On Decision Boundaries of Naïve Bayes in Continuous Domains

Tapio Elomaa and Juho Rousu

Department of Computer Science, University of Helsinki, Finland
{elomaa,rousu}@cs.helsinki.fi

Abstract. Naïve Bayesian classifiers assume the conditional independence of attribute values given the class. Despite this in practice often violated assumption, these simple classifiers have been found efficient, effective, and robust to noise.

Discretization of continuous attributes in naïve Bayesian classifiers has achieved a lot of attention recently. Continuous attributes need not necessarily be discretized, but it unifies their handling with nominal attributes and can lead to improved classifier performance.

We show that optimal partitioning results from decision tree learning carry over to Naïve Bayes as well. In particular, it sets decision boundaries on borders of segments with equal class frequency distribution. An optimal univariate discretization with respect to the Naïve Bayes rule can be found in linear time but, unfortunately, optimal multivariate optimization is intractable.

1 Introduction

The naïve Bayesian classifier, or Naïve Bayes, is surprisingly effective in classification tasks. Therefore, even if it does not belong to state-of-the-art methods, it plays an important role — alongside decision tree learning — as standard baseline methods of inductive algorithms. Naïve Bayesian classifiers have been studied extensively over the years [18,19,7].

In Naïve Bayes numerical attributes can be handled without explicit discretization of the value range [7,15] unlike in, e.g., decision tree induction. An often made assumption is that within each class the data is generated by a single Gaussian distribution. To model actual distributions more faithfully one can abandon the normality assumption and, rather, use nonparametric density estimation [7,15].

Treating numerical attributes by density estimation, Gaussian or other, indicates that numerical and discrete attributes are handled differently. Furthermore, discretization has been observed to increase the prediction accuracy and make the method more efficient [2,6]. There are discretization methods that are specific to Naïve Bayes [2,6,25,4,26] as well as general approaches that are often used with naïve Bayesian classifiers [13]. A particularly interesting fact is that Naïve Bayes permits overlapping discretization [16,27] unlike many other classification learning algorithms.

N. Lavrač et al. (Eds.): PKDD 2003, LNAI 2838, pp. 144–155, 2003.

In decision tree setting the line of research stemming from Fayyad and Irani's [12] seminal work on optimal discretizations for evaluation functions of ID3 has led to more efficient preprocessing approaches and a better understanding of the necessary and sufficient preprocessing needed to guarantee finding optimal partitions with respect to common evaluation functions [8,9,10,11]. In this paper we show that this type of analysis carries over to naïve Bayesian classifiers despite the difference of univariate inspection in decision trees and multivariate one in naïve Bayesian classifiers. The *decision boundaries* separating decision regions — class prediction changes — of naïve Bayesian classifiers fall exactly on the so-called segment borders. No other cut point candidates need to be considered in order to find the error-minimizing discretization.

We show that with respect to one numerical attribute, a partition that optimizes the naïve Bayes rule can be found in linear time using the same algorithm as in connection with decision trees. However, simultaneously satisfying the optimality with respect to more than one attribute, unfortunately, has recently turned out to be NP-complete [23]. This does not leave us with possibilities to solve the problem efficiently.

In Sect. 2 we first recapitulate the basics on naïve Bayesian classification. In Sect. 3 the optima-preserving preprocessing of numerical value ranges is reviewed. In Sect. 4 we prove that the same line of analysis applies to naïve Bayesian classifiers as well. We also briefly consider multivariate discretization in this section. Finally, Sect. 5 concludes this article by summarizing the work and discussing further research possibilities.

2 Naïve Bayes

Naïve Bayes gives an instance $\boldsymbol{x} = \langle a_1, \ldots, a_n \rangle$ the label

$$\arg\max_{c \in C} \mathbf{Pr}\left(c \mid \boldsymbol{x}\right), \tag{1}$$

where C is the set of classes. In other words, the classifier assigns for the given instance the class that is most probable. The computation of the conditional probability $\mathbf{Pr}\left(c \mid \boldsymbol{x}\right)$ is based on the Bayes rule

$$\mathbf{Pr}\left(c \mid \boldsymbol{x}\right) = \frac{\mathbf{Pr}\left(c\right)\,\mathbf{Pr}\left(\boldsymbol{x} \mid c\right)}{\mathbf{Pr}\left(\boldsymbol{x}\right)}$$

and the (naïve) assumption that the attributes $A_1, \ldots, A_n$ are independent of each other given the class, which indicates that

$$\mathbf{Pr}\left(\boldsymbol{x} \mid c\right) = \prod_{i=1}^{n} \mathbf{Pr}\left(A_i = a_i \mid c\right).$$

The denominator $\mathbf{Pr}\left(\boldsymbol{x}\right)$ of the Bayes rule is the same for all classes in C. Therefore, it is convenient to consider the quantity

$$\arg\max_{c \in C} \mathbf{Pr}\left(c \cap \boldsymbol{x}\right) = \arg\max_{c \in C} \mathbf{Pr}\left(c \mid \boldsymbol{x}\right) \mathbf{Pr}\left(\boldsymbol{x}\right)$$

instead of (1). The two formulas, of course, always predict the same label.

Probability estimation is based on a training set of classified examples $E = \{\, \langle \boldsymbol{x}_i, y_i \rangle \,\}_{i=1}^{m}$, where $y_i \in C$ for all i. Let m_c denote the number of instances from class c in E. Then, the data prior for class c is $\widehat{P}(c) = m_c/m$. Discrete attributes are easy to handle: we just estimate $\mathbf{Pr}(\boldsymbol{x} \mid c)$ based on the training set by estimating the conditional marginals $\widehat{P}(A_i = a_i \mid c)$ by counting the fraction of occurrences of each value $A_i = a_i$ in m_c.

Unless discretized, continuous values are harder to take care of and require a different strategy. It is common to assume that within each class c the values of numeric attributes are normally distributed. Then by estimating from the training set the mean μ_c and standard deviation σ_c of the continuous attribute given c, one can compute the probability of the observed value. After obtaining μ_c and σ_c for an attribute A_i the estimation boils down to calculating the probability density function for a Gaussian distribution:

$$\mathbf{Pr}(A_i = a_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{(a_i - \mu_c)^2}{2\sigma_c^2} \right).$$

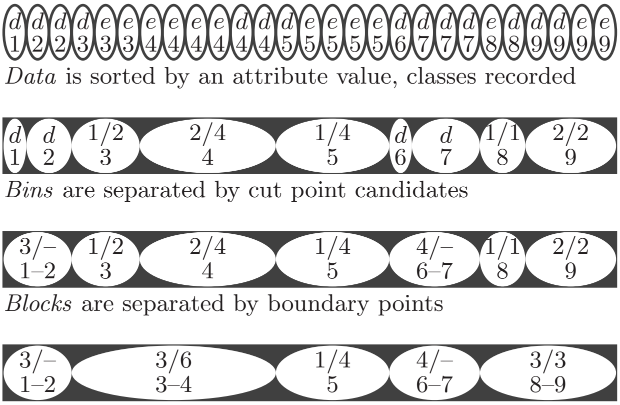
Using Dirichlet prior, or more generally Bayesian estimation methods [17,4], and kernel density estimation [15] are some alternatives to the straightforward normality assumption for estimating a model for the distribution of the continuous attribute. In this paper, however, we are only concerned with probability estimates computed as data priors $\widehat{P}(\cdot)$.

Despite the unrealistic attribute independence assumption underlying Naïve Bayes it is a very successful classifier in practical situations. Some explanations have been offered by Domingos and Pazzani [5], who showed that Naïve Bayes may be globally optimal even though the attribute independence assumption is violated. It was shown that, under 0–1 loss, Naïve Bayes is globally optimal for the concept classes conjunctions and disjunctions of literals. Gama [14] discusses Naïve Bayes and quadratic loss.

It is well-known that the naïve Bayesian classifier is equivalent to a linear machine and, hence, for nominal attributes its decision boundary is a hyperplane [7,21,5]. Thus, Naïve Bayes can only be globally optimal for linearly separable concept classes. Ling and Zhang [20] consider the representational power of Naïve Bayes and more general Bayesian networks further. They characterize the representational power through the maximum XOR contained in a function.

3 Discretizing Continuous Attributes

The dominating discretization techniques for continuous attributes in Naïve Bayes are unsupervised equal-width binning [24] and the greedy top-down approach of Fayyad and Irani [13]. These straightforward heuristic approaches have also been offered some analytical backing [4]. However, in other classifier learners — decision trees in particular — analysis of discretization has been taken much further. In the following we recapitulate briefly the line of analysis



Data is sorted by an attribute value, classes recorded

Bins are separated by cut point candidates

Blocks are separated by boundary points

Segments have different relative class distributions

Fig. 1. The original set of 27 examples (top) can only be partitioned at bin borders (second from top) where the value of the attribute changes. Class uniform bins can be combined into blocks (second from bottom). Block borders are the boundary points of the numerical range. Furthermore, partitions can only happen between blocks with different relative class distribution. Thus, we may arrange the data into segments (below). Segment borders are a subset of boundary points

initiated by Fayyad and Irani [12]. The goal is to reduce the number of examined cut points without losing the possibility to recover optimal partitions.

In decision tree learning the processing of a numerical value range usually starts with sorting of the data points [1,22]. If one could make its own partition interval out of each data point in the sorted sequence, this discretization would have zero training error. However, only those points that differ in their value can be separated from each other. Therefore, we can preprocess the data into *bins*, one bin for each existing data point value. Within each bin we record the class distribution of the instances that belong to it (see Fig. 1). The class distribution information suffices to evaluate the goodness of the partition; the actual data set does not need to be maintained.

The sequence of bins attains the minimal misclassification rate. However, the same rate can usually be obtained with a smaller number of intervals. The analysis of the entropy function by Fayyad and Irani [12] has shown that cut points embedded into class-uniform intervals need not be taken into account, only the end points of such intervals — the *boundary points* — need to be considered to find the optimal discretization. Elomaa and Rousu [8] showed that the same is true for several commonly-used evaluation functions.

Subsequently, a more general property was also proved for some evaluation functions [9]: *segment borders* — points that lie in between two adjacent bins with different relative class distributions — are the only points that need to be taken into account. It is easy to see that segment borders are a subset of boundary points.

For strictly convex evaluation functions it was shown later that examining segment borders is necessary as well as sufficient in order to be able to discover the optimal partition. For Training Set Error, which is not strictly convex, it suffices to only examine a subset of the segment borders. These points are called *alternations* and they are placed on segment borders where the frequency ordering of the classes changes [10,11].

These analyses can be used in preprocessing in a straightforward manner: we merge together, in linear time, adjacent class uniform bins with the same class label to obtain example *blocks* (see Fig. 1). The boundary points of the value range are the borders of its blocks. Example *segments* are easily obtained from bins by comparing the relative class distributions of adjacent bins (see Fig. 1). This can be accomplished on the same left-to-right scan that is required to identify bins. Also alternations can be detected during the same scan.

4 Decision Boundaries of Naïve Bayes

We show now that segments in the domain of a continuous attribute are the locations where Naïve Bayes changes its class prediction, i.e., its decision boundaries. We start from undiscretized domains, go on to error-minimizing discretizations, and finally consider optimal partitions with respect to several attributes.

4.1 Decision Boundaries for Undiscretized Attributes

We start by examining the decision boundaries that Naïve Bayes sets when the continuous attribute is not discretized, but each numerical value is treated separately. When we consider the decision boundaries from the point of view of one attribute, we assume an arbitrary fixed value setting for the other attributes. In the following, notation $\widehat{P}(\cdot)$ is used to denote the probability estimates computed by Naïve Bayes, to distinguish them from true probabilities.

Theorem 1. *The decision boundaries of the naïve Bayesian classifier are situated on segment borders.*

Proof. Let A_i be a numerical attribute and let V' and V'' be two adjacent intervals in its range separated by cut point v_i. For any other attribute A_j, $i \neq j$, let V_j be an arbitrary subset of the values of A_j. Let us denote by

$$T = V_1 \times \cdots \times V_{i-1} \times V_{i+1} \times \cdots \times V_n$$

the Cartesian product of these subsets. We assume that the prediction of naïve Bayesian classifier within $V' \times T$ is $c' \in C$, and the prediction within $V'' \times T$ is $c'' \in C$. In other words, looking at the situation only from the point of view of A_i and taking all other attributes to have an arbitrary (but fixed) value combination, the decision boundary is set between intervals V' and V''. Then

$$\begin{aligned}\widehat{P}(c' \mid V' \times T) &= \widehat{P}(c')\,\widehat{P}(V' \mid c') \prod_{j \neq i} \widehat{P}(V_j \mid c') \\ &> \widehat{P}(c'')\,\widehat{P}(V' \mid c'') \prod_{j \neq i} \widehat{P}(V_j \mid c'') = \widehat{P}(c'' \mid V' \times T).\end{aligned}$$

By reorganizing the middle inequality we get

$$\frac{\widehat{P}\,(V' \mid c')}{\widehat{P}\,(V' \mid c'')} > \frac{\widehat{P}\,(c'')\prod_{j\neq i}\widehat{P}\,(V_j \mid c'')}{\widehat{P}\,(c')\prod_{j\neq i}\widehat{P}\,(V_j \mid c')}. \tag{2}$$

On the other hand, within $V'' \times T$ we obtain, by similar manipulation,

$$\frac{\widehat{P}\,(V'' \mid c')}{\widehat{P}\,(V'' \mid c'')} < \frac{\widehat{P}\,(c'')\prod_{j\neq i}\widehat{P}\,(V_j \mid c'')}{\widehat{P}\,(c')\prod_{j\neq i}\widehat{P}\,(V_j \mid c')}. \tag{3}$$

Put together, (2) and (3) imply

$$\frac{\widehat{P}\,(V'' \mid c')}{\widehat{P}\,(V'' \mid c'')} < \frac{\widehat{P}\,(c'')\prod_{j\neq i}\widehat{P}\,(V_j \mid c'')}{\widehat{P}\,(c')\prod_{j\neq i}\widehat{P}\,(V_j \mid c')} < \frac{\widehat{P}\,(V' \mid c')}{\widehat{P}\,(V' \mid c'')}.$$

By using the Bayes rule to the conditional probabilities and canceling out equal factors we get

$$\frac{\widehat{P}\,(c' \mid V'')}{\widehat{P}\,(c'' \mid V'')} < \frac{\widehat{P}\,(c' \mid V')}{\widehat{P}\,(c'' \mid V')}.$$

Hence, the relative class distributions must be strictly different within the intervals V' and V'' making v_i thus a segment border.

The above result does not, of course, mean that all segment borders would be places for class prediction change. However, the class prediction changes of an undiscretized domain are confined to segment borders. Consequently, no loss is incurred in grouping the examples in segments of equal class distribution. On the contrary, we expect to benefit from the more accurate probability estimation.

4.2 Decision Boundaries in *k*-Interval Discretization

Let us now turn to the case where the continuous range has been discretized into k intervals. We will prove that in this case too segment borders are the only potential points for the decision boundaries.

The following proof has the same setting as the proofs in connection with decision trees [9]. The sample contains three subsets, P, Q, and R, with class frequency distributions

$$p = \sum_{j=1}^{m} p_j,\ q = \sum_{j=1}^{m} q_j,\ \text{and}\ r = \sum_{j=1}^{m} r_j,$$

where p is the number of examples in P and p_j is the number of instances of class j in P. Furthermore, m is the number of classes. The notation is similar also for Q and R.

We consider the k-ary partition $\{\, S_1, \ldots, S_k \,\}$ of the sample, where subsets S_h and S_{h+1}, $1 \le h \le k-1$, consist of the set $P \cup Q \cup R$, so that the split point

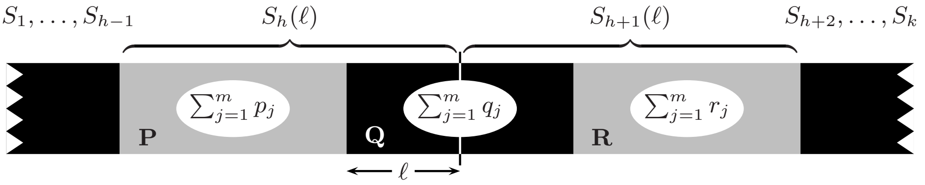


Fig. 2. The following proofs consider partitioning of the example set $P \cup Q \cup R$ into two subsets S_h and S_{h+1} within Q. No matter where, within Q, the cut point is placed, equal class distributions result

is inside Q, on the border of P and Q, or that of Q and R (see Fig. 2). Let ℓ be a real value, $0 \leq \ell \leq q$ [1]. Let $S_h(\ell)$ denote partition interval S_h that contains P and the ℓ first examples from Q. In the same situation $S_{h+1}(\ell)$ denotes the interval S_{h+1}. We assume that splitting the set Q so that ℓ examples belong to $S_h(\ell)$ and $q - \ell$ to $S_{h+1}(\ell)$ results in identical class frequency distributions for both subsets of Q regardless of the value of ℓ.

Let T again be the Cartesian product of the (arbitrary) subsets in dimensions other than the one under consideration. In this setting we can prove the following result which will be put to use later.

Lemma 1. *The sum* $\max_{c \in C} \widehat{P}(c \cap S_h(\ell) \times T) + \max_{c \in C} \widehat{P}(c \cap S_{h+1}(\ell) \times T)$ *is convex over* $\ell \in [0, q]$.

Proof. Let $\ell_1, \ldots, \ell_{r-1}$ be the class prediction change points within $[0, q]$. Without loss of generality, let us denote by c_i, $1 \leq i \leq r-1$, the class predicted within $S_h(\ell)$, $\ell \in]\ell_i, \ell_{i+1}]$. The probability of instances of class c within $S_h(\ell) \times T$ can be expressed as

$$\widehat{P}(c \cap S_h(\ell) \times T) = \frac{p_c \cdot \widehat{P}(T \mid c)}{n} + \ell \cdot \frac{q_c/q \cdot \widehat{P}(T \mid c)}{n},$$

which describes a line with offset $(p_c/n)\widehat{P}(T \mid c)$ and slope $((q_c/q)/n)\widehat{P}(T \mid c)$ (see Fig. 3). Now, it must be that the offsets satisfy

$$\frac{p_{c_1} \cdot \widehat{P}(T \mid c_1)}{n} \geq \cdots \geq \frac{p_{c_{r-1}} \cdot \widehat{P}(T \mid c_{r-1})}{n}$$

and the slopes of the lines satisfy

$$\frac{q_{c_1}/q \cdot \widehat{P}(T \mid c_1)}{n} \leq \cdots \leq \frac{q_{c_{r-1}}/q \cdot \widehat{P}(T \mid c_{r-1})}{n}.$$

Interpreting the situation geometrically, we see that $\max_c \widehat{P}(c \cap S_h(\ell) \times T)$ forms a convex curve (Fig. 3). By symmetry, $\max_c \widehat{P}(c \cap S_{h+1}(\ell) \times T)$ also is convex, and the claim follows by the convexity of the sum of convex functions.

[1] No harm is done considering splitting Q in other points than those corresponding to integral number of examples, since we are proving absence of local extrema.

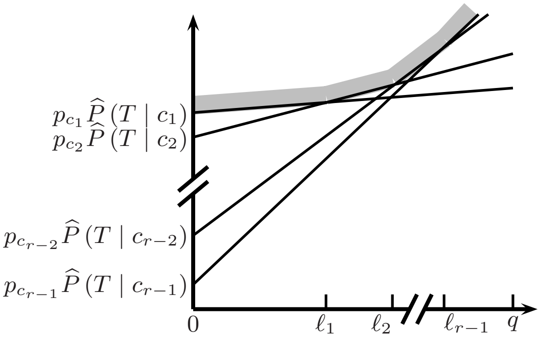


Fig. 3. The maxima of the sum of the most probable classes in $S_h(\ell)$ and $S_{h+1}(\ell)$ forms a convex curve over $[0, q]$

The following proof shows that a cut point in between two adjacent subsets S_h and S_{h+1} in one dimension is on a segment border, regardless of the context induced by the other attributes. Due to the additivity of the error, the result also holds in the multisplitting case, where a number of cut points are chosen in each dimension.

Theorem 2. *The error-minimizing cut points of Naïve Bayes are located on segment borders.*

Proof. Let $c_L(\ell) = \arg\max_{c \in C} \widehat{P}(c \cap S_h(\ell) \times T)$ be the most probable class for $S_h(\ell)$ according to Naïve Bayes criterion in this situation. In other words,

$$\widehat{P}(c_L(\ell) \cap S_h(\ell) \times T) = \max_{c \in C} \widehat{P}(c \cap S_h(\ell))\, \widehat{P}(T \mid c)\,.$$

Similarly, let $c_R(\ell)$ denote the most probable class in $S_{h+1}(\ell)$.

The minimum-error partition is the one that has the smallest combined error in the subsets $S_h(\ell)$ and $S_{h+1}(\ell)$. Thus, we want to optimize

$$\min_{\ell \in [0,q]} (\widehat{P}(S_h(\ell) \times T) - \widehat{P}(c_L(\ell) \cap S_h(\ell) \times T) + \widehat{P}(S_{h+1}(\ell) \times T) - \widehat{P}(c_R(\ell) \cap S_{h+1}(\ell) \times T)),$$

which is equal to

$$\min_{\ell \in [0,q]} (\widehat{P}(S \times T) - (\widehat{P}(c_L(\ell) \cap S_h(\ell) \times T) + \widehat{P}(c_R(\ell) \cap S_{h+1}(\ell) \times T))),$$

where $S = P \cup Q \cup R$. By Lemma 1 this is a concave function of $\ell \in [0, q]$. Hence, it minimizes at one of the extreme values of ℓ, which are the locations of the segment borders. Thus, we have proved the claim.

In principle it might be possible to reduce the number of examined points by leaving some segment borders without attention. Can we identify such a subset efficiently?

In univariate setting the answer is affirmative: Only the set of *alternation points* need to be considered. These points are those in between adjacent bins V' and V'' with a conflict in the frequency ordering of the classes: $\widehat{P}(c \mid V') < \widehat{P}(c' \mid V')$ and $\widehat{P}(c' \mid V'') < \widehat{P}(c \mid V'')$. This is a direct consequence of the fact that training error minimizes on such points. The set of alternation points can be found in linear time, so it can speed up the discretization process [10].

In multivariate setting, the other attributes need to be taken into account. Let $V' \times T$ and $V'' \times T$ be two adjacent hyperrectangles. One can show that there is no decision boundary in between them if for all class pairs c' and c'' we have either

1. $\widehat{P}(c' \mid V' \times T) \leq \widehat{P}(c \mid V' \times T)$ and $\widehat{P}(c' \mid V'' \times T) \leq \widehat{P}(c \mid V'' \times T)$, or
2. $\widehat{P}(c' \mid V' \times T) \geq \widehat{P}(c \mid V' \times T)$ and $\widehat{P}(c' \mid V'' \times T) \geq \widehat{P}(c \mid V'' \times T)$.

The problem with using this criterion to prune the set of candidate cut points is that the definition depends on the context T, and there is an exponential number of such contexts. So, even if all segment borders are not useful, deciding which of them can be discarded seems difficult.

Thus, in practice, finding a linear-time preprocessing scheme to reduce the set of potential cut points to a proper subset of segment borders is difficult.

4.3 Decision Boundaries of Naïve Bayes in Multiple Dimensions

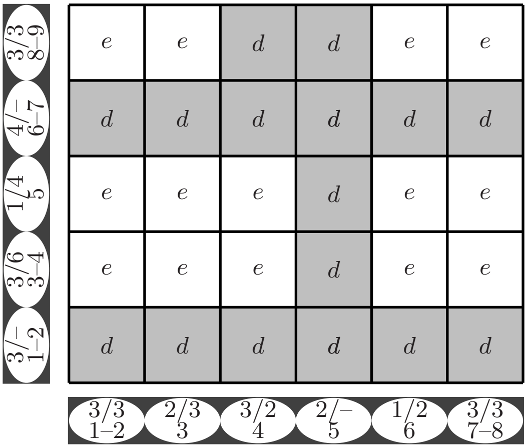
It is well known that in the discrete (two-class) case the decision boundary is a (single) hyperplane in the input space [7,21]. In case of continuous attributes the situation is much more difficult: The decision regions and their boundaries may have arbitrary shape [7]. However, from preceding results we know that decision boundaries in reality can only occur at segment borders of each continuous attribute. Therefore, we actually can consider discretized ranges instead of truly continuous attributes.

In Fig. 4 the example set of Fig. 1 has been augmented with another (arbitrary) dimension. The segments of these two dimensions divide the input space (a plane) into a 6×5 grid, where each grid cell gets assigned a class label. Class uniform rows and columns get a uniform labeling but otherwise one cannot determine the labeling of grid cells based only on one dimension. Values of both attributes are needed to determine the class label. For example, when the attribute depicted on the y-axis of Fig. 4 has a value in its last segment, depending on the value of the attribute along the x-axis, there are two segments where the most probable prediction would be d and two segment where it would be e.

In general the discretized input space is divided into hyperrectangular cells, each assigned the class label according to the relevant segment statistics.

4.4 On Finding Optimal Discretizations for Naïve Bayes

Theorem 2 tells us that the decision boundaries of Naïve Bayes are always located on segment borders, which makes it possible to preprocess the data into



3/3 8–9	*e*	*e*	*d*	*d*	*e*	*e*
4/– 6–7	*d*	*d*	*d*	*d*	*d*	*d*
1/4 5	*e*	*e*	*e*	*d*	*e*	*e*
3/6 3–4	*e*	*e*	*e*	*d*	*e*	*e*
3/– 1–2	*d*	*d*	*d*	*d*	*d*	*d*
	3/3 1–2	2/3 3	3/2 4	2/– 5	1/2 6	3/3 7–8

Fig. 4. The segments of two continuous attributes divide the input space into a rectangular grid. All grid cells are assigned the class label determined by the residual sums of the corresponding segments

segments prior to discretization. By this result, one can use the same linear-time optimization algorithm to find the univariate Naïve Bayes optimal multisplits as in the case of decision trees [9,11]. This so-called Auer-Birkendorf algorithm is based on dynamic programming. During a left-to-right scan over the segments, one can maintain the information required to decide into how many intervals should the data be split and where to locate the interval borders to obtain as good value as possible for the partition.

However, the situation in the multivariate setting is much more difficult. Even with the data preprocessed into segments we may still have a daunting amount of possible discretizations: $O(2^T)$ to be exact, where $T = \sum_{i=1}^{n} T_i$ and T_i is the number of cut points candidates along the i-th dimension. Could there, nevertheless, exist an efficient algorithm for optimal discretization? Unfortunately, we have to answer in the negative, as shown by the next theorem [23].

Theorem 3. *Finding the Naïve Bayes optimal discretization of the real plane* $\mathbb{R}^2$ *is NP-complete.*

This can be proved by a reduction from *Minimum Set Cover* using a similar construction as Chlebus and Nguyen [3] to show that already optimal consistent splitting of the real plane $\mathbb{R}^2$ is NP-complete. We construct a configuration of points in the 2D plane corresponding to the set covering instance and show two properties [23]:

1. The plane can be consistently discretized with k cut lines if and only if there is a set cover of size k for the given set cover instance.
2. The optimal Naïve Bayes discretization coincides with the consistent discretization.

Since the hypothesis class of Naïve Bayes is the set of product distributions of marginal likelihoods, the above theorem strengthens the negative result of Chlebus and Nguyen [3], which holds for general axis-parallel partitions of $\mathbb{R}^2$. Observe that the result easily generalizes to cases with more than two dimensions by embedding the 2D plane corresponding to the set covering instance into the higher dimensional space. The problem remains equally hard when there are more than two classes.

From the point of view of finding the optimal multivariate splits, exhaustive search over the segment borders of all dimensions is the remaining possibility for optimization, which becomes prohibitively time-consuming on larger datasets.

5 Conclusion

Examining segment borders is necessary and sufficient in searching for the optimal partition of a value range with respect to a strictly convex evaluation function [11]. The same set of cut point candidates is relevant for Naïve Bayes: Their decision boundaries (in disjoint partitioning) fall exactly on segment borders.

On the other hand, it seems that for an algorithm to rule out some segment borders from among the decision boundary candidates, it would have to examine too many contexts to be efficient. Therefore, preprocessing the value ranges of continuous attributes into segments appears necessary if one wants to detect all class prediction changes. Such preprocessing, naturally, is sufficient.

As future work we leave the empirical evaluation of the usefulness of segment borders and their accuracy in probability estimation as well as studying possibilities to approximate optimal multivariate discretization and the utility of segment borders therein.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Pacific Grove, CA (1984)
2. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Proc. Fifth European Working Session on Learning. Lecture Notes in Computer Science, Vol. 482. Springer-Verlag, Berlin Heidelberg New York (1991) 164–178
3. Chlebus, B.S., Nguyen, S.H.: On finding optimal discretizations for two attributes. In: Polkowski. L., Skowron, A. (eds.): Rough Sets and Current Trends in Computing, Proc. First International Conference. Lecture Notes in Artificial Intelligence, Vol. 1424, Springer-Verlag, Berlin Heidelberg New York (1998) 537–544
4. Chu, C.-N., Huang, H.-J., Wong, T.-T.: Why discretization works for naïve Bayesian classifiers. In: Langley, P. (ed.): Proc. Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA (2000) 399–406
5. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29** (1997) 103–130
6. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Proc. Twelfth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA (1995) 194–202

7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, Second Edition. John Wiley & Sons, New York (2001)
8. Elomaa, T., Rousu, J.: General and efficient multisplitting of numerical attributes. Mach. Learn. **36** (1999) 201–244
9. Elomaa, T., Rousu, J.: Generalizing boundary points. In: Proc. Seventeenth National Conf. on Artificial Intelligence. MIT Press, Cambridge, MA (2000) 570–576
10. Elomaa, T., Rousu, J.: Fast minimum error discretization. In: Sammut, C., Hoffmann, A. (eds.): Proc. Nineteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA (2002) 131–138
11. Elomaa, T., Rousu, J.: Necessary and sufficient pre-processing in numerical range discretization. Knowl. Information Systems **5** (2003) in press
12. Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. Mach. Learn. **8** (1992) 87–102
13. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco (1993) 1022–1027
14. Gama, J.: Iterative Bayes. Theor. Comput. Sci. **292** (2003) 417–430
15. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: Proc. Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA (1995) 338–345
16. Kononenko, I.: Naive Bayesian classifier and continuous attributes. Informatica **16** (1992) 1–8
17. Kontkanen, P., Myllymõki, P., Silander, T., Tirri, H.: A Bayesian approach to discretization. In: European Symposium on Intelligent Techniques. ELITE Foundation, Aachen (1997) 265–268
18. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proc. Tenth National Conference on Artificial Intelligence. MIT Press, Cambridge, MA (1992) 223–228
19. Langley, P., Sage, S.: Tractable average-case analysis of naive Bayesian classifiers. In: Bratko, I., Džeroski, S. (eds.): Proc. Sixteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA (1999) 220–228
20. Ling, C.X., Zhang, H.: The representational power of discrete Bayesian networks. J. Mach. Learn. Res. **3** (2002) 709–721
21. Peot, M.A.: Geometric implications of the naive Bayes assumption. In: Horvitz, E., Jensen, F. (eds.): Proc. Twelfth Annual Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA (1996) 414–419
22. Quinlan, J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA (1993)
23. Rousu, J.: Optimal multivariate discretization for naive Bayesian classifiers is NP-hard. Tech. Rep. C-2003-8, Dept. of Computer Science, Univ. of Helsinki (2003)
24. Wong, A.K.C., Chiu, D.K.Y.: Synthesizing statistical knowledge from incomplete mixed-mode data. IEEE Trans. Pattern Anal. Mach. Intell. **9** (1987) 796–805
25. Wu, X.: A Bayesian discretizer for real-valued attributes. Computer J. **39** (1996) 688–691
26. Yang, Y., Webb, G.I.: Proportional *k*-interval discretization for naive-Bayes classifiers. In: De Raedt, L., Flach, P. (eds.): Proc. Twelfth European Conference on Machine Learning. Lecture Notes in Artificial Intelligence, Vol. 2167, Springer-Verlag, Berlin Heidelberg New York (2001) 564–575
27. Yang, Y., Webb, G.I.: Non-disjoint discretization for naive-Bayes classifiers. In: Sammut, C., Hoffmann, A. (eds.): Proc. Nineteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA (2002) 666–673