# Naïve Bayes vs. Logistic Regression: An Assessment of the Impact of the Misclassification Cost

**Vidit Jain**
University of Massachusetts Amherst
Amherst MA USA
vidit@cs.umass.edu

## Abstract

Recent advances in the asymptotic characterization of generative and discriminative learning have suggested several ways to develop more effective hybrid models. An application of these suggested approaches to a practical problem domain remains non-trivial, perhaps due to the violation of various underlying assumptions. One common assumption corresponds to the choice of equal misclassification cost or the ability to estimate such a cost. Here, we investigate the effect of this misclassification cost on the comparison between naïve Bayes and logistic regression. To assess the utility of this comparison for practical domains, we include a comparison of mean average precision values for our experiments. We present the empirical comparison patterns on the LETOR data sets to encourage the development of supporting theoretical results.

## 1 Introduction

An overwhelming number of probabilistic models have been developed recently for application domains such as computer vision and information retrieval. For a given task, selecting an appropriate model from such a large pool requires a common framework to compare different models. Such a comparison would not only aid in model selection, but also is likely to lead to the design of more effective hybrid models.

Generative and discriminative models are two classes of graphical models that have attracted much attention recently. As a result, there has been significant effort spent on comparative studies of generative and discriminative learning. One such comparison is due to Ng & Jordan [5]. In their theoretical analysis, they compared the empirical risk minimization for linear classifiers with naïve Bayes classifier. Their results suggest that even though the discriminative model (logistic regression) has a lower asymptotic error, the generative model (naïve Bayes) has a faster convergence towards the asymptotic error. Thus, with only a handful of training examples, naïve Bayes performs better than logistic regression, but with more training examples, the latter outperform the former. Later, Liang & Jordan [2] presented a unified framework for studying the comparison between generative and discriminative estimators, and concluded that when the model is well-specified, the asymptotic error for generative models is less than that of discriminative models, whereas when the model is mis-specified ( i.e., the approximation error is not zero), discriminative models have lower approximation and asymptotic estimation errors.

Applying the insights developed through these theoretical results to practical data sets remains difficult, perhaps due to the violation of some of the underlying assumptions made in these analyses. For instance, while evaluating different classification algorithms, the misclassification costs are often assumed to be equal. In other words, the cost of misclassifying a positive example as negative (false negative) is assumed to be the same as the cost of misclassifying a negative example as positive (false positive). This assumption of equal costs has not been useful for the evaluation of various practical tasks such as learning to rank in the context of information retrieval. Moreover, the choice of misclassification costs depends on the task and the domain, and does *not* depend on the data. In this work, we investigate the impact of the choice of these cost parameters on the comparison between two classifiers in a generative-discriminative pair.

Furthermore, the relatively small sizes of the data sets used in previous experiments have obscured the effects of the factors (e.g., the lower-order terms in the asymptotic analysis) that could be manifested only in large data sets or data sets with complex data distributions. Thus the generalizability of the conclusions drawn from those experiments is limited. Hence, to develop further insights about the comparison between generative and discriminative learning, we perform our experiments on the LETOR data sets [3] that are used as benchmarks in information retrieval research.

Although information retrieval research has a great emphasis on comparative studies, there has been little work focused on studying the generative-discriminative distinction. One such comparison was done by Gey [1], which includes a comparison of logistic regression and the vector space model (a generative model) for the ad-hoc retrieval task. The difference in performances of these two approaches was not found to be statistically significant. It was argued that this observation was due to large differences between the frequencies of classes in the train and test collections. The data sets used in these experiments were very small, and it is unclear if these results would hold for larger data sets. Later, Nallapati [4] did a comparison between SVM (discriminative) and language modeling (generative) for ad-hoc retrieval task, and observed similar performances for both of these approaches. In their experiments, an under-sampling of the majority class is done to circumvent the problem of unbalanced data. This artifact in their data set raises concern about the generalizability of these results in practical scenarios, where the expected class frequencies are unknown.

In this work, we investigate the effect of the misclassification costs on the comparison between naïve Bayes and logistic regression.[1] To assess the relevance of the theoretical results for practical applications, we present a comparison of these approaches using two different evaluation metrics: area under ROC curve (AUC) and mean average precision (MAP), where MAP refers to the mean of the area under the precision-recall curve, over all the queries. AUC is a standard evaluation metric for classification, whereas MAP is commonly used to evaluate ranked results (or rankings).

## 2 Experiments

To benchmark learning to rank algorithms, Liu et al. [3] compiled data from two TREC data sets (TD2003 and TD2004) and the OHSUMED corpus, and referred to the collection as the LETOR data sets. These data sets have 50, 75, and 106 distinct queries, and a total of 16141, 49172, and 74171 documents, respectively. Each query-document pair in these data sets was annotated with one of the three relevance levels: *non-relevant*, *probably relevant*, and *relevant*. A set of pre-computed features (e.g., *tf-idf* terms) that were found to be useful for predicting relevance are also computed for each of these pairs. For our experiments[2], we treat *probably relevant* as *relevant*. Here, we are comparing three approaches: naïve Bayes using a normal distribution for each feature, naïve Bayes with kernel density estimation, and logistic regression. Note that the first approach assumes an independent normal distribution for each of the features.

We first examine the relative performances of different approaches against the percentage of the training data used. in particular, we under-sampled the training set at different fractions to make observations similar to Ng & Jordan [5] . Figure 1 shows the AUC and MAP values as a function of the fraction of the original train set.

Despite several efforts during the collection stage, some characteristics of the resulting data sets do not match the related observations in the practical domain. For instance, the relative frequencies of non-relevant documents to relevant documents for OHSUMED, TD2003, and TD2004 is 2.3, 94.2, and 166.0, respectively. There may exist some application domains where this value may be representative of real observations, but at the same time, for document collections in domains such as the internet, this value is expected to be much more extreme. Thus, to draw conclusions about the performances of different approaches that can be generalized across domains, this disparity between training and test sets must be carefully investigated. One way to characterize the evaluation

[1] These two classifiers are chosen because they form a generative-discriminative pair [5].

[2] The results for 10-fold cross-validation are reported on all three of the LETOR data sets.

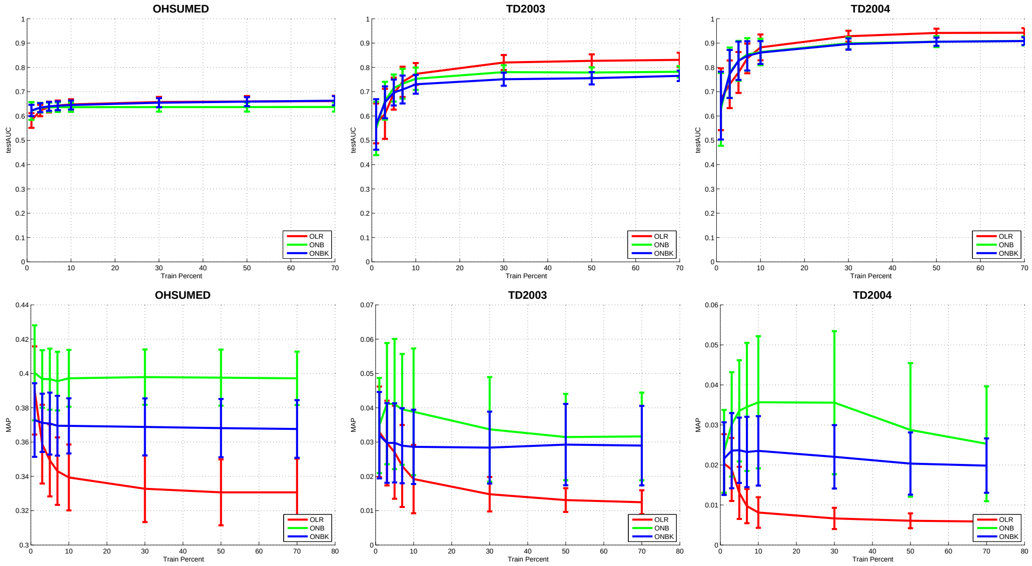

Figure 1: **Area under ROC curve** (top) and **Mean Average Precision** (bottom) curves as a function of the number of training examples. The three columns correspond to OHSUMED, TD2003, and TD2004 data sets, respectively. Using AUC for evaluation, logistic regression (red) performs better than naïve Bayes (green) and naïve Bayes with kernel density estimation (blue) approaches, but the differences are not statistically significant. The trend is reversed when MAP is used as the evaluation metric.

requirements of a particular domain is through the use of a cost matrix representing the relative costs of different kind of errors. Here, we investigate the effect of choosing different cost matrices on the comparative evaluation of naïve Bayes and logistic regression. In particular, we used a diagonal cost-matrix $[\lambda_1\ 0;\ 0\ \lambda_2]$ and varied the ratio of the non-zero entries $\frac{\lambda_1}{\lambda_2}$. Figure 2 shows the AUC and MAP values as a function of this parameter.

## 3 Discussion

The experiment evaluating the performance of the trained classifier with varying number of training examples (as seen in the Figure 1) displays similar AUC values for all the three approaches, with logistic regression doing marginally better than other approaches for the TREC data sets. This is perhaps due to the greater number of examples in these data sets than in the OHSUMED corpus. This observation is consistent with that of Ng & Jordan [5]. The MAP values, however, show a reverse trend.

The AUC values for the experiments with varying misclassification cost are shown in Figure 2. As is expected, the MAP values for this experiment demonstrate a non-monotonic behavior. One consistent observation is that the AUC and MAP values for both naïve Bayes with kernel density estimation and logistic regression drops as the difference between the misclassification cost and empirical relative class-frequency increases. For the most part, logistic regression outperforms naïve Bayes with kernel density estimation. These results appear to be consistent with Liang & Jordan's conclusions [2].

One standing intuition in the information retrieval community is that most of the evaluation metrics track each other closely. While this intuition is supported by our experiments, to the best of our knowledge, a theoretical characterization (beyond the selection of a single cost ratio as the maximum) for the expected behavior of these metrics, particularly the MAP values, does not exist. We present these results with the hope that they will: (a) provide the experimental basis for a theoretical analysis of mean average precision values and other measures used for tasks other than classification, and (b) lead to discussions about the practical implications of these theoretical results.

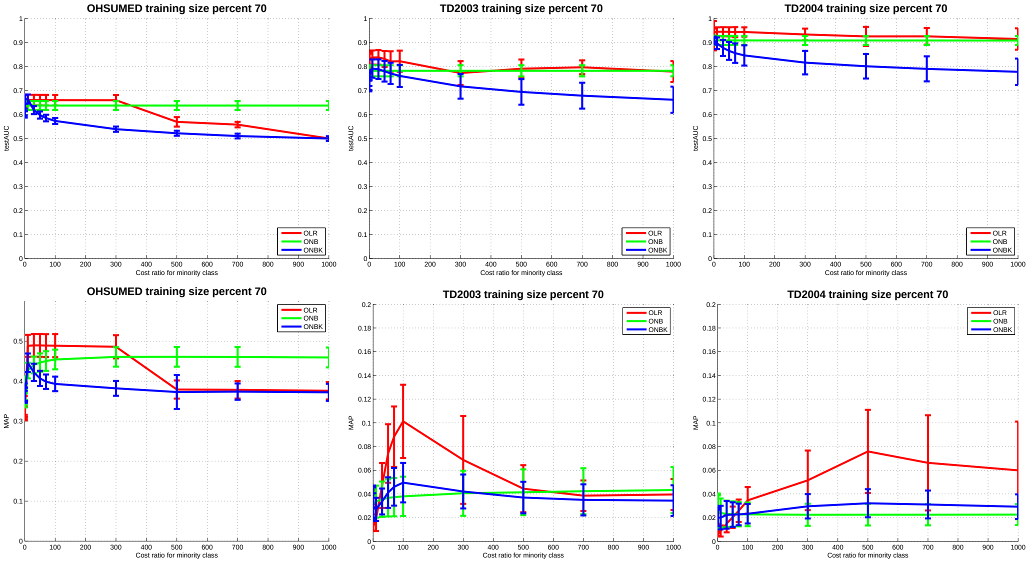

Figure 2: **Area under ROC curve** (top) and **Mean Average Precision** (bottom) curves as a function of the ratio of the misclassification costs used for training the classifier. The three columns correspond to OHSUMED, TD2003, and TD2004 data sets, respectively. For the TREC data sets (last two columns), logistic regression (red) performs better than the naïve Bayes (green) and naïve Bayes with kernel density estimation (blue) approaches. For the OHSUMED data set, which has relatively more balanced class frequencies, the performance of logistic regression dips sharply with increasing model mismatch. Also note that the performance for naïve Bayes classifier (green) has low dependency on the assumed ratio of the misclassification costs.

### Acknowledgments

The author thanks James Allan for providing a sounding board and useful advice, Erik Learned-Miller for invaluable discussions, and Niranjan Balasubramanian for help with experimental setup. This work was supported in part by National Science Foundation under CAREER award IIS-0546666.

## References

[1] F. C. Gey. Inferring probability of relevance using the method of logistic regression. In W. B. Croft and C. J. van Rijsbergen, editors, *ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 222–231. ACM/Springer, 1994.

[2] P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning*, 2008.

[3] T. Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2007.

[4] R. Nallapati. Discriminative models for information retrieval. In *ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2004.

[5] A. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*. MIT Press, 2002.