On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes

Andrew Y. Ng
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720

Michael I. Jordan
C.S. Div. & Dept. of Stat.
University of California, Berkeley
Berkeley, CA 94720

Abstract

We compare discriminative and generative learning as typified by logistic regression and naive Bayes. We show, contrary to a widely-held belief that discriminative classifiers are almost always to be preferred, that there can often be two distinct regimes of performance as the training set size is increased, one in which each algorithm does better. This stems from the observation—which is borne out in repeated experiments—that while discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster.

1 Introduction

Generative classifiers learn a model of the joint probability, $p(x, y)$, of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $p(y|x)$, and then picking the most likely label y. Discriminative classifiers model the posterior $p(y|x)$ directly, or learn a direct map from inputs x to the class labels. There are several compelling reasons for using discriminative rather than generative classifiers, one of which, succinctly articulated by Vapnik [6], is that "one should solve the [classification] problem directly and never solve a more general problem as an intermediate step [such as modeling $p(x|y)$]." Indeed, leaving aside computational issues and matters such as handling missing data, the prevailing consensus seems to be that discriminative classifiers are almost always to be preferred to generative ones.

Another piece of prevailing folk wisdom is that the number of examples needed to fit a model is often roughly linear in the number of free parameters of a model. This has its theoretical basis in the observation that for "many" models, the VC dimension is roughly linear or at most some low-order polynomial in the number of parameters (see, e.g., [1, 3]), and it is known that sample complexity *in the discriminative setting* is linear in the VC dimension [6].

In this paper, we study empirically and theoretically the extent to which these beliefs are true. A parametric family of probabilistic models $p(x, y)$ can be fit either to optimize the joint likelihood of the inputs and the labels, or fit to optimize the conditional likelihood $p(y|x)$, or even fit to minimize the 0-1 training error obtained by thresholding $p(y|x)$ to make predictions. Given a classifier h_{Gen} fit according to the first criterion, and a model h_{Dis} fit according to either the second or the third criterion (using the same parametric family of models), we call h_{Gen} and h_{Dis} a *Generative-Discriminative pair*. For example, if $p(x|y)$ is Gaussian and $p(y)$ is multinomial, then the corresponding Generative-Discriminative pair is Normal Discriminant Analysis and logistic regression. Similarly, for the case of discrete inputs it is also well known that the naive Bayes classifier and logistic regression form a Generative-Discriminative pair [4, 5].

To compare generative and discriminative learning, it seems natural to focus on such pairs. In this paper, we consider the naive Bayes model (for both discrete and continuous inputs) and its discriminative analog, logistic regression/linear classification, and show: (a) The generative model does indeed have a higher asymptotic error (as the number of training examples becomes large) than the discriminative model, but (b) The generative model may also approach its asymptotic error much faster than the discriminative model—possibly with a number of training examples that is only *logarithmic*, rather than linear, in the number of parameters. This suggests—and our empirical results strongly support—that, as the number of training examples is increased, there can be two distinct regimes of performance, the first in which the generative model has already approached its asymptotic error and is thus doing better, and the second in which the discriminative model approaches its lower asymptotic error and does better.

2 Preliminaries

We consider a binary classification task, and begin with the case of discrete data. Let $\mathcal{X} = \{0,1\}^n$ be the n-dimensional input space, where we have assumed binary inputs for simplicity (the generalization offering no difficulties). Let the output labels be $\mathcal{Y} = \{T, F\}$, and let there be a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ from which a training set $S = \{x^{(i)}, y^{(i)}\}_{i=1}^m$ of m iid examples is drawn. The generative naive Bayes classifier uses S to calculate estimates $\hat{p}(x_i|y)$ and $\hat{p}(y)$ of the probabilities $p(x_i|y)$ and $p(y)$, as follows:

$$\hat{p}(x_i = 1|y = b) = \tfrac{\#_S\{x_i=1, y=b\}+l}{\#_S\{y=b\}+2l} \tag{1}$$

(and similarly for $\hat{p}(y = b)$,) where $\#_S\{\cdot\}$ counts the number of occurrences of an event in the training set S. Here, setting $l = 0$ corresponds to taking the empirical estimates of the probabilities, and l is more traditionally set to a positive value such as 1, which corresponds to using Laplace smoothing of the probabilities. To classify a test example x, the naive Bayes classifier $h_{\text{Gen}} : \mathcal{X} \mapsto \mathcal{Y}$ predicts $h_{\text{Gen}}(x) = T$ if and only if the following quantity is positive:

$$l_{\text{Gen}}(x) = \log \frac{(\prod_{i=1}^n \hat{p}(x_i|y = T))\hat{p}(y = T)}{(\prod_{i=1}^n \hat{p}(x_i|y = F))\hat{p}(y = F)} = \sum_{i=1}^n \log \frac{\hat{p}(x_i|y = T)}{\hat{p}(x_i|y = F)} + \log \frac{\hat{p}(y = T)}{\hat{p}(y = F)}. \tag{2}$$

In the case of continuous inputs, almost everything remains the same, except that we now assume $\mathcal{X} = [0,1]^n$, and let $\hat{p}(x_i|y = b)$ be parameterized as a univariate Gaussian distribution with parameters $\hat{\mu}_{i|y=b}$ and $\hat{\sigma}_i^2$ (note that the $\hat{\mu}$'s, but not the $\hat{\sigma}$'s, depend on y). The parameters are fit via maximum likelihood, so for example $\hat{\mu}_{i|y=b}$ is the empirical mean of the i-th coordinate of all the examples in the training set with label $y = b$. Note that this method is also equivalent to Normal Discriminant Analysis assuming diagonal covariance matrices. In the sequel, we also let $\mu_{i|y=b} = \mathrm{E}[x_i|y = b]$ and $\sigma_i^2 = \mathrm{E}_y[\mathrm{Var}(x_i|y)]$ be the "true" means and variances (regardless of whether the data are Gaussian or not).

In both the discrete and the continuous cases, it is well known that the discriminative analog of naive Bayes is logistic regression. This model has parameters $[\beta, \theta]$, and posits that $p(y = T|x; \beta, \theta) = 1/(1 + \exp(-\beta^T x - \theta))$. Given a test example x, the discriminative logistic regression classifier $h_{\mathrm{Dis}} : \mathcal{X} \mapsto \mathcal{Y}$ predicts $h_{\mathrm{Dis}}(x) = T$ if and only if the linear discriminant function

$$l_{\mathrm{Dis}}(x) = \sum_{i=1}^{n} \beta_i x_i + \theta \tag{3}$$

is positive. Being a discriminative model, the parameters $[\beta, \theta]$ can be fit either to maximize the conditional likelikood on the training set $\sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)}; \beta, \theta)$, or to minimize 0-1 training error $\sum_{i=1}^{n} 1\{h_{\mathrm{Dis}}(x^{(i)}) \neq y^{(i)}\}$, where $1\{\cdot\}$ is the indicator function ($1\{\text{True}\} = 1, 1\{\text{False}\} = 0$). Insofar as the error metric is 0-1 classification error, we view the latter alternative as being more truly in the "spirit" of discriminative learning, though the former is also frequently used as a computationally efficient approximation to the latter. In this paper, we will largely ignore the difference between these two versions of discriminative learning and, with some abuse of terminology, will loosely use the term "logistic regression" to refer to either, though our formal analyses will focus on the latter method.

Finally, let $\mathcal{H}$ be the family of all linear classifiers (maps from $\mathcal{X}$ to $\mathcal{Y}$); and given a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$, define its generalization error to be $\varepsilon(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$.

3 Analysis of algorithms

When $\mathcal{D}$ is such that the two classes are far from linearly separable, neither logistic regression nor naive Bayes can possibly do well, since both are linear classifiers. Thus, to obtain non-trivial results, it is most interesting to compare the performance of these algorithms to their asymptotic errors (cf. the agnostic learning setting). More precisely, let $h_{\mathrm{Gen},\infty}$ be the population version of the naive Bayes classifier; i.e. $h_{\mathrm{Gen},\infty}$ is the naive Bayes classifier with parameters $\hat{p}(x|y) = p(x|y), \hat{p}(y) = p(y)$. Similarly, let $h_{\mathrm{Dis},\infty}$ be the population version of logistic regression. The following two propositions are then completely straightforward.

Proposition 1 *Let h_{Gen} and h_{Dis} be any generative-discriminative pair of classifiers, and $h_{\mathrm{Gen},\infty}$ and $h_{\mathrm{Dis},\infty}$ be their asymptotic/population versions. Then*[1] $\varepsilon(h_{\mathrm{Dis},\infty}) \leq \varepsilon(h_{\mathrm{Gen},\infty})$.

Proposition 2 *Let h_{Dis} be logistic regression in n-dimensions. Then with high probability*

$$\varepsilon(h_{\mathrm{Dis}}) \leq \varepsilon(h_{\mathrm{Dis},\infty}) + O\left(\sqrt{\tfrac{n}{m} \log \tfrac{m}{n}}\right)$$

Thus, for $\varepsilon(h_{\mathrm{Dis}}) \leq \varepsilon(h_{\mathrm{Dis},\infty}) + \epsilon_0$ to hold with high probability (here, $\epsilon_0 > 0$ is some fixed constant), it suffices to pick $m = \Omega(n)$.

Proposition 1 states that aymptotically, the error of the discriminative logistic regression is smaller than that of the generative naive Bayes. This is easily shown by observing that, since $\varepsilon(h_{\mathrm{Dis}})$ converges to $\inf_{h\in\mathcal{H}} \varepsilon(h)$ (where $\mathcal{H}$ is the class of all linear classifiers), it must therefore be asymptotically no worse than the linear classifier picked by naive Bayes. This proposition also provides a basis for what seems to be the widely held belief that discriminative classifiers are better than generative ones.

Proposition 2 is another standard result, and is a straightforward application of Vapnik's uniform convergence bounds to logistic regression, and using the fact that $\mathcal{H}$ has VC dimension n. The second part of the proposition states that the sample complexity of discriminative learning—that is, the number of examples needed to approach the asymptotic error—is at most on the order of n. Note that the worst case sample complexity is also lower-bounded by order n [6].

[1] Under a technical assumption (that is true for most classifiers, including logistic regression) that the family of possible classifiers h_{Dis} (in the case of logistic regression, this is $\mathcal{H}$) has finite VC dimension.

The picture for discriminative learning is thus fairly well-understood: The error converges to that of the best linear classifier, and convergence occurs after on the order of n examples. How about generative learning, specifically the case of the naive Bayes classifier? We begin with the following lemma.

Lemma 3 *Let any $\epsilon_1, \delta > 0$ and any $l \geq 0$ be fixed. Assume that for some fixed $\rho_0 > 0$, we have that $\rho_0 \leq p(y = T) \leq 1 - \rho_0$. Let $m = O\left((1/\epsilon_1^2)\log(n/\delta)\right)$. Then with probability at least $1 - \delta$:*

1. *In case of discrete inputs, $|\hat{p}(x_i|y = b) - p(x_i|y = b)| \leq \epsilon_1$ and $|\hat{p}(y = b) - p(y = b)| \leq \epsilon_1$, for all $i = 1, \ldots, n$ and $b \in \mathcal{Y}$.*
2. *In the case of continuous inputs, $|\hat{\mu}_{i|y=b} - \mu_{i|y=b}| \leq \epsilon_1$, $|\hat{\sigma}_i^2 - \sigma_i^2| \leq \epsilon_1$, and $|\hat{p}(y = b) - p(y = b)| \leq \epsilon_1$ for all $i = 1, \ldots, n$ and $b \in \mathcal{Y}$.*

Proof (sketch). Consider the discrete case, and let $l = 0$ for now. Let $\epsilon_1 \leq \rho_0/2$. By the Chernoff bound, with probability at least $1 - \delta_1 = 1 - 2\exp(-2\epsilon_1^2 m)$, the fraction of positive examples will be within ϵ_1 of $p(y = T)$, which implies $|\hat{p}(y = b) - p(y = b)| \leq \epsilon_1$, and we have at least γm positive and γm negative examples, where $\gamma = \rho_0 - \epsilon_1 = \Omega(1)$. So by the Chernoff bound again, for specific i, b, the chance that $|\hat{p}(x_i|y = b) - p(x_i|y = b)| > \epsilon_1$ is at most $\delta_2 = 2\exp(-2\epsilon_1^2 \gamma m)$. Since there are $2n$ such probabilities, the overall chance of error, by the Union bound, is at most $\delta_1 + 2n\delta_2$. Substituting in δ_1 and δ_2's definitions, we see that to guarantee $\delta_1 + 2n\delta_2 \leq \delta$, it suffices that m is as stated. Lastly, smoothing ($l > 0$) adds at most a small, $O(1/m)$ perturbation to these probabilities, and using the same argument as above with (say) $\epsilon_1/2$ instead of ϵ_1, and arguing that this $O(1/m)$ perturbation is at most $\epsilon_1/2$ (which it is as m is at least order $1/\epsilon_1^2$), again gives the result. The result for the continuous case is proved similarly using a Chernoff-bounds based argument (and the assumption that $x_i \in [0, 1]$). $\square$

Thus, with a number of samples that is only *logarithmic*, rather than linear, in n, the parameters of the generative classifier h_{Gen} are uniformly close to their asymptotic values in $h_{\text{Gen},\infty}$. Is is tempting to conclude therefore that $\varepsilon(h_{\text{Gen}})$, the error of the generative naive Bayes classifier, also converges to its asymptotic value of $\varepsilon(h_{\text{Gen},\infty})$ after this many examples, implying only $O(\log n)$ examples are required to fit a naive Bayes model. We will shortly establish some simple conditions under which this intuition is indeed correct. Note that this implies that, even though naive Bayes converges to a higher asymptotic error of $\varepsilon(h_{\text{Gen},\infty})$ compared to logistic regression's $\varepsilon(h_{\text{Dis},\infty})$, it may also approach it significantly faster—after $O(\log n)$, rather than $O(n)$, training examples.

One way of showing $\varepsilon(h_{\text{Gen}})$ approaches $\varepsilon(h_{\text{Gen},\infty})$ is by showing that the parameters' convergence implies that h_{Gen} is very likely to make the same predictions as $h_{\text{Gen},\infty}$. Recall h_{Gen} makes its predictions by thresholding the discriminant function l_{Gen} defined in (2). Let $l_{\text{Gen},\infty}$ be the corresponding discriminant function used by $h_{\text{Gen},\infty}$. On every example on which both l_{Gen} and $l_{\text{Gen},\infty}$ fall on the same side of zero, h_{Gen} and $h_{\text{Gen},\infty}$ will make the same prediction. Moreover, as long as $l_{\text{Gen},\infty}(x)$ is, with fairly high probability, far from zero, then $l_{\text{Gen}}(x)$, being a small perturbation of $l_{\text{Gen},\infty}(x)$, will also be usually on the same side of zero as $l_{\text{Gen},\infty}(x)$.

Theorem 4 *Define $G(\tau) = \Pr_{(x,y)\sim\mathcal{D}}[(l_{\text{Gen},\infty}(x) \in [0, \tau n] \wedge y = T) \vee (l_{\text{Gen},\infty}(x) \in [-\tau n, 0] \wedge y = F)]$. Assume that for some fixed $\rho_0 > 0$, we have $\rho_0 \leq p(y = T) \leq 1 - \rho_0$, and that either $\rho_0 \leq p(x_i = 1|y = b) \leq 1 - \rho_0$ for all i, b (in the case of discrete inputs), or $\sigma_i^2 \geq \rho_0$ (in the continuous case). Then with high probability,*

$$\varepsilon(h_{\text{Gen}}) \leq \varepsilon(h_{\text{Gen},\infty}) + G\left(O\left(\sqrt{\tfrac{1}{m}\log n}\right)\right). \quad (4)$$

Proof (sketch). $\varepsilon(h_{\text{Gen}}) - \varepsilon(h_{\text{Gen},\infty})$ is upperbounded by the chance that $h_{\text{Gen},\infty}$ correctly classifies a randomly chosen example, but h_{Gen} misclassifies it.

Lemma 3 ensures that, with high probability, all the parameters of h_{Gen} are within $O(\sqrt{(\log n)/m})$ of those of $h_{\text{Gen},\infty}$. This in turn implies that every one of the $n+1$ terms in the sum in l_{Gen} (as in Equation 2) is within $O(\sqrt{(\log n)/m})$ of the corresponding term in $l_{\text{Gen},\infty}$, and hence that $|l_{\text{Gen}}(x) - l_{\text{Gen},\infty}(x)| \leq O(n\sqrt{(\log n)/m})$. Letting $\tau = O(\sqrt{(\log n)/m})$, we therefore see that it is possible for $h_{\text{Gen},\infty}$ to be correct and h_{Gen} to be wrong on an example (x, y) only if $y = T$ and $l_{\text{Gen},\infty}(x) \in [0, \tau n]$ (so that it is possible that $l_{\text{Gen},\infty}(x) \geq 0$, $l_{\text{Gen}}(x) \leq 0$), or if $y = F$ and $l_{\text{Gen},\infty}(x) \in [-\tau n, 0]$. The probability of this is exactly $G(\tau)$, which therefore upperbounds $\varepsilon(h_{\text{Gen}}) - \varepsilon(h_{\text{Gen},\infty})$. $\square$

The key quantity in the Theorem is the $G(\tau)$, which must be small when τ is small in order for the bound to be non-trivial. Note $G(\tau)$ is upper-bounded by $\Pr_x[l_{\text{Gen},\infty}(x) \in [-\tau n, \tau n]]$—the chance that $l_{\text{Gen},\infty}(x)$ (a random variable whose distribution is induced by $x \sim \mathcal{D}$) falls near zero. To gain intuition about the scaling of these random variables, consider the following:

Proposition 5 *Suppose that, for at least an $\Omega(1)$ fraction of the features i ($i = 1, \ldots, n$), it holds true that $|p(x_i = 1|y = T) - p(x_i = 1|y = F)| \geq \gamma$ for some fixed $\gamma > 0$ (or $|\mu_{i|y=T} - \mu_{i|y=F}| \geq \gamma$ in the case of continuous inputs). Then $\mathrm{E}[l_{\text{Gen},\infty}(x)|y = T] = \Omega(n)$, and $-\mathrm{E}[l_{\text{Gen},\infty}(x)|y = F] = \Omega(n)$.*

Thus, as long as the class label gives information about an $\Omega(1)$ fraction of the features (or less formally, as long as most of the features are "relevant" to the class label), the expected value of $|l_{\text{Gen},\infty}(x)|$ will be $\Omega(n)$. The proposition is easily proved by showing that, conditioned on (say) the event $y = T$, each of the terms in the summation in $l_{\text{Gen},\infty}(x)$ (as in Equation (2), but with $\hat{p}$'s replaced by p's) has non-negative expectation (by non-negativity of KL-divergence), and moreover an $\Omega(1)$ fraction of them have expectation bounded away from zero.

Proposition 5 guarantees that $|l_{\text{Gen},\infty}(x)|$ has large expectation, though what we want in order to bound G is actually slightly stronger, namely that the random variable $|l_{\text{Gen},\infty}(x)|$ further be large/far from zero with high probability. There are several ways of deriving sufficient conditions for ensuring that G is small. One way of obtaining a loose bound is via the Chebyshev inequality. For the rest of this discussion, let us for simplicity implicitly condition on the event that a test example x has label T. The Chebyshev inequality implies that $\Pr[l_{\text{Gen},\infty}(x) \leq \mathrm{E}[l_{\text{Gen},\infty}(x)] - t] \leq \mathrm{Var}(l_{\text{Gen},\infty}(x))/t^2$. Now, $l_{\text{Gen},\infty}(x)$ is the sum of n random variables (ignoring the term involving the priors $p(y)$). If (still conditioned on y), these n random variables are independent (i.e. if the "naive Bayes assumption," that the x_i's are conditionally independent given y, holds), then its variance is $O(n)$; even if the n random variables were not completely independent, the variance may still be not much larger than $O(n)$ (and may even be smaller, depending on the signs of the correlations), and is at most $O(n^2)$. So, if $\mathrm{E}[l_{\text{Gen},\infty}(x)|y = T] = \alpha n$ (as would be guaranteed by Proposition 5) for some $\alpha > 0$, by setting $t = (\alpha - \tau)n$, Chebyshev's inequality gives $\Pr[l_{\text{Gen},\infty}(x) \leq \tau n] \leq O(1/(\alpha - \tau)^2 n^\eta)$ ($\tau < \alpha$), where $\eta = 0$ in the worst case, and $\eta = 1$ in the independent case. This thus gives a bound for $G(\tau)$, but note that it will frequently be very loose. Indeed, in the unrealistic case in which the naive Bayes assumption really holds, we can obtain the much stronger (via the Chernoff bound) $G(\tau) \leq \exp(-O((\alpha - \tau)^2 n))$, which is exponentially small in n. In the continuous case, if $l_{\text{Gen},\infty}(x)$ has a density that, within some small interval $[-\epsilon n, \epsilon n]$, is uniformly bounded by $O(1/n)$, then we also have $G(\tau) = O(\tau)$. In any case, we also have the following Corollary to Theorem 4.

Corollary 6 *Let the conditions of Theorem 4 hold, and suppose that $G(\tau) \leq \epsilon_0/2 + F(\tau)$ for some function $F(\tau)$ (independent of n) that satisfies $F(\tau) \to 0$ as $\tau \to 0$, and some fixed $\epsilon_0 > 0$. Then for $\varepsilon(h_{\text{Gen}}) \leq \varepsilon(h_{\text{Gen},\infty}) + \epsilon_0$ to hold with high*

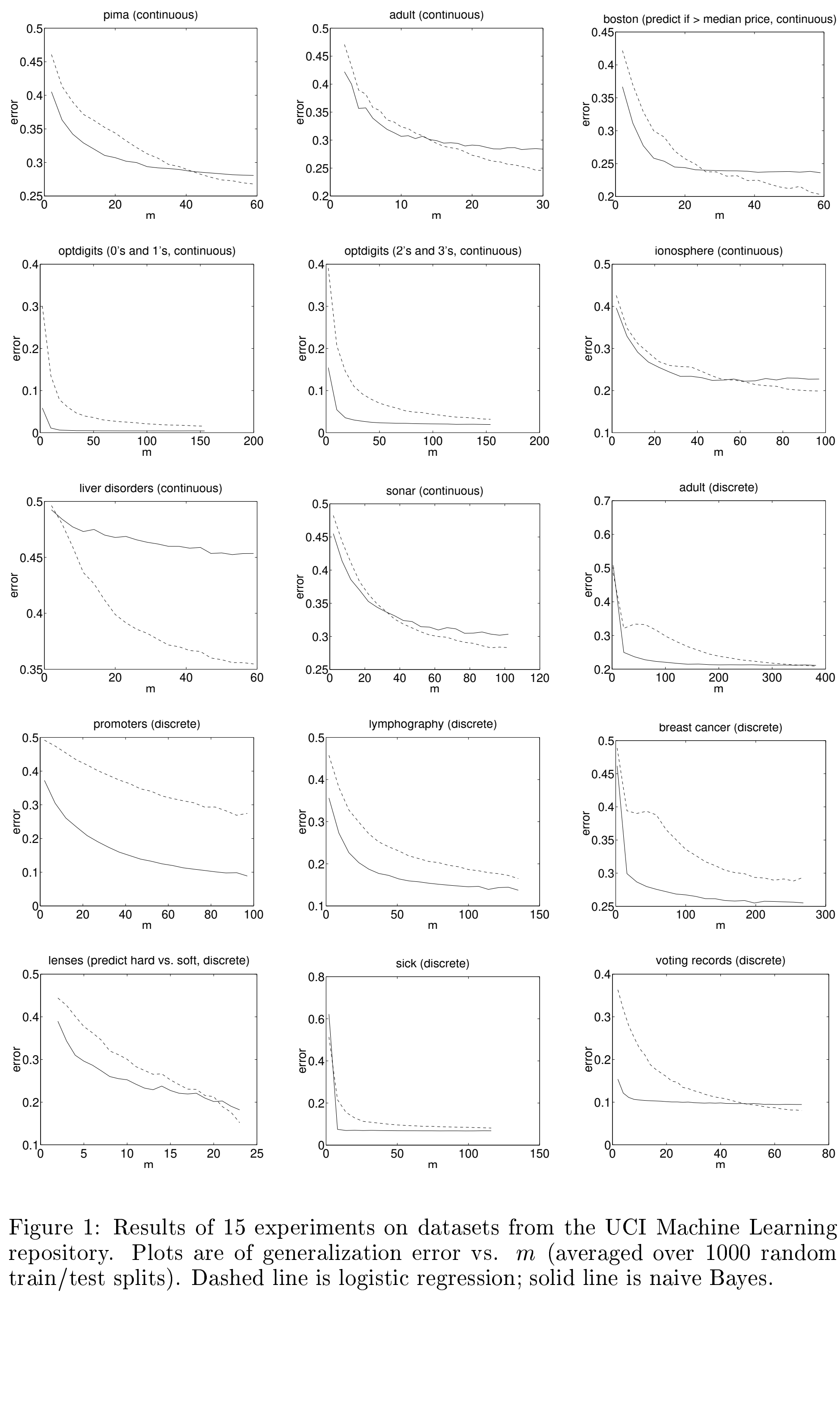

Figure 1: Results of 15 experiments on datasets from the UCI Machine Learning repository. Plots are of generalization error vs. m (averaged over 1000 random train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

probability, it suffices to pick $m = \Omega(\log n)$.

Note that the previous discussion implies that the preconditions of the Corollary do indeed hold in the case that the naive Bayes (and Proposition 5's) assumption holds, for any constant ϵ_0 so long as n is large enough that $\epsilon_0 \geq \exp(-O(\alpha^2 n))$ (and similarly for the bounded $\mathrm{Var}(l_{\mathrm{Gen},\infty}(x))$ case, with the more restrictive $\epsilon_0 \geq O(1/(\alpha^2 n^\eta))$). This also means that either of these (the latter also requiring $\eta > 0$) is a sufficient condition for the asymptotic sample complexity to be $O(\log n)$.

4 Experiments

The results of the previous section imply that even though the discriminative logistic regression algorithm has a lower asymptotic error, the generative naive Bayes classifier may also converge more quickly to its (higher) asymptotic error. Thus, as the number of training examples m is increased, one would expect generative naive Bayes to initially do better, but for discriminative logistic regression to eventually catch up to, and quite likely overtake, the performance of naive Bayes.

To test these predictions, we performed experiments on 15 datasets, 8 with continuous inputs, 7 with discrete inputs, from the UCI Machine Learning repository.[2]

The results of these experiments are shown in Figure 1. We find that the theoretical predictions are borne out surprisingly well. There are a few cases in which logistic regression's performance did not catch up to that of naive Bayes, but this is observed primarily in particularly small datasets in which m presumably cannot grow large enough for us to observe the expected dominance of logistic regression in the large m limit.

5 Discussion

Efron [2] also analyzed logistic regression and Normal Discriminant Analysis (for continuous inputs), and concluded that the former was only asymptotically very slightly (1/3–1/2 times) less statistically efficient. This is in marked contrast to our results, and one key difference is that, rather than assuming $P(x|y)$ is Gaussian with a diagonal covariance matrix (as we did), Efron considered the case where $P(x|y)$ is modeled as Gaussian with a full convariance matrix. In this setting, the estimated covariance matrix is singular if we have fewer than linear in n training examples, so it is no surprise that Normal Discriminant Analysis cannot learn much faster than logistic regression here. A second important difference is that Efron considered only the special case in which the $P(x|y)$ is truly Gaussian. Such an asymptotic comparison is not very useful in the general case, since the only possible conclusion, if $\varepsilon(h_{\mathrm{Dis},\infty}) < \varepsilon(h_{\mathrm{Gen},\infty})$, is that logistic regression is the superior algorithm. In contrast, as we saw previously, it is in the non-asymptotic case that the most interesting "two-regime" behavior is observed.

Practical classification algorithms generally involve some form of regularization—in particular logistic regression can often be improved upon in practice by techniques

[2] To maximize the consistency with the theoretical discussion, these experiments avoided discrete/continuous hybrids by considering only the discrete or only the continuous-valued inputs for a dataset where necessary. Train/test splits were random subject to there being at least one example of each class in the training set, and continuous-valued inputs were also rescaled to $[0, 1]$ if necessary. In the case of linearly separable datasets, logistic regression makes no distinction between the many possible separating planes. In this setting we used an MCMC sampler to pick a classifier randomly from them (i.e., so the errors reported are empirical averages over the separating hyperplanes). Our implementation of Normal Discriminant Analysis also used the (standard) trick of adding ϵ to the diagonal of the covariance matrix to ensure invertibility, and for naive Bayes we used $l = 1$.

such as shrinking the parameters via an L_1 constraint, imposing a margin constraint in the separable case, or various forms of averaging. Such regularization techniques can be viewed as changing the model family, however, and as such they are largely orthogonal to the analysis in this paper, which is based on examining particularly clear cases of Generative-Discriminative model pairings. By developing a clearer understanding of the conditions under which pure generative and discriminative approaches are most successful, we should be better able to design hybrid classifiers that enjoy the best properties of either across a wider range of conditions.

Finally, while our discussion has focused on naive Bayes and logistic regression, it is straightforward to extend the analyses to several other models, including generative-discriminative pairs generated by using a fixed-structure, bounded fan-in Bayesian network model for $P(x|y)$ (of which naive Bayes is a special case).

Acknowledgments

We thank Andrew McCallum for helpful conversations. A. Ng is supported by a Microsoft Research fellowship. This work was also supported by a grant from Intel Corporation, NSF grant IIS-9988642, and ONR MURI N00014-00-1-0637.

References

[1] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[2] B. Efron. The efficiency of logistic regression compared to Normal Discriminant Analysis. *Journ. of the Amer. Statist. Assoc.*, 70:892–898, 1975.

[3] P. Goldberg and M. Jerrum. Bounding the VC dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

[4] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[5] Y. D. Rubinstein and T. Hastie. Discriminative vs. informative learning. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 49–53. AAAI Press, 1997.

[6] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.