr(e)(x) = (a-h)\cdot x + h$$

For $c = c'$, we thus find for the sensitivity function that

$$f_{\Pr(c'|e)}(x) = \frac{\Pr(c',e)(x)}{\Pr(e)(x)} = \frac{a\cdot x}{(a-h)\cdot x + h} = \frac{(a/(a-h))\cdot x}{x + h/(a-h)} = \frac{(1-s)\cdot x}{x-s}$$

and for $c \neq c'$ that

$$f_{\Pr(c|e)}(x) = \frac{\Pr(c,e)(x)}{\Pr(e)(x)} = \frac{-b\cdot x + b}{(a-h)\cdot x + h} = \frac{t\cdot(x-1)}{x-s}$$

where the constant s defines the vertical asymptote that is shared by the sensitivity functions $f_{\Pr(c'|e)}(x)$ and $f_{\Pr(c|e)}(x)$, $c \neq c'$. The constant t defines the horizontal asymptote for the function $f_{\Pr(c|e)}(x)$ with $c \neq c'$; note that the horizontal asymptote of the function $f_{\Pr(c'|e)}(x)$ is defined by $t = 1 - s$.

The value s for the sensitivity functions $f_{\Pr(c'|e)}(x)$ and $f_{\Pr(c|e)}(x)$, $c \neq c'$, equals

$$\begin{aligned}
s &= -\frac{h}{a-h} = -\frac{(\Pr(e) - \Pr(c',e))/(1-x_0)}{\Pr(c',e)/x_0 - (\Pr(e) - \Pr(c',e))/(1-x_0)} = \\
&= \frac{\Pr(c',e) - \Pr(e)}{((1-x_0)/x_0 + 1)\cdot\Pr(c',e) - \Pr(e)} = \frac{\Pr(c',e) - \Pr(e)}{(1/x_0)\cdot(\Pr(c',e) - x_0\cdot\Pr(e))} = \\
&= \frac{x_0\cdot\Pr(e)\cdot(\Pr(c' \mid e) - 1)}{\Pr(e)\cdot(\Pr(c' \mid e) - x_0)} = \frac{x_0\cdot(p_0' - 1)}{p_0' - x_0}
\end{aligned}$$

The value t for the sensitivity function $f_{\Pr(c|e)}(x)$ with $c \neq c'$ equals

$$\begin{aligned}
t &= \frac{-b}{a-h} = -\frac{\Pr(c,e)}{(1-x_0)\cdot\Big(\Pr(c',e)/x_0 - (\Pr(e) - \Pr(c',e))/(1-x_0)\Big)} = \\
&= -\frac{\Pr(c \mid e)\cdot\Pr(e)\cdot x_0}{(1-x_0)\cdot\Pr(c' \mid e)\cdot\Pr(e) - \Pr(e)\cdot x_0 + \Pr(c' \mid e)\cdot\Pr(e)\cdot x_0} = \\
&= \frac{p_0\cdot x_0}{-(1-x_0)\cdot p_0' + x_0 - p_0'\cdot x_0} = \frac{p_0\cdot x_0}{x_0 - p_0'} = \\
&= \frac{p_0\cdot x_0}{x_0\cdot(1-p_0')}\cdot\frac{x_0\cdot(1-p_0')}{x_0 - p_0'} = \frac{p_0}{1-p_0'}\cdot s
\end{aligned}$$

The properties stated in the proposition now follow. □

A.2 The proof of Proposition 2

Proof. Let $x = \theta(c')$ be a parameter with original value x_0, pertaining to the value c' of class variable C. Let p_0' be the original value of $\Pr(c' \mid e)$, let $p_0^{\mathsf{T}} = \text{argmax}_{c \neq c'}\{\Pr(c \mid e)\}$, and let c^{T} be a value of C for which $\Pr(c^{\mathsf{T}} \mid e) = p_0^{\mathsf{T}}$.

We begin by observing that all sensitivity functions $f_{\Pr(c|e)}(x)$ share the same vertical asymptote, located at $x = s$, regardless of the class value c.

The first property stated in the proposition now follows from the observation that, regardless of the location of the vertical asymptote, all functions $f_{\Pr(c|e)}(x)$ for the class values $c \neq c'$ are monotonically decreasing and intersect only at $f(1) = 0$. As a consequence, either c' or c^{T} is the most likely value of the class variable C, regardless of the value of the parameter x.

For the admissible deviation for the class parameter $x = \theta(c')$, we observe that regardless of the location of the vertical asymptote, the function $f_{\Pr(c'|e)}(x)$ includes the two points $f_{c'}(0) = 0$ and $f_{c'}(1) = 1$, while the function $f_{\Pr(c^{\mathsf{T}}|e)}(x)$ includes $f_{c^{\mathsf{T}}}(1) = 0$. The two functions therefore intersect at some value $x_m \in [0, 1\rangle$ for the parameter x within the unit window. The value of x_m now follows from $f_{\Pr(c'|e)}(x) = f_{\Pr(c^{\mathsf{T}}|e)}(x)$:

$$\frac{(1-s)\cdot x}{x-s} = \frac{p_0^{\mathsf{T}}}{1-p_0'} \cdot \frac{s\cdot(x-1)}{x-s} \iff (1-s)\cdot x = \frac{p_0^{\mathsf{T}}}{1-p_0'} \cdot s \cdot (x-1)$$

$$\iff x = \frac{-s \cdot p_0^{\mathsf{T}}}{(1-p_0')\cdot(1-s) - s\cdot p_0^{\mathsf{T}}}$$

Substituting the value $(1-p_0')\cdot x_0/(x_0-p_0')$ for s in the above formula and multiplying both the numerator and the denominator by $(x_0-p_0')/(1-p_0')$ results in:

$$x_m = \frac{p_0^{\mathsf{T}} \cdot x_0}{(1-x_0)\cdot p_0' + p_0^{\mathsf{T}} \cdot x_0}$$

Note that our assumption of hyperbolic sensitivity functions implies that $x_0 \neq p_0'$. The admissible deviations stated in the proposition now follow immediately from the functional forms. □

A.3 The proof of Proposition 3

Proof. Let E_v be a feature variable with value e_v in instance e. Let $x = \theta(e_v' \mid c')$ be a parameter with original value x_0, pertaining to the value e_v' of E_v and the class value c'. Let $\Pr(c \mid e)$ be an output probability of interest with the original value p_0, and let p_0' be the original value of $\Pr(c' \mid e)$. We prove the proposition for $e_v = e_v'$; the proof for $e_v \neq e_v'$ is analogous.

We begin by writing the marginal probability $\Pr(c, e)$ in terms of the network's parameters:

$$\Pr(c, e) = \prod_{E_i \in E \setminus \{E_v\}} \theta(e_i \mid c) \cdot \theta(c) \cdot \theta(e_v' \mid c)$$

where e_i is the value of the feature variable E_i in the input instance e. Building upon this expression, the probability $\Pr(c, e)$ relates to the parameter $x = \theta(e'_v \mid c')$ as

$$\Pr(c,e)(x) \;=\; \begin{cases} \displaystyle\prod_{E_i \in E \setminus \{E_v\}} \theta(e_i \mid c') \cdot \theta(c') \cdot x & \text{if } c = c' \\ \displaystyle\prod_{E_i \in E} \theta(e_i \mid c) \cdot \theta(c) & \text{otherwise} \end{cases}$$

Similarly, using the definition of marginalisation, the probability $\Pr(e)$ can be written as

$$\Pr(e)(x) = \Pr(c', e)(x) + \sum_{c \neq c'} \Pr(c, e)$$

where the marginal probability $\Pr(c', e)$ relates to the parameter x as indicated above and the sum $\sum_{c \neq c'} \Pr(c, e)$ is constant with respect to x. Note that the parameters $\theta(e_v \mid c')$ for $e_v \neq e'_v$, which co-vary with x, are no part of the above expression since such values e_v do not occur in the input instance e, as a result of our assumption $e_v = e'_v$.

We define the following constants:

$$a \;=\; \prod_{E_i \in E \setminus \{E_v\}} \theta(e_i \mid c') \cdot \theta(c') = \frac{\Pr(c', e)}{x_0}$$

$$b \;=\; \prod_{E_i \in E} \theta(e_i \mid c) \cdot \theta(c)$$

$$h \;=\; \sum_{c \neq c'} \Pr(c, e) = \Pr(e) - \Pr(c', e)$$

For the class value $c = c'$ we now find for the sensitivity function $f_{\Pr(c'|e)}(x)$ that

$$f_{\Pr(c'|e)}(x) = \frac{\Pr(c', e)(x)}{\Pr(e)(x)} = \frac{a \cdot x}{a \cdot x + h} = \frac{x}{x - s}$$

and for the class value $c \neq c'$ we find that

$$f_{\Pr(c|e)}(x) = \frac{f_{\Pr(c,e)}(x)}{f_{\Pr(e)}(x)} = \frac{b}{a \cdot x + h} = \frac{r}{x - s}$$

where the constant $s = -h/a$ defines the vertical asymptote that is shared by the functions $f_{\Pr(c'|e)}(x)$ and $f_{\Pr(c|e)}(x)$, $c \neq c'$; its value equals

$$s \;=\; -x_0 \cdot \frac{\Pr(e) - \Pr(c', e)}{\Pr(c', e)} \;=\; x_0 \cdot \left(1 - \frac{\Pr(e)}{\Pr(c', e)}\right) \;=\; x_0 - \frac{x_0}{p'_0}$$

The constant $r = b/a$ of the function $f_{\Pr(c|e)}(x)$ with $c \neq c'$ now directly follows from $f_c(x_0) = p_0 = r/(x_0 - s)$. Finally, the function $f_{\Pr(c'|e)}(x)$ has a horizontal asymptote defined by $t = a/a = 1$. The function $f_{\Pr(c|e)}(x)$ with $c \neq c'$ has $t = 0/a = 0$. The proposition summarises these properties. □

A.4 The proof of Proposition 4

Proof. Let E_v be a feature variable with value e_v in instance e. Let $x = \theta(e'_v \mid c')$ be a parameter with original value x_0, pertaining to the value e'_v of E_v and the class value c'. Let p'_0 be the original value of $\Pr(c' \mid e)$; in addition, let $p^{\top}_0 = \text{argmax}_{c \neq c'}\{\Pr(c \mid e)\}$ and let $c^{\top}$ be a value of C for which $\Pr(c^{\top} \mid e) = p^{\top}_0$.

The first property stated in the proposition follows from the observation that all functions $f_{\Pr(c|e)}$ with $c \neq c'$ have the same horizontal and vertical asymptotes and therefore do not intersect. As a consequence, either c' or $c^{\top}$ is the most likely value of C, regardless of the value of x.

For the admissible deviation for the parameter $\theta(e'_v \mid c')$, we consider the value x_m at which the two functions $f_{\Pr(c'|e)}(x)$ and $f_{\Pr(c^{\top}|e)}(x)$ intersect. We establish this value for the situation where $e_v = e'_v$; for $e_v \neq e'_v$ the proof is analogous. For $e_v = e'_v$ we find that

$$f_{\Pr(c'|e)}(x) = f_{\Pr(c^{\top}|e)}(x) \iff \frac{x}{x-s} = p^{\top}_0 \cdot \frac{x_0 - s}{x-s} \iff x = p^{\top}_0 \cdot (x_0 - s)$$

Substituting s in the above formula with its value $x_0 - x_0/p'_0$ results in $x_m = p^{\top}_0 \cdot x_0/p'_0 \in [0, \infty\rangle$. Note that the intersection of the two functions lies within the unit-window if $x_m < 1$, that is, if $x_0 < p'_0/p^{\top}_0$. We further note that the sensitivity function $f_{\Pr(c'|e)}$ is a IVth-quadrant hyperbola branch, and therefore $p^{\top}_0 < p'_0$ if and only if $x_0 > x_m$. The admissible deviations given $e_v = e'_v$ now follow immediately from the functional forms. □

References

[1] E. Castillo, J.M. Gutiérrez, A.S. Hadi: Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics* 27: 412 – 423, 1997.

[2] H. Chan, A. Darwiche: A distance measure for bounding probabilistic belief change. In: R. Dechter, M. Kearns, R.S. Sutton (eds.), *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, AAAI Press 2002, pp. 539 – 545.

[3] H. Chan, A. Darwiche: Reasoning about Bayesian network classifiers. In: C. Meek, U. Kjærulff (eds), *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann 2003, pp. 107 – 115.

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

[5] P. Domingos, M. Pazzani: On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29: 103 – 130, 1997.

[6] F.V. Jensen: *Bayesian Networks and Decision Graphs*. Springer-Verlag 2001.

[7] U. Kjærulff, L.C. van der Gaag: Making sensitivity analysis computationally efficient. In: C. Boutilier, M. Goldszmidt (eds), *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann 2000, pp. 317 – 325.

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

[9] P. Liu, L. Lei, N. Wu: A quantitative study of the effect of missing data in classifiers. In: N. Gu, D. Wei, Z. Xie, H. Wang, S.X. Wang, B. Shi (eds), *IEEE Proceedings of the 5th International Conference on Computer and Information Technology*, pp. 28 – 33, 2005.

[10] J. Pearl: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann 1988.

[11] S. Renooij, L.C. van der Gaag: Evidence-invariant sensitivity bounds. In: M. Chickering, J. Halpern (eds), *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, AUAI Press 2004, pp. 479 – 486.

[12] I. Rish: An empirical study of the naive Bayes classifier. In: *IJCAI Workshop on Empirical Methods in Artificial Intelligence*, pp. 41 – 46, 2001.

[13] L.C. van der Gaag, S. Renooij: Analysing sensitivity data from probabilistic networks. In: J.S. Breese, D. Koller (eds), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann 2001, pp. 530 – 537.

[14] L.C. van der Gaag, S. Renooij: On the sensitivity of probabilistic networks to reliability characteristics. *Modern Information Processing: From Theory to Applications*, Elsevier 2006, pp. 385 – 405.