IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Pragmatic Study on Naive Bayes Classifier

Khushal singh
IILM College of Engineering & Technology
Greater Noida, India

Praveenkumarsengar
IIMT College of Engineering
Greater Noida, India

Uday krishan kol
IILM College of Engineering & Technology
New Delhi , India

Abstract

The naive Bayes classifier greatly simplify learning by assuming that features are independent given class. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers.

Our broad goal is to understand the data characteristics which affect the performance of naive Bayes. Our approach uses Monte Carlo simulations that allow a systematic study of classification accuracy for several classes of randomly generated problems. We analyze the impact of the distribution entropy on the classification error, showing that low-entropy feature distributions yield good performance of naive Bayes. We also demonstrate that naive Bayes works well for certain nearlyfunctional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features (as expected) and functionally dependent features (which is surprising). Another surprising result is that the accuracy of naive Bayes is not directly correlated with the degree of feature dependencies measured as the classconditional mutual information between the features. Instead, a better predictor of naive Bayes accuracy is the amount of information about the class that is lost because of the independence assumption.

1Introduction

Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent given class, that is, , where is a feature vector and is a class.

Despite this unrealistic assumption, the resulting classifier known as *naive Bayes* is remarkably successful in practice, often competing with much more sophisticated techniques [6; 8; 4; 2]. Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management [2; 9; 5].

T.J. Watson Research Center, 30 Saw Mill River Road, Hawthorne, NY 10532. Phone +1 (914) 784-7431

The success of naive Bayes in the presence of feature dependencies can be explained as follows: optimality in terms of zero-one loss (classification error) is not necessarily related to the quality of the fit to a probability distribution (i.e., the appropriateness of the independence assumption). Rather, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class [2]. For example, [2] provenaive Bayes optimality for some problems classes that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts.

However, this explanation is too general and therefore not very informative. Ultimately, we would like to understand the data characteristics which affect the performance of naive Bayes. While most of the work on naive Bayes compares its performance to other classifiers on particular benchmark problems (e.g., UCI benchmarks), our approach uses Monte Carlo simulations that allow a more systematic study of classification accuracy on parametric families of randomly generated problems. Also, our current analysis is focused only on the *bias* of naive Bayes classifier, not on its *variance*. Namely, we assume an infinite amount of data (i.e., a perfect knowledge of data distribution) which allows us to separate the approximation error (bias) of naive Bayes from the error induced by training sample set size (variance).

2 Definitions and Background

Let be a vector of observed random variables, called *features*, where each feature takes values from its *domain* . The set of all feature vectors (*examples*, or *states*), is denoted . Let be an unobserved random variable denoting the *class* of an example, where can take one of values Capital letters, such as , will denote variables, while lower-case letters, such as , will denote their values; boldface letters will denote vectors.

A function , where , denotes a *concept* to be learned. Deterministic corresponds to a concept without noise, which always assigns the same class to a given example (e.g., disjunctive and conjunctive concepts are deterministic). In

general, however, a concept can be *noisy*, yielding a random function .

A classifier is defined by a (deterministic) function

(a *hypothesis*) that assigns a class to any given example. A common approach is to associate each class with a discriminant function ,

, and let the classifier select the class with maximum discriminant function on a given example:

.

The *Bayes* classifier (that we also call *Bayes-optimal* classifier and denote), uses as discriminant functions the class posterior probabilities given a feature vector, i.e.

Applying Bayes rule gives_____ , whereis identical for all classes, and therefore can be ignored. This yields Bayes discriminant functions

(1)

where is called the *class-conditional probability distribution (CPD)*. Thus, the Bayes classifier

3 When does naive Bayes work well? Effects of some nearly-deterministic dependencies

In this section, we discuss known limitations of naive Bayes and then some conditions of its optimality and nearoptimality, that include low-entropy feature distributions and nearly-functional feature dependencies.

3.1 Concepts without noise

We focus first on concepts with or for any and (i.e. no noise), which therefore have zero Bayes risk. The features are assumed to have finite domains (-th feature has values), and are often called *nominal*. (A nominal feature can be transformed into a numeric one by imposing an order on its domain.) Our attention will be restricted to binary classification problems where the class is either 0 or 1.

Some limitations of naive Bayes are well-known: in case of binary features (for all), it can only learn linear discriminant functions [3], and thus it is always suboptimal for non-linearly separable concepts (the classical example is XOR function; another one is -of- concepts [7; 2]). When for some features, naive Bayes is able to learn (some) polynomial discriminant functions [3]; thus, polynomial separability is a necessary, although not sufficient, condition of naive Bayes optimality for concepts with finite-domain features.

Despite its limitations, naive Bayes was shown to be optimal for some important classes of concepts that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts [2]. These results can be generalized to concepts with any nominal features (see [10] for details):

Theorem 1 *[10] The naive Bayes classifier is optimal for any two-class concept with nominal features that assigns class 0 to exactly one example, and class 1 to the other examples, with probability 1.* [1]

The performance of naive Bayes degrades with increasing number of class-0 examples (i.e., with increasing prior

, also denoted), as demonstrated in Figure 1a. This figure plots average naive Bayes error computed over 1000 problem instances generated randomly for each value of . The problem generator, called **ZeroBayesRisk**, assumes features (here we only consider two features), each having values, and varies the number of class-0 examples from 1 to (so that varies from to 0.5; the results for are symmetric)[2]. As expected, larger (equivalently, larger), yield a wider range of problems with various dependencies among features, which result into increased errors of Bayes; a closer look at the data shows no other cases of optimality besides .

3.2 Noisy concepts

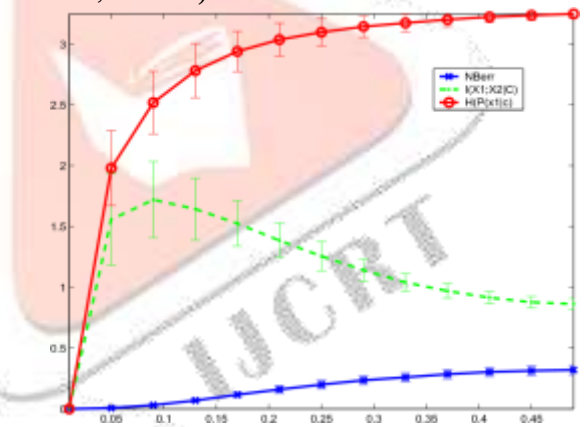
Low-entropy feature distributions

Generally, concepts can be noisy, i.e. can have nondeterministic and thus a non-zero Bayes risk.

A natural extension of the conditions of Theorem 1 to noisy concepts yields low-entropy, or "extreme", probability distributions, having almost all the probability mass concentrated in one state. Indeed, as shown in [10], the independence assumption becomes more accurate with decreasing entropy which yields an asymptotically optimal performance of naive Bayes. Namely,

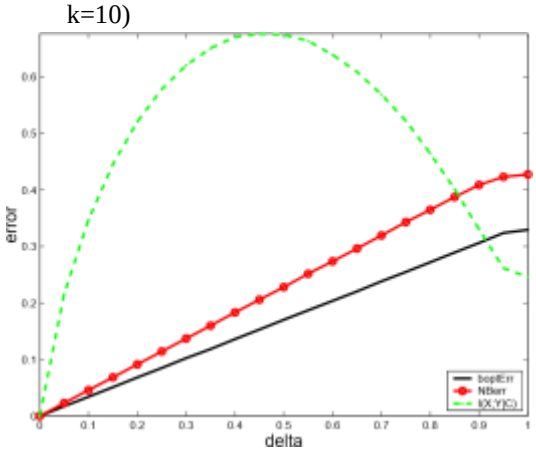
Theorem 2 *[10] Given that one of the following conditions hold:*

NBerror, I(X1;X2|C), and H(P(x1|c) vs. P(0) (n=2, m=2, k=10, N=1000)



(a)

Average errors vs. mutual information (n=2, m=2, k=10)

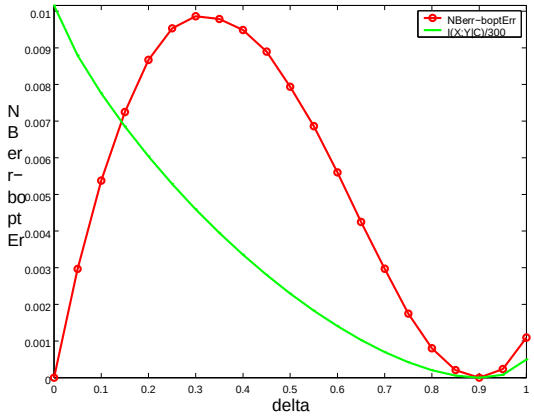


(b)

[1] Clearly, this also holds in case of a single example of class 1.

[2] Note that in all experiments perfect knowledge of data distribution (i.e., infinite amount of data)is assumed in order to avoid the effect of finite sample size.

Average error difference vs. mutual information (n=2, m=2, k=10)



(c)

Figure 1: (a) results for the generator **ZeroBayesRisk** (k=10, 1000 instances): average naive Bayes error (NBerr), class-conditional mutual information between features ( ), and entropy of marginal distribution,

; the error bars correspond to the standard deviation of each measurement across 1000 problem instances; (b) Results for the generator **EXTREME**: average Bayes and naive Bayes errors and average ; (c) results for the generator **FUNC1**: average difference between naive Bayes error and Bayes error (- constant for all), and scaled I(X1;X2—C) (divided by 300).

1. *a joint probability distribution is such that for some state , or*
2. *a set of marginal probability distributions is such that for each , for some , then .*

The performance of naive Bayes on low-entropy distributions is demonstrated using a random problem generator called **EXTREME**. This generator takes the number of classes, , number of features, , number of values per feature, , and the parameter , and creates class-conditional feature distributions, each satisfying the condition

if , where the are different states randomly selected from possible states. For each class , the remaining probability mass in is randomly distributed among the remaining states. Class prior distributions are uniform. Once is generated, naive Bayes classifier (NB) is compared against the Bayes-optimal classifier (BO).

Figure 1b shows that, as expected, the naive Bayes error (both the average and the maximum) converges to zero with

(simulation performed on a set of 500 problems with , ,). Note that, similarly to the previous observations, the error of naive Bayes is not a monotone function of the strength of feature dependencies; namely, the average class-conditional mutual information plotted in Figure 1b is a concave function reaching its maximum between and , while the decrease of average naive Bayes error is monotone in .

Almost-functional feature dependencies

Another ”counterintuitive” example that demonstrates the non-monotonic relation between the feature dependence and the naive Bayes accuracy is the case of certain functional and nearly-functional dependencies among features. Formally,

Theorem 3 *[10] Given equal class priors, Naive Bayes is optimal if for every feature , , where is a one-to-one mapping* [3].

Namely, naive Bayes can be optimal in situations just opposite to the class-conditional feature independence (when mutual information is at minimum) - namely, in cases of completely deterministic dependence among the features (when mutual information achieves its maximum). For example, Figure 1c plots the simulations results obtained using an ”nearly-functional” feature distribution generator called **FUNC1**, which assumes uniform class priors, two features, each having values, and ”relaxes” functional dependencies between the features using the noise parameter conditional joint feature distributions satisfying the following conditions:

and

This way the states satisfying functional dependence obtain probability mass, so that by controlling we can get as close as we want to the functional dependence described before, i.e. the generator relaxes the conditions of Theorem 3. Note that, on the other hand, gives us uniform distributions over the second feature

, which makes it independent of (given class). Thus varying from 0 to 1 explores the whole range from deterministic dependence to complete independence between the features given class.

The results for 500 problems with are summarized in Figure 1c, which plots the difference between the average naive Bayes error and average Bayes risk (which turned out to be , a constant for all) is plotted against . We can see that naive Bayes is optimal when (functional dependence) and when (complete independence), while its maximum error is reached between the two extremes. On the other hand, the class-conditional mutual information decreases monotonically in , from its maximum at (functional dependencies) to its minimum at (complete independence)[4].

4 Conclusions

Despite its unrealistic independence assumption, the naive Bayes classifier is surprisingly effective in practice since its classification decision may often be correct even if its probability estimates are inaccurate. Although some optimality conditions of naive Bayes have been already identified in the past [2], a deeper understanding of data characteristics that affect the performance of naive Bayes is still required.

Our broad goal is to understand the data characteristics which affect the performance of naive Bayes. Our approach uses Monte Carlo simulations that allow a systematic study of classification accuracy for several classes of randomly generated problems. We analyze the impact of the distribution entropy on the classification error, showing that certain almostdeterministic, or low-entropy, dependencies yield good performance of naive

[3] A similar observation was made in [11], but the important ”oneto-one” condition on functional dependencies was not mentioned there. However, it easy to construct an example of a non-ne-toone functional dependence between the features that yields non-zero error of naive Bayes.

[4] Note that the mutual information in Figure 1c is scaled (divided by 300) to fit the error range.

Bayes. Particularly, we demonstrate that naive Bayes works best in two cases: completely independent features (as expected) and functionally dependent features (which is surprising). Naive Bayes has its worst performance between these extremes.

Acknowledgements

We wish to thank Mark Brodie, Vittorio Castelli, Joseph Hellerstein, Jayram Thathachar, Daniel Oblinger, and Ricardo Vilalta for many insightful discussions that contributed to the ideas of this paper.

5 References

[1] T.M. Cover and J.A. Thomas. *Elements of information theory*. New York:John Wiley & Sons, 1991.

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

[3] R.O. Duda and P.E.Hart. *Pattern classification and scene analysis*. New York: John Wiley and Sons, 1973.

[4] N. Friedman, D. Geiger, and Goldszmidt M. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[5] J. Hellerstein, Jayram Thathachar, and I. Rish. Recognizing end-user transactions in performance management. In *Proceedings of AAAI-2000*, pages 596–602, Austin, Texas, 2000.

[6] J. Hilden. Statistical diagnosis based on conditional independence does not require it. *Comput. Biol. Med.*, 14(4):429–435, 1984.