The Learnability of Naive Bayes

Huajie Zhang*, Charles X. Ling*, and Zhiduo Zhao*

Department of Computer Science
The University of Western Ontario
London, Ontario, Canada N6A 5B7
{hzhang, ling, zdzhao}@csd.uwo.ca

Abstract. Naive Bayes is an efficient and effective learning algorithm, but previous results show that its representation ability is severely limited since it can only represent certain linearly separable functions in the binary domain. We give necessary and sufficient conditions on linearly separable functions in the binary domain to be learnable by Naive Bayes under uniform representation. We then show that the learnability (and error rates) of Naive Bayes can be affected dramatically by sampling distributions. Our results help us to gain a much deeper understanding of this seemingly simple, yet powerful learning algorithm.

1 Introduction

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. It is time efficient, for its time complexity is only linear of the training data. It is space efficient since, after discretization, it builds up a frequency table in size of the product of the number of attributes, number of class values, and the number of values per attribute. It does not need to store the training data in memory when it builds the frequency table, but just scans the data once from the disk.

Naive Bayes is also surprisingly effective in the classification task. Many empirical comparisons between Naive Bayes and modern decision tree algorithms such as C4.5 (Quinlan, 1993) showed that Naive Bayes predicts equally well as C4.5 (see, for example, Langley, Iba & Thomas, 1992; Kononenko, 1990; Pazzani, Muramatsu & Billsus, 1996). The good performance of Naive Bayes is surprising because it makes an unrealistic assumption that is almost always violated in real-world applications: given class values, all attributes are independent.

Domingos and Pazzani (1997) presented an important work explaining the good predictive performance of Naive Bayes. Basically, even though the assumption that Naive Bayes makes alters the probability distribution of the class label, the class with the maximum probability can still be the same. That is, under the MAP (Maximum A Posteriori) Principle, the 0/1 loss of Naive Bayes can be very small (Domingos & Pazzani, 1997).

However, Naive Bayes has a strong limitation in its representation ability, since it is well-known that Naive Bayes can represent only linearly separable functions in the binary domain (Duda & Hart, 1973). Further, Naive Bayes cannot represent *all* linearly separable functions. As Domingos and Pazzani (1997) showed, several specific linear functions are not learnable with Naive Bayes under uniform distribution.

We extend (Domingos & Pazzani, 1997)'s work further in this paper. We derive necessary and sufficient conditions for linearly separable functions to be learnable by Naive Bayes under uniform distribution, and we verify our results with empirical experiments. We then show, surprisingly, that the learnability and error rates of Naive Bayes on linearly separable functions can be affected dramatically by the sampling distribution: learnable functions under uniform distribution can become unlearnable under other distributions, and vice versa.

Our results help us to gain a deeper understanding of this seemingly simple yet powerful learning algorithm. Naive Bayes' learning capacities are determined not only by the target function, but also by the sampling distribution, and by how attribute values are represented.

2 Naive Bayes Basics

We start with the Boolean classification problem in the binary domain. Assume X_1, X_2,..., X_n are n binary Boolean variables taking values 1 or 0. For convenience, we use x_i to represent the value that X_i takes. Thus, an example E can be represented by a vector $(x_1, x_2, , ..., x_n)$. Let Ψ, called the example space, be a set of all possible (2^n) examples. We use C to represent the target (classification) variable, which also takes values 1 or 0 (again, we use use c to represent the value C takes).

According to the Bayesian Theorem, the probability of an example $E = (x_1, x_2, , ..., x_n)$ being class c is

$$p(c|E) = \frac{p(x_1, x_2, ..., x_n|c)p(c)}{p(x_1, x_2, ..., x_n)}$$

The prediction which Bayes makes is taken as the class value with largest probability. That is, E belongs to the class $C = 1$ iff

$$g(E) = \frac{p(C = 1|x_1, x_2, ..., x_n)}{p(C = 0|x_1, x_2, ..., x_n)} \geq 1$$

where $g(E)$ is called *the discriminant function.*

Naive Bayes assumes variable independence given class values. That is:

$$p(x_1, x_2, ..., x_n|c) = p(x_1|c)p(x_2|c)...p(x_n|c)$$

The corresponding discriminant function $g(E)$ of Naive Bayes is then:

$$g(E) = \frac{p(C = 1)p(x_1|C = 1)p(x_2|C = 1)...p(x_n|C = 1)}{p(C = 0)p(x_1|C = 0)p(x_2|C = 0)...p(x_n|C = 0)} \tag{1}$$

The value of $p(x_i|c)$ can be estimated from the training examples.

According to Duda and Hart (1973), Naive Bayes can only represent linearly separable (binary) functions. A linearly separable function is defined as $E \in C = 1$ iff $g_l(E) \geq 0$, where

$$g_l(E) = \sum_{i=1}^{n} a_i x_i - a_0 \tag{2}$$

Assume p_i and q_i represent $p(X_i = 1|C = 1)$ and $p(X_i = 1|C = 0)$ respectively. When the logarithm is applied to Equation 1, we obtain a linear discriminant function, denoted by g_v, as below (Duda & Hart, 1973):

$$g_v(E) = \sum_{i=1}^{n} a_i x_i - a_0 \tag{3}$$

where

$$a_i = log\frac{p_i(1-q_i)}{(1-p_i)q_i} \tag{4}$$

$$a_0 = -(\sum_{i=1}^{n} log\frac{(1-p_i)}{(1-q_i)} + log\frac{p(C=1)}{p(C=0)}) \tag{5}$$

Clearly, $E \in C = 1$ iff $g(E) \geq 1$ iff $g_v(E) \geq 0$. Thus, Naive Bayes can only represent linearly separable functions.

However, Naive Bayes can not represent *all* linearly separable functions. A special case is that Naive Bayes can not learn some m-of-n concepts (Domingos & Pazzani,1997). An m-of-n concept is a Boolean concept that is true if m or more out of n Boolean variables are true. Clearly, it is a linearly separable function. Domingos & Pazzani (1997) showed that for the concept 8-of-25, when the input Boolean variables have just six or seven 1's, Naive Bayes gives an incorrect answer of 1 (instead of 0). Their result is based on the assumption that the example space Ψ consists of all 2^n examples of the 8-of-25 concept, or is under uniform sample distribution.

3 Naive Bayes Under Uniform Distribution

In this section, we assume that the sampling distribution is uniform. We start with a few necessary definitions:

Definition 1: Two discriminant functions f_1 and f_2 are said to be equal under zero-one loss if for every example E in the example space, $f_1(E) \geq 0$ if only if $f_2(E) \geq 0$. We denote this $f_1 \doteq f_2$.

Definition 2: A linearly separable function defined by the linear function g_l is called *Naive-Bayes linearly separable (NBLS)*, if for every example E in example space, $g_l(E) \doteq g_v(E)$. In short, we say this concept is NBLS.

Let us consider a special class of linear functions g_l

$$g_l(E) = a\sum_{i=1}^{n} x_i - a_0 \tag{6}$$

That is, $a_i = a$ for all $1 \leq i \leq n$. Let x denote $\sum_{i=1}^{n} x_i$, which is the number of $x_i = 1$ in the function. We have $g_l(E) = g_l(x) = ax - a_0$. We consider $a, a_0 \geq 0$ without losing generality.

For this kind of linearly separable functions, it is obvious that $p_1 = p_2 = ... = p_n$ and $q_1 = q_2 = ... = q_n$ under uniform distribution, where p_i and q_i are defined earlier. Assume $p = p_i$, and $q = q_i$. It can be shown easily that $p \geq 0.5$, and $q \leq 0.5$.[1] Then for Naive Bayes, the discriminant function can be represented as:

$$g_v(E) = g_v(x) = a'x - a'_0 \tag{7}$$

where

$$a' = log\frac{p(1-q)}{(1-p)q} \tag{8}$$

$$a'_0 = -(nlog\frac{(1-p)}{(1-q)} + log\frac{p(C=1)}{p(C=0)}) \tag{9}$$

Clearly, we have $a' \geq 0$. We have the following theorem:
Theorem 1: For binary functions defined by Equation 6, $g_v(x) \doteq g_l(x)$ if only if $x \geq max(\frac{a_0}{a}, \frac{a'_0}{a'})$ or $x < min(\frac{a_0}{a}, \frac{a'_0}{a'})$, where $0 \leq x \leq n$, a, a_0, a', and a'_0 are given by Equation 6, 8 and 9.
Proof:
Suppose $x \geq max(\frac{a_0}{a}, \frac{a'_0}{a'})$ or $x < min(\frac{a_0}{a}, \frac{a'_0}{a'})$, then whenever $g_l(x) \geq 0$, $g_v(x) \geq 0$, and whenever $g_l(x) < 0$, $g_v < 0$ too. That is, under zero-one loss, $g_l(x)$ is equal to $g_v(x)$.

Suppose

$$g_v(x) \doteq g_l(x)$$

then $g_l(x) \geq 0$ whenever $g_v(x) \geq 0$, and $g_l(x) < 0$ whenever $g_v(x) < 0$. That is, $x \geq \frac{a_0}{a}$ and $x \geq \frac{a'_0}{a'}$, or $x < \frac{a_0}{a}$ and $x < \frac{a'_0}{a'}$.

So, we have $x \geq max(\frac{a_0}{a}, \frac{a'_0}{a'})$ or $x < min(\frac{a_0}{a}, \frac{a'_0}{a'})$.

Essentially, **Theorem 1** tells us that for any Boolean function defined by linear equation 6, if there exists some integer j in the following left-closed and right-open interval:

$$[min(\frac{a_0}{a}, \frac{a'_0}{a'}), max(\frac{a_0}{a}, \frac{a'_0}{a'}))$$

then for any examples with $x = \sum x_i = j$, Naive Bayes will make an incorrect prediction; for all other examples, Naive Bayes agrees with the target function values.

The Boolean functions defined by Equation 6 are essentially the m-of-n concepts where $a = 1$ and $a_0 = m$. The discriminant function of the m-of-n concept, denoted by g_{m-n}, is thus:

$$g_{m-n}(E) = g_{m-n}(x) = x - m$$

[1] Under uniform distribution, $p(X_i = 1) = 0.5$. Since some X_i must be 1 for $C = 1$, $p = p(X_i = 1|C = 1) \geq 0.5$, and $q = p(X_i = 1|C = 0) \leq 0.5$.

Applying **Theorem 1** to m-of-n concept, we have the following corollary.
Corollary 1: Naive Bayes discriminant function $g_v(x) \doteq g_{m-n}(x)$ on an arbitrary m-of-n concept if only if $x \geq max(m, \frac{a_0'}{a'})$ or $x < min(m, \frac{a_0'}{a'})$, where $0 \leq x \leq n$, a, a_0, a', and a_0' are given by Equation 6, 8 and 9. The corresponding probabilities can now be obtained explicitly (Domingos & Pazzani, 1997):

$$p(C=1) = \frac{\sum_{i=m}^{n} \binom{n}{i}}{2^n}$$

$$p(C=0) = \frac{\sum_{i=0}^{m-1} \binom{n}{i}}{2^n}$$

$$p = p(x_i = 1|C=1) = \frac{\sum_{i=m-1}^{n-1} \binom{n-1}{i}}{\sum_{i=m}^{n} \binom{n}{i}}$$

$$q = p(x_i = 1|C=0) = \frac{\sum_{i=0}^{m-2} \binom{n-1}{i}}{\sum_{i=0}^{m-1} \binom{n}{i}}$$

Corollary 1 presents necessary and sufficient condition for training examples of the m-of-n concept to be equivalent to the target functions. The next theorem tells us exactly what kinds of m-of-n concepts are NBLS; that is, $g_v(x) \doteq g_{m-n}(x)$ for all x.
Theorem 2: An m-of-n concept is NBLS, if only if its corresponding Naive Bayes discriminant function Equation 7 satisfies the following condition:

$$\frac{a_0'}{a'} \in (m-1, m]$$

Proof: Let N={1,2,...,n},Δ represents a condition. $S_N(\Delta)$ denotes a set $\{y|y \in N$ and y satisfies $\Delta\}$. Suppose $\frac{a_0'}{a'} \in (m-1, m]$.
Then $max(m, \frac{a_0'}{a'}) = m$, $min(m, \frac{a_0'}{a'}) = \frac{a_0'}{a'}$. So

$$S_N(x \geq max(m, \frac{a_0'}{a'})) = S_N(x \geq m) = \{m, m+1, ..., n\}$$

$$S_N(x < min(m, \frac{a_0'}{a'})) = S_N(x < m) = \{m-1, m-2, ..., 1\}$$

$$S_N(x \geq max(m, \frac{a_0'}{a'})) \bigcup S_N(x < min(m, \frac{a_0'}{a'}) = N.$$

According to **Corollary 1**, for any $x \in N$, $g_v(x) \doteq g_{m-n}(x)$. That is, the m-of-n concept is NBLS.

Suppose a m-of-n concept Ω is NBLS. Then for any $x \in N$, $g_v(x) \doteq g_{m-n}(x)$. There are three cases as below:
(1) $\left| m - \frac{a'_0}{a'} \right| > 1$
(2) $\left| m - \frac{a'_0}{a'} \right| = 1$
(2) $\left| m - \frac{a'_0}{a'} \right| < 1$
In the first case, there is at least one integer $j \in N$ such that $m < j < \frac{a'_0}{a'}$ or $\frac{a'_0}{a'} < j < m$. From **Corollary 1**, $g_v(j)$ not $\doteq g_{m-n}(j)$.
In the second case, $\frac{a'_0}{a'} = m - 1$ or $\frac{a'_0}{a'} = m + 1$.
For the former one, we have $g_v(m-1)$ not $\doteq g_{m-n}(m-1)$.
For the latter one, we have $g_v(m)$ not $\doteq g_{m-n}(m)$.
In the third case, $\frac{a'_0}{a'} \in (m-1, m]$ or $\frac{a'_0}{a'} \in [m, m+1)$
If $\frac{a'_0}{a'} \in [m, m+1)$, $max(m, \frac{a'_0}{a'}) = \frac{a'_0}{a'}$, $min(m, \frac{a'_0}{a'}) = m$.
It is easy to show $g_v(m)$ not $\doteq g_{m-n}(m)$. Therefore $\frac{a'_0}{a'} \in (m-1, m]$.

Since $\frac{a'_0}{a'}$ is a function of only m and n, **Theorem 2** gives us a necessary and sufficient condition on m and n for any m-of-n concept to be representable by Naive Bayes under uniform distribution. Let L_n denote a set of m such that the corresponding m-of-n concepts are NBLS. Here are several examples: $L_{24} = \{12, 13\}$, $L_{25} = \{13\}$, $L_{60} = \{30, 31\}$, $L_{61} = \{31\}$, $L_{134} = \{67, 68\}$, and $L_{135} = \{68\}$. From the above results, we have the following:
Conjecture: Any m-of-n concept is NBLS if and only if $m = \frac{n}{2}$ or $\frac{n}{2} + 1$ when n is even, and $m = \frac{n+1}{2}$ when n is odd.

We have verified this conjecture to be true for $6 \leq n \leq 1322$, but we have not been able to formally prove it. Nevertheless, it is clear that only a small portion of the m-of-n concepts is NBLS.

4 Sampling Distribution

The above discussions on the representation ability of Naive Bayes are based on an assumption that the sampling distribution is uniform. One might tend to think that such learnability results can be naturally extended to other sampling distributions. After all, the representation power of most other learning algorithms, such as decision tree and neural network learning algorithms, is not affected by the sampling distribution of the training examples. Perceptions, for example, can always represent any linearly separable functions no matter how the training (and testing) sets are sampled.

Surprisingly, this is not true for Naive Bayes.

When the sampling distribution is changed, the probabilities p_i, q_i and $p(C = 1)/p(C = 0)$ in Equation 4 5 can be changed, causing the discrimination function of Naive Bayes to be different from the one under uniform distribution. This may alter the learnability (and the error rate) of Naive Bayes.

A simple way to see the effect of various sampling distributions on Naive Bayes is to use what we call the "two-tiered" sampling: uniform sampling is still applied on positive and negative examples respectively, but with various ratios (under uniform distribution, the ratio is fixed and is determined by the target function). For example, positive examples may be sampled 10 (or any number of) times more likely than the negative examples. That way, all of the p_i and q_i are kept the same as the ones under uniform distribution, but the term $p(C = 1)/p(C = 0)$ can be changed to arbitrary values between $(0, \infty)$. This allows us to "manipulate" the constant term a_0 in the linear discriminant function (Equation 5) *alone* to study the learnability of Naive Bayes.

Figure 1 shows the error rate under uniform distribution on three m-of-25 concepts when $p(C = 1)$ ranges from 0.01 to 0.99 (and $p(C = 0) = 1-p(C = 1)$). The small dark dot on each curve represents the result of that concept under uniform distribution.

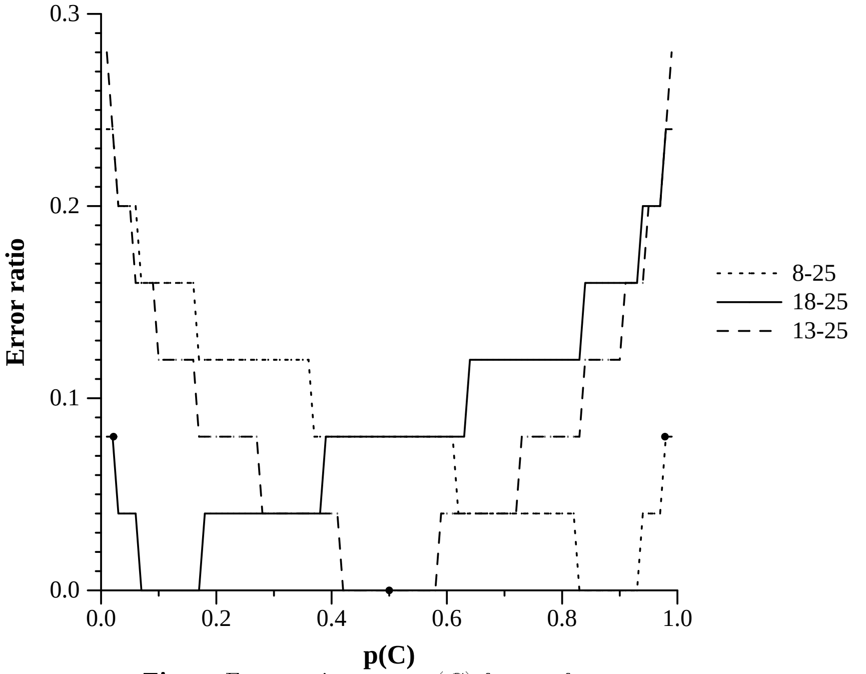

Fig. 1. Error ratio versus $p(C)$ for m-of-n concepts.

8-25: m=8, n=25; 18-25: m=18, n=25; 13-25: m=13, n=25.
•: the point on uniform sampling.

Several interesting observations are obtained. First of all, as we discussed earlier, 13-of-25 concept is NBLS under uniform distribution, and we can verify this from the figure: the dot representing uniform distribution on its curve has an error rate of 0. But 8-of-25 and 18-of-25 are not NBLS under uniform distribution, and their dots indicate an error rate of about 8%. Second, NBLS functions under uniform distribution can become non-NBLS under other distributions. For example, 13-of-25, which is NBLS under uniform distribution, becomes non-NBLS when $p(C = 1)$ is greater than 0.59 or less than 0.42. In fact, the error rate of 13-of-25 can become quite large (as high as 50%; not shown in the figure) under extreme distributions when $p(C = 1)$ is approaching 0 or 1. Third, non-NBLS functions under uniform distribution (such as 8-of-25 and 18-of-25) can become NBLS if the sampling distribution changes. As we can see from their curves, not far from their respective sampling rate of positive examples $p(C = 1)$ under uniform distribution, the curves reach 0% error rate under other sampling rates.

Figure 3 is similar to Figure 1, we can see clearly again that some linear function (such as $m, n, a_0 = 19, 25, 0$) that is NBLS under uniform distribution is non-NBLS under other distributions, and vice versa.

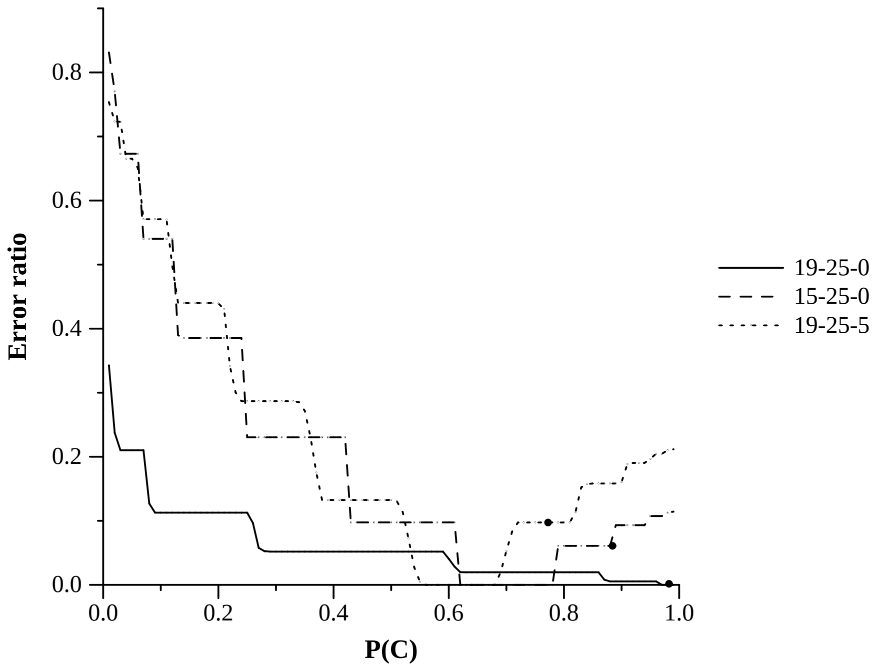

Fig. 2. Error ratio versus $p(C)$ for example 1.

19-25-0: m=19, n=25, a_0=0; 15-25-0: m=15, n=25, a_0=0;
19-25-5: m=19, n=25, a_0=5; •: the point on uniform sampling.

The error curve for linearly separable function $\sum ix_i - a_0 \geq 0$ (Example 2 in Section 3) is plotted in Figure 2. It is very interesting to observe that under uniform distribution, the functions are not NBLS. When $p(C = 1)$ changes, the error rates may become very small (1.8%) to very large (75.1%), but never 0%. Of course it does not mean that this class of functions are not NBLS under any distribution, since changing $p(C=1)$ is just one way of changing the sampling distribution. It is unknown to us if these functions can become NBLS under some other distributions.

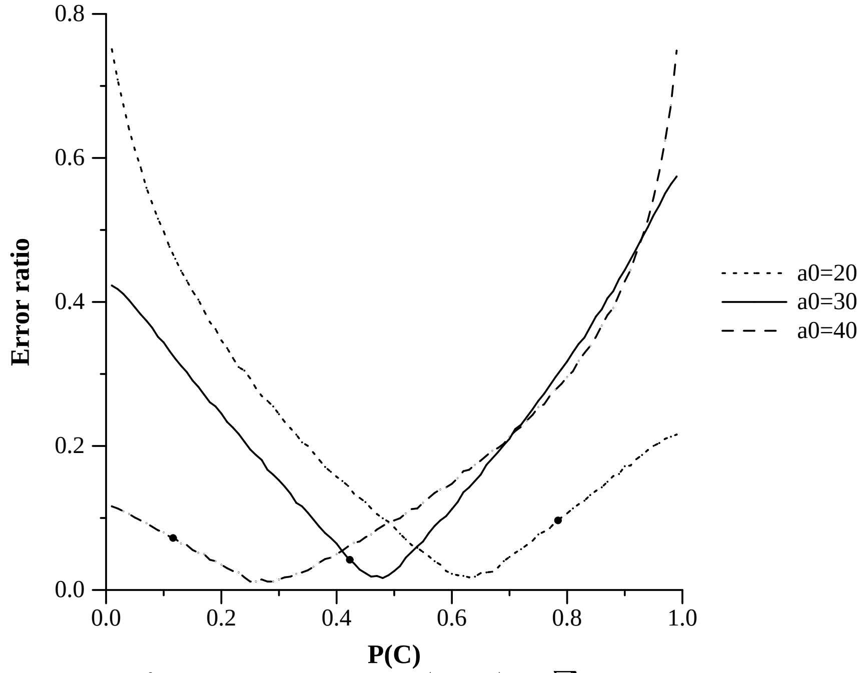

Fig. 3. Error ratio versus $p(C=1)$ for $\sum ix_i - a_0 \geq 0$.

•: the point on uniform sampling.

Nevertheless, Naive Bayes cannot learn any non-linearly separable Boolean functions in the binary domain under any non-zero sampling distribution[2]. This is simply because no matter how we alter the sampling distribution, the discriminating function of Naive Bayes is always a linear function.

5 Conclusions

Naive Bayes is an efficient and effective learning algorithm. But previous results show that its representation ability is severely limited since it cannot represent any non-linearly separable functions in the binary domain. Further, it cannot represent all linearly separable functions, as several counter examples given

[2] Non-zero distributions guarantee a non-zero sampling probability on each example in the example space.

in (Domingos & Pazzani,1997). In this paper we extend Domingos & Pazzani (1997)'s results, and give necessary and sufficient conditions on linearly separable functions in the binary domain to be learnable by Naive Bayes under uniform representation. We demonstrate further that the learnability and error rates of Naive Bayes can be affected dramatically by the sampling distribution of the training examples. NBLS functions can become non-NBLS, and vice versa.

Our results help us to gain a much deeper understanding of this seemingly simple but powerful learning algorithm. Naive Bayes' learning capacities and error rates are not only determined by the target function, but also by the sampling distribution, and how the attributes are represented.

References

1. Domingos, P., & Pazzani, M. (1997): Beyond independence: Conditions of the optimality of the simple bayesian classifier. Machine Learning, 29, 103-130.
2. Dougherty, J., Kohavi, R., & Sahami, M.(1995): Supervised and unsupervised discretization of Continuous Features. In Proceedings of the Twelfth International Conference on Machine Learning, pp. 194-202. Morgan Kaufmann.
3. Duda, R. O., & Hart, P. E. (1973): Pattern classification and scene analysis. A Wiley Interscience Publication.
4. Kononenko, I. (1990): Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In Wielinga, B. (Ed.), Current Trends in Knowledge Acquisition. IOS Press.
5. Langley, P., Iba, W., & Thomas, K.(1992). An analysis of Bayesian classifier. In proceedings of the Tenth National Conference of Artificial Intelligence, pp. 223-228. AAAI Press.
6. Pazzani, M., Muramatsu, J., & Billsus, B. (1996). Syskill & webert: Identifying interesting websites. In Proceedings of the Thirteenth National Conference of Artificial Intelligence. pp. 54-62. AAAI Press.
7. Quinlan, J. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA.