5 Risks of Semi-Supervised Learning: How Unlabeled Data Can Degrade Performance of Generative Classifiers

Empirical and theoretical results have often testified favorably towards the semi-supervised learning of generative classifiers, as described in other chapters of this book. However the literature has also brought to light a number of situations where semi-supervised learning fails to produce good generative classifiers. Here some clarification is due. We are not simply concerned with classifiers that produce high classification error — this can also happen in supervised learning. Our concern is this: it is frequently the case that we would be better off just discarding the unlabeled data and employing a supervised method, rather than taking a semi-supervised route. Thus we worry about the embarrassing situation where the addition of unlabeled data *degrades* the performance of a classifier.

performance degradation

How can this be? Typically we do not expect to be better off by discarding data; how can we understand this aspect of semi-supervised learning? In this chapter we focus on the effect of modeling errors in semi-supervised learning, and show how modeling errors can lead to performance degradation.

5.1 Do unlabeled data improve or degrade classification performance?

Perhaps it would be reasonable to expect an average improvement in classification performance for any increase in the number of samples (labeled or unlabeled): the more data, the better. In fact, existing literature presents empirical findings that attribute positive value to unlabeled data; other chapters present some of these results. O'Neill's statement that "unclassified observations should certainly not be discarded" [O'Neill, 1978] seems to be confirmed by theoretical studies, most notably by Castelli [1994], Castelli and Cover [1995, 1996] and Ratsaby and Venkatesh [1995].

The gist of these previous theoretical investigations is this. Suppose samples (x_i, y_i) are realizations of random variables X_v and Y_v that are distributed according to distribution $p(X_v, Y_v)$. Suppose one learns a parametric model $p(X_v, Y_v|\theta)$ such that $p(X_v, Y_v|\theta)$ is equal to $p(X_v, Y_v)$ for some value of θ — that is, the "model is

positive results: "correct" model

correct" in the sense that it can exactly represent $p(X_v, Y_v)$.[1] Then one is assured to have an expected reduction in classification error as more and more data are collected (labeled or unlabeled). Moreover, labeled data are exponentially more effective in reducing classification error than unlabeled data. In these optimistic results, unlabeled data can be profitably used whenever available.

examples of performance degradation

However, a more detailed analysis of current empirical results does reveal some puzzling aspects of unlabeled data. For example, Shahshahani and Landgrebe [1994] report experiments where unlabeled data degraded the performance of Naive Bayes classifiers with Gaussian variables. They attribute such cases to deviations from modeling assumptions, such as outliers and "samples of unknown classes" — they even suggest that unlabeled samples should be used with care, and only when the labeled data alone produce a poor classifier. Another representative example is the work by Nigam et al. [2000] on text classification, where classifiers sometimes display performance degradation. They suggest several possible sources of difficulties: numerical problems in the learning algorithm, mismatches between the natural clusters in feature space and the actual labels. Additional examples are easy to find. Baluja [1998] used Naive Bayes and Tree-Augmented Naive Bayes (TAN) classifiers [Friedman et al., 1997] to detect faces in images, but there were cases where unlabeled data degraded performance. Bruce [2001] used labeled and unlabeled data to learn Bayesian network classifiers, from Naive Bayes classifiers to fully connected networks; the Naive Bayes classifiers displayed bad classification performance, and in fact the performance degraded as more unlabeled data were used (more complex networks also displayed performance degradation as unlabeled samples were added). A final example: Bengio and Grandvalet [2004] describes experiments where outliers are added to a Gaussian model, causing generative classifiers to degrade with unlabeled data.

Figure 5.1 shows a number of experiments that corroborate this anecdotal evidence. All of them involve binary classification with categorical variables; in all of them X_v is actually a vector containing several attributes X_{vi}. In all experiments the generative classifiers were learned by maximum likelihood using the EM algorithm (Chapters 2, 4). Figure 5.1.a shows the performance of Naive Bayes classifiers learned with increasing amounts of unlabeled data (for several fixed amounts of labeled data), where the data are distributed according to Naive Bayes assumptions. That is, the data were generated by randomly generated statistical models that comply with the independence assumptions of Naive Bayes classifiers. In the Naive Bayes model, all attributes X_v are independent of each other given the class Y_v: $p(X_v, Y_v) = p(Y_v) \prod p(X_{vi})$. The result is simple: the more unlabeled data, the better. Figure 5.1.b shows an entirely different picture. Here a series of Naive Bayes classifiers were learned with data distributed according to TAN assumptions: each

1. Note that here and in the remainder of the paper we employ p to denote distributions and densities (for discrete/continuous variables using appropriate measures); we indicate the type of object we deal with whenever it is not clear from the context.

attribute is directly dependent on the class and on at most another attribute — the attributes form a "tree" of dependencies, hence the name Tree-Augmented Naive Bayes [Friedman et al., 1997]. That is, in Figure 5.1.b the "model is incorrect." The graphs in Figure 5.1.b indicate performance degradation with increasing amounts of unlabeled data.

Figure 5.1.c depicts a more complex scenario. Again a series of Naive Bayes classifiers were learned with data distributed according to TAN assumptions, so the "model is incorrect." Note that two of the graphs show a trend of decreasing error (as the number of unlabeled sample increases), while the other graph shows a trend of increasing error. Here unlabeled data improve performance in the presence of a few labeled samples, but unlabeled data degrade performance when added to a larger number of labeled samples. A larger set of experiments with artificial data is described by Cozman and Cohen [2002].

Figure 5.1.d shows the result of learning Naive Bayes classifiers using different combinations of labeled and unlabeled datasets for the Adult classification problem (using the training and testing datasets available in the UCI repository). We see that adding unlabeled data can improve classification when the labeled data set is small (30 labeled data), but degrade performance as the labeled data set becomes larger. Thus the properties of this real dataset lead to behavior similar to Figure 5.1.c.

Finally, Figure 5.1.e and Figure 5.1.f show the result of learning Naive Bayes and TAN classifiers using Data Set 8 in the benchmark data (Chapter 22). Both show similar trends as those displayed in previous graphs.

5.2 Understanding unlabeled data: Asymptotic bias

We can summarize the previous section as follows. First, there are results that guarantee benefits from unlabeled data when the learned generative classifier is based on a "correct" model. Second, there is strong empirical evidence that unlabeled data may degrade performance of classifiers. Performance degradation may occur whenever the modeling assumptions adopted for a particular classifier do not match the characteristics of the distribution generating the data.[2] This is troubling because it is usually difficult, if not impossible, to guarantee a priori that a particular statistical model is a "correct" one.

key: asymptotic bias

The key to understand the vagaries of semi-supervised learning is to study asymptotic bias. In this section we present an intuitive discussion, leaving more formal analysis to Section 5.3. Our arguments here and in the remainder of this chapter focus on generative classifiers learned by maximum likelihood methods. As

2. As we show in this and subsequent sections, performance degradation occurs even in the absence of numerical errors or existence of local optima for parameter estimation. In fact our presentation is independent of numerical techniques, so that results are not clouded by the intricacies of numerical analysis.

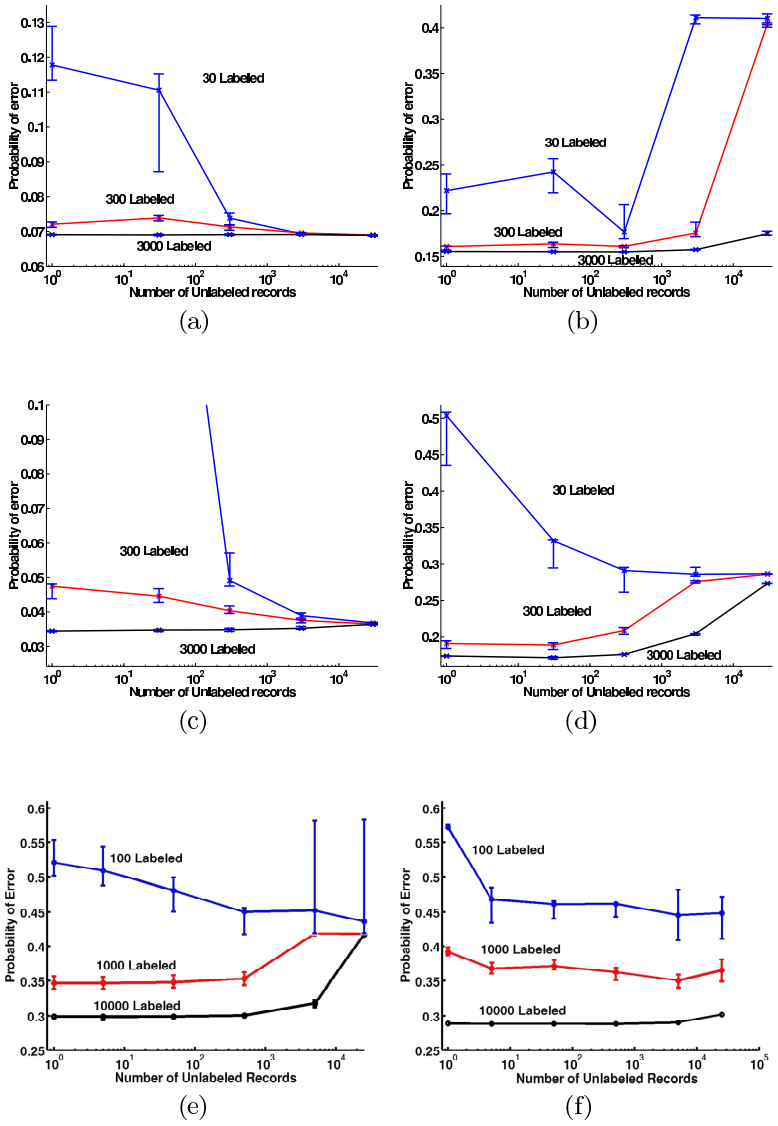

Figure 5.1 (a) Naive Bayes classifiers learned from data distributed according to Naive Bayes assumptions with 10 attributes; attributes with 2 to 4 values. (b) Naive Bayes classifiers learned from data distributed according to TAN assumptions with 10 attributes. (c) Naive Bayes classifiers learned from data distributed according to TAN assumptions with 49 attributes. (d) Naive Bayes classifiers generated from the Adult database. (e) Naive Bayes classifiers generated from the Data Set 8, benchmark data (Chapter 22). (f) TAN classifiers generated from the Data Set 8, benchmark data (Chapter 22). In all graphs, points summarize 10 runs of each classifier on testing data (bars cover 30th to 70th percentiles).

most of our arguments are asymptotic, the same rationale will apply to maximum a posteriori and other Bayesian estimators, as their asymptotic behavior is dominated by the likelihood function [DeGroot, 1970].

The gist of the argument is as follows. As we formally show in Section 5.3, the asymptotic bias of the maximum likelihood estimator produced with labeled data *can be different* from the asymptotic bias of the maximum likelihood estimator produced with unlabeled data, for the same classifier. Suppose then that one learns a classifier with a reasonable amount of labeled data. The resulting classifier may be relatively close to its asymptotic limit, yielding some classification error. Now suppose one takes a much larger amount of unlabeled data, and learns the same classifier with all available data. Now the classifier may be tending to the asymptotic limit *for unlabeled data* — and the performance for this limiting classifier may be worse than the performance for the first "labeled" limiting classifier. The net result is that by adding a large number of unlabeled samples one produces a worse classifier.

However puzzling, this situation can be found even in seemingly innocent situations, and does not require sophisticated modeling errors. We now discuss a simple example where unlabeled data degrades the performance of a generative classifier; this (fictitious) example may help the reader grasp the sometimes unexpected effects of unlabeled data.

classifying baby's gender

Consider the following classification problem. We are interested in predicting a baby's gender ($G = \text{Boy}$ or $G = \text{Girl}$) at the 20'th week of pregnancy based on two attributes: whether the mother craved chocolate in the first trimester ($Ch = \text{Yes}$ or $Ch = \text{No}$), and whether the mother's weight gain was more or less than 15lbs ($W = \text{More}$ or $W = \text{Less}$). Suppose that W and G are independent conditional on Ch; that is, the direct dependencies in the domain are expressed by the graph $G \rightarrow Ch \rightarrow W$, leading to the following decomposition of the joint distribution: $\text{P}(G, Ch, W) = \text{P}(G)\text{P}(Ch|G)\text{P}(W|Ch)$. Suppose also that data are distributed according to:

$$\begin{aligned} \text{P}(G = \text{Boy}) &= 0.5, \\ \text{P}(Ch = \text{No}|G = \text{Boy}) &= 0.1, \\ \text{P}(Ch = \text{No}|G = \text{Girl}) &= 0.8, \\ \text{P}(W = \text{Less}|Ch = \text{No}) &= 0.7, \\ \text{P}(W = \text{Less}|Ch = \text{Yes}) &= 0.2. \end{aligned}$$

Note that from the above distribution we can compute the probabilities of W given G to get:

$$\begin{aligned} \text{P}(W = \text{Less}|G = \text{Boy}) &= 0.25, \\ \text{P}(W = \text{Less}|G = \text{Girl}) &= 0.6. \end{aligned}$$

To classify the baby's gender given weight gain and chocolate craving, we compute the a posteriori probability of G given W and Ch (which, from the independence

stated above, depends only on Ch):

$$\begin{aligned} \mathrm{P}(G = \mathrm{Girl}|Ch = \mathrm{No}) &= 0.89, \\ \mathrm{P}(G = \mathrm{Boy}|Ch = \mathrm{No}) &= 0.11, \\ \mathrm{P}(G = \mathrm{Girl}|Ch = \mathrm{Yes}) &= 0.18, \\ \mathrm{P}(G = \mathrm{Boy}|Ch = \mathrm{Yes}) &= 0.82. \end{aligned}$$

Bayes rule

From the a posteriori probabilities, the optimal classification rule (the Bayes rule, discussed in the next section) is:

$$\text{if } Ch = \mathrm{No}, \text{ choose } G = \mathrm{Girl}; \qquad \text{if } Ch = \mathrm{Yes}, \text{ choose } G = \mathrm{Boy}. \tag{5.1}$$

The Bayes error rate (that is, the probability of error under the Bayes rule) for this problem can be easily computed and found to be at about 15%.

assuming Naive Bayes

Suppose that we incorrectly assume a Naive Bayes model for the problem; that is, we assume that dependencies are expressed by the graph $Ch \leftarrow G \rightarrow W$. Thus we incorrectly assume that weight gain is independent of chocolate craving given the gender, thus we incorrectly assume that the factorization of the joint probability distribution can be written as: $\mathrm{P}(G, Ch, W) = \mathrm{P}(G)\mathrm{P}(Ch|G)\mathrm{P}(W|G)$. Suppose that a friend gave us the "true" values of $\mathrm{P}(Ch|G)$, so we do not have to estimate these quantities. We wish to estimate $\mathrm{P}(G)$ and $\mathrm{P}(W|G)$ using maximum likelihood techniques.

In the case where only labeled data are available, estimators are obtained by relative frequencies, with zero bias and variance inversely proportional to the size of the database. Thus even a relatively small database will produce excellent estimates of probability values. The estimate for $\mathrm{P}(G)$ will most likely be close to 0.5; likewise, estimates of $\mathrm{P}(W = \mathrm{Less}|G = \mathrm{Girl})$ will be close to 0.6 and estimates of $\mathrm{P}(W = \mathrm{Less}|G = \mathrm{Boy})$ will be close to 0.25. With these estimated parameters and the assumed decomposition of the joint probability distribution, the a posteriori probabilities for G will likely be close to:

	$\mathrm{P}(G = \mathrm{Girl}\|Ch, W)$	$\mathrm{P}(G = \mathrm{Boy}\|Ch, W)$
$Ch = \mathrm{No}$, $W = \mathrm{Less}$	0.95	0.05,
$Ch = \mathrm{No}$, $W = \mathrm{More}$	0.81	0.19,
$Ch = \mathrm{Yes}$, $W = \mathrm{Less}$	0.35	0.65,
$Ch = \mathrm{Yes}$, $W = \mathrm{More}$	0.11	0.89.

the "labeled" classifier

Suppose we take these estimates and classify incoming observations using the maximum a posteriori value of G. Even though the bias from the "true" a-posteriori probabilities is not zero, this will *produce the same optimal Bayes rule (5.1)*; that is, the "labeled" classifier is likely to yield the minimum classification error.

Now suppose that unlabeled data are available. As more and more unlabeled samples are collected, the ratio between the number of labeled samples and the total number of samples goes to zero. In Section 5.3 we show how to compute the

asymptotic estimates in this case. The computation, which is performed in closed form for this case, yields the following asymptotic estimates: $\mathrm{P}(G = \mathrm{Boy}) = 0.5$, $\mathrm{P}(W = \mathrm{Less}|G = \mathrm{Girl}) = 0.78$, $\mathrm{P}(W = \mathrm{Less}|G = \mathrm{Boy}) = 0.07$. The a posteriori probabilities for G will therefore tend to:

	$\mathrm{P}(G = \mathrm{Girl}\|Ch, W)$	$\mathrm{P}(G = \mathrm{Boy}\|Ch, W)$
$Ch = \mathrm{No}$, $W = \mathrm{Less}$	0.99	0.01,
$Ch = \mathrm{No}$, $W = \mathrm{More}$	0.55	0.45,
$Ch = \mathrm{Yes}$, $W = \mathrm{Less}$	0.71	0.29,
$Ch = \mathrm{Yes}$, $W = \mathrm{More}$	0.05	0.95.

Classification using the maximum a posteriori value of G yields:

$$\begin{aligned}
&\text{if } \{Ch = \mathrm{No}, W = \mathrm{Less}\}, \text{ choose } G = \mathrm{Girl};\\
&\text{if } \{Ch = \mathrm{No}, W = \mathrm{More}\}, \text{ choose } G = \mathrm{Girl};\\
&\text{if } \{Ch = \mathrm{Yes}, W = \mathrm{Less}\}, \text{ choose } G = \mathrm{Girl};\\
&\text{if } \{Ch = \mathrm{Yes}, W = \mathrm{More}\}, \text{ choose } G = \mathrm{Boy}.
\end{aligned}$$

the "unlabeled" classifier

Here we see that the prediction has changed from the optimal in the case $\{Ch = \mathrm{Yes}, W = \mathrm{Less}\}$; instead of predicting $\{G = \mathrm{Boy}\}$, we predict $\{G = \mathrm{Girl}\}$. We can easily find the expected error rate to be at 22%, an *increase* of 7% in error.

What happened? The labeled data take us to a particular asymptotic limit, and the unlabeled data take us to a distinct limit. In Section 5.3 we will see that this transition is smooth as unlabeled samples are collected. Because the latter limit is worse (from the point of view of classification) than the former, the gradual addition of unlabeled degrades performance.

Consider again Figure 5.1.a. The graphs there illustrate the situation where the "model is correct": labeled and unlabeled data lead to identical asymptotic estimates. The other graphs in Figure 5.1 illustrate situations where the "model is incorrect". In these cases the asymptotic estimates tend to the "unlabeled" classifier as more and more unlabeled data are available — depending on the amount of labeled data, the graphs start above or below this "unlabeled" limit.

5.3 The asymptotic analysis of generative semi-supervised learning

We start by collecting a few assumptions in this section, at the cost of repeating definitions already stated in previous chapters. The goal here is to classify a vector of attributes X_v. Each instantiation x of X_v is a *sample*. There exists a *class variable* Y_v that takes values in a set of *labels*. To simplify the discussion, we assume that Y_v is a binary variable with values -1 and $+1$. We assume 0-1 loss, hence our objective is to minimize the probability of classification errors. If we knew exactly the joint distribution $p(X_v, Y_v)$, the optimal rule would be to select the label with highest posterior probability; this is the *Bayes rule*, and it produces the smallest

classification error, referred to as the *Bayes error* [Devroye et al., 1996]. A classifier is learned using n independent samples in a database; there are l labeled samples and u labeled samples ($n = l + u$), and without loss of generality we assume that the samples are ordered with labeled ones coming first. We assume that a sample has probability $(1 - \lambda)$ of having its label hidden (the same distribution $p(X_v|Y_v)$ generates the labeled and the unlabeled samples).

parametric model and assumptions

Consider that a generative model is adopted as a representation for the joint distribution $p(X_v, Y_v)$. Suppose that a parametric representation $p(X_v, Y_v|\theta)$ with parameters θ is employed, and a database containing *training* samples is available to produce estimates $\hat{\theta}$. All samples x_i are collected in a database denoted by X, and all samples y_i are collected in a database denoted by Y. We consider "plug-in" classification: compute the optimal rule pretending that $p(Y_v|X_v, \hat{\theta})$ is the correct posterior density of Y_v.

Throughout the chapter we denote the distributions/densities generating the data by $p(\cdot)$ and the statistical models that are employed to learn the distribution by $p(\cdot|\theta)$. Several smoothness and measurability assumptions on these distributions/densities are necessary to proceed with asymptotic analysis and are adopted throghout.[3]

Two principles often used to generate estimates are *maximum likelihood* and *maximization of posterior loss* [DeGroot, 1970]; the computation of estimates using these principles generally requires iterative methods, the most popular of which is the EM algorithm [Dempster et al., 1977]. Generative models are well suited for semi-supervised learning by maximum likelihood, because the likelihood is directly affected by unlabeled data — as opposed to discriminative models, where the associated likelihood is not affected by unlabeled data [Zhang and Oles, 2000].

likelihood

We take that estimates $\hat{\theta}$ are produced by maximizing the likelihood $L(\theta) = \prod_{i=1}^{l} p(x_i, y_i|\theta) \prod_{j=l+1}^{n} p(x_j|\theta)$. When a sample is unlabeled, its likelihood can be written as a mixture $p(X_v|Y_v = +1, \theta)p(Y_v = +1|\theta) + p(X_v|Y_v = -1, \theta)p(Y_v = -1|\theta)$; we assume that such mixtures are identifiable [Redner and Walker, 1984].

We use the following known result [Berk, 1966, Huber, 1967, White, 1982]. Consider a parametric model $p(Z|\theta)$ and a sequence of maximum likelihood estimates $\hat{\theta}_n$, obtained by maximization of $\sum_{i=1}^{n} \log p(z_i|\theta)$, with an increasing number n of independent samples z_i, all identically distributed according to $p(Z)$. Then $\hat{\theta}_n \to \theta^*$ as $n \to \infty$ for θ in an open neighborhood of θ^*, where θ^* maximizes $\mathbf{E}_{p(Z)}[\log p(Z|\theta)]$. If θ^* is interior to the parameter space, then estimates are asymptotically Gaussian.

central result

Extending the result above to semi-supervised learning we have:

---

3. Distributions must be defined on measurable Euclidean spaces, with measurable Radon-Nikodym densities. The dependence of $p(X_v, Y_v|\theta)$ on θ must be continuous so that second derivatives exist (and first derivatives must be measurable). Likelihoods, their derivatives and second derivatives must be dominated by integrable functions. Finally, expected values $\mathbf{E}_{p(Z)}[\log p(Z|\theta)]$ must exist for Z equal to X_v, Y_v and (X_v, Y_v). These conditions are listed in detail by Cozman et al. [2003a].

Theorem 5.1 *The limiting value θ^* of maximum likelihood estimates is:*

$$\arg\max_{\theta} \left(\lambda \mathbf{E}_{p(X_v,Y_v)} \left[\log p(X_v, Y_v|\theta)\right] + (1-\lambda)\mathbf{E}_{p(X_v,Y_v)} \left[\log p(X_v|\theta)\right]\right). \quad (5.2)$$

Proof. In semi-supervised learning, the samples are realizations of (X_v, Y_v) with probability λ, and of X_v with probability $(1-\lambda)$. Denote by $\tilde{Y}_v$ a random variable that assumes the same values of Y_v plus the "unlabeled" value 0. We have $p(\tilde{Y}_v \neq 0) = \lambda$. The actually observed samples are realizations of $(X_v, \tilde{Y}_v)$, thus

$$\tilde{p}(X_v, \tilde{Y}_v = y) = (\lambda p(X_v, Y_v = y))^{I_{\{\tilde{Y}_v \neq 0\}}(y)} \left((1-\lambda)p(X_v)\right)^{I_{\{\tilde{Y}_v = 0\}}(y)},$$

where $p(X_v)$ is a mixture density. Accordingly, the parametric model adopted for $(X_v, \tilde{Y}_v)$ has the same form:

$$\tilde{p}(X_v, \tilde{Y}_v = y|\theta) = (\lambda p(X_v, Y_v = y|\theta))^{I_{\{\tilde{Y}_v \neq 0\}}(y)} \left((1-\lambda)p(X_v|\theta)\right)^{I_{\{\tilde{Y}_v = 0\}}(y)}.$$

The value θ^* that maximizes $\mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[\log \tilde{p}(X_v, \tilde{Y}_v|\theta)\right]$ is

$$\arg\max_{\theta} \mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[I_{\{\tilde{Y}_v \neq 0\}}(\tilde{Y}_v) \left(\log \lambda p(X_v, Y_v|\theta)\right) + I_{\{\tilde{Y}_v = 0\}}(\tilde{Y}_v) \left(\log(1-\lambda)p(X_v|\theta)\right)\right].$$

Hence θ^* maximizes

$$\beta + \mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[I_{\{\tilde{Y}_v \neq 0\}}(\tilde{Y}_v) \log p(X_v, Y_v|\theta)\right] + \mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[I_{\{\tilde{Y}_v = 0\}}(\tilde{Y}_v) \log p(X_v|\theta)\right],$$

where $\beta = \lambda \log \lambda + (1-\lambda)\log(1-\lambda)$. As β does not depend on θ, we must only maximize the last two terms, which are equal to $\lambda \mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[\log p(X_v, Y_v|\theta)|\tilde{Y}_v \neq 0\right] + (1-\lambda)\mathbf{E}_{\tilde{p}(X_v,\tilde{Y}_v)} \left[\log p(X_v|\theta)|\tilde{Y}_v = 0\right]$. As we have $\tilde{p}(X_v, \tilde{Y}_v|\tilde{Y}_v \neq 0) = p(X_v, Y_v)$ and $\tilde{p}(X_v|\tilde{Y}_v = 0) = p(X_v)$, the last expression is equal to $\lambda \mathbf{E}_{p(X_v,Y_v)} \left[\log p(X_v, Y_v|\theta)\right] + (1-\lambda)\mathbf{E}_{p(X_v,Y_v)} \left[\log p(X_v|\theta)\right]$. Thus we obtain Expression (5.2). ■

Results by White [1982] can also be adapted to the context of semi-supervised learning to prove that generally the variance of estimates decreases with increasing n. The asymptotic variance depends on the inverse of the Fisher information; the Fisher information is typically larger for larger proportions of labeled data [Castelli, 1994], [Castelli and Cover, 1995, 1996].

semi-supervised learning as "convex" combination

Expression (5.2) indicates that the objective function in semi-supervised learning can be viewed asymptotically as a "convex" combination of objective functions for supervised learning ($\mathbf{E}\left[\log p(X_v, Y_v|\theta)\right]$) and for unsupervised learning ($\mathbf{E}\left[\log p(X_v|\theta)\right]$). Denote by θ^*_λ the value of θ that maximizes Expression (5.2) for a given λ. Denote by θ^*_l the "labeled" limit θ^*_1 and by θ^*_u the "unlabeled" limit θ^*_0.[4] We note that, with a few additional assumptions on the modeling densities, Theorem 5.1 and the implicit function theorem can be used to prove that θ^*_λ is

4. We have to handle a difficulty with the classification error for θ^*_u: given only unlabeled data, there is no information to decide the labels for decision regions, and the classification error is 1/2 [Castelli, 1994]. Thus we always reason with $\lambda \to 0$ instead of $\lambda = 0$.

a continuous function of λ — that is, the "path" followed by the solution is a continuous one.

model is correct

We can now present more formal versions of the arguments sketched in Section 5.2. Suppose first that the family of distributions $p(X_v, Y_v|\theta)$ contains the distribution $p(X_v, Y_v)$; that is, $p(X_v, Y_v|\theta_\top) = p(X_v, Y_v)$ for some $\theta_\top$, so the "model is correct." When such a condition is satisfied, $\theta_l^* = \theta_u^* = \theta_\top$ given identifiability, and then $\theta_\lambda^* = \theta_\top$, for any $0 < \lambda \leq 1$, is a maximum likelihood estimate. In this case, maximum likelihood is consistent, the asymptotic bias is zero, and classification error converges to the Bayes error. As variance decreases with increasing numbers of labeled and unlabeled data, the addition of both kinds of data eventually reaches the "correct" distribution and the Bayes error.

model is incorrect

We now study the scenario that is more relevant to our purposes, where the distribution $p(X_v, Y_v)$ does not belong to the family of distributions $p(X_v, Y_v|\theta)$. Denote by $e(\theta)$ the classification error with parameter θ, and suppose $e(\theta_u^*) > e(\theta_l^*)$ (as in the Boy-Girl example and in the other examples presented later). If we observe a large number of labeled samples, the classification error is approximately $e(\theta_l^*)$. If we then collect more samples, most of which unlabeled, we eventually reach a point where the classification error approaches $e(\theta_u^*)$. So, the net result is that we started with classification error close to $e(\theta_l^*)$, and by adding a great number of unlabeled samples, classification performance degraded towards $e(\theta_u^*)$. A labeled dataset can be dwarfed by a much larger unlabeled dataset: the classification error using the whole dataset can be larger than the classification error using only labeled data.

summary

To summarize, we have the following conclusions. First, labeled and unlabeled data contribute to a reduction in variance in semi-supervised learning under maximum likelihood estimation. Second, when the model is "correct," maximum likelihood methods are asymptotically unbiased both with labeled and unlabeled data. Third, when the model is "incorrect," there may be different asymptotic biases for different values of λ. Asymptotic classification error may also vary with λ — an increase in the number of unlabeled samples may lead to a larger estimation asymptotic bias and to a larger classification error. If the performance obtained with a given set of labeled data is better than the performance with infinitely many unlabeled samples, then at some point the addition of unlabeled data must decrease performance.

5.4 The value of labeled and unlabeled data

The previous discussion alluded to the possibility that $e(\theta_u^*) > e(\theta_l^*)$ when the model is "incorrect." To understand a few important details about this phenomenon, consider another example.

Gaussian example

Suppose we have attributes X_{v1} and X_{v2} from two classes -1 and $+1$. We know that (X_{v1}, X_{v2}) is a Gaussian vector with mean $(0, 3/2)$ conditional on $\{Y_v = -1\}$, and mean $(3/2, 0)$ conditional on $\{Y_v = +1\}$; variances for X_{v1} and for X_{v2} conditional on Y_v are equal to 1. We believe that X_{v1} and X_{v2} are independent

given Y_v, but actually X_{v1} and X_{v2} are *dependent* conditional on $\{Y_v = +1\}$: the correlation $\rho = \mathbf{E}\left[(X_{v1} - \mathbf{E}\left[X_{v1}|Y_v = +1\right])(X_{v2} - \mathbf{E}\left[X_{v2}|Y_v = +1\right])|Y_v = +1\right]$ is equal to $4/5$ (X_{v1} and X_{v2} are independent conditional on $\{Y_v = -1\}$). Data are sampled from a distribution such that $\eta = \mathrm{P}(Y_v = -1) = 3/5$, but we do not know this probability. If we knew the value of ρ and η, we would easily compute the optimal classification boundary on the plane $X_{v1} \times X_{v2}$ (this optimal classification boundary is quadratic). By mistakenly assuming that ρ is zero we are generating a Naive-Bayes classifier that approximates $\mathrm{P}(Y_v|X_{v1}, X_{v2})$.

Under the incorrect assumption that $\rho = 0$, the "optimal" classification boundary is linear: $x_{v2} = x_{v1} + 2\log((1-\hat{\eta})/\hat{\eta})/3$. With labeled data we can easily obtain $\hat{\eta}$ (a sequence of Bernoulli trials); then $\eta_l^* = 3/5$ and the classification boundary is given by $x_{v2} = x_{v1} - 0.27031$. Note that this (linear) boundary obtained with labeled data and the generative Naive Bayes classifier assumption is not the best possible linear boundary minimizing the classification error. We can in fact find the best possible linear boundary of the form $x_{v2} = x_{v1} + \gamma$. The classification error can be written as a function of γ that has positive second derivative; consequently the function has a single minimum that can be found numerically (the minimizing γ is -0.45786). If we consider the set of lines of the form $x_{v2} = x_{v1} + \gamma$, we see that the farther we go from the best line, the larger the classification error. Figure 5.2 shows the linear boundary obtained with labeled data and the best possible linear boundary. The boundary from labeled data is "above" the best linear boundary.

the "labeled" classifier

the best linear classifier

Now consider the computation of η_u^*, the asymptotic estimate with unlabeled data. By Theorem 5.1, we must obtain:

$$\arg\max_{\eta\in[0,1]} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_0(x_{v1}, x_{v2}) \log(\eta g_1(x_{v1}, x_{v2}) + (1-\eta) g_3(x_{v1}, x_{v2}))\, dx_{v2} dx_{v1},$$

where

$$\begin{aligned}
g_0(x_{v1}, x_{v2}) &= (3/5) g_1(x_{v1}, x_{v2}) + (2/5) g_2(x_{v1}, x_{v2}),\\
g_1(x_{v1}, x_{v2}) &= \mathcal{N}([0, 3/2]^T, \mathrm{diag}[1, 1]),\\
g_2(x_{v1}, x_{v2}) &= \mathcal{N}\left([3/2, 0]^T, \begin{bmatrix} 1 & 4/5 \\ 4/5 & 1 \end{bmatrix}\right),\\
g_3(x_{v1}, x_{v2}) &= \mathcal{N}([3/2, 0]^T, \mathrm{diag}[1, 1]).
\end{aligned}$$

The second derivative of this double integral is always negative (as can be seen by interchanging differentiation with integration), so the function is concave and there is a single maximum. We can search for the zero of the derivative of the double integral with respect to η. We obtain this value numerically, $\eta_u^* = 0.54495$. Using this estimate, the linear boundary from unlabeled data is $x_{v2} = x_{v1} - 0.12019$. This line is "above" the linear boundary from labeled data, and, given the previous discussion, leads to a larger classification error than the boundary from labeled data. The boundary obtained from unlabeled data is also shown in Figure 5.2. The classification error for the best linear boundary is 0.06975, while $e(\eta_l^*) = 0.07356$ and $e(\eta_u^*) = 0.08141$.

the "unlabeled" classifier

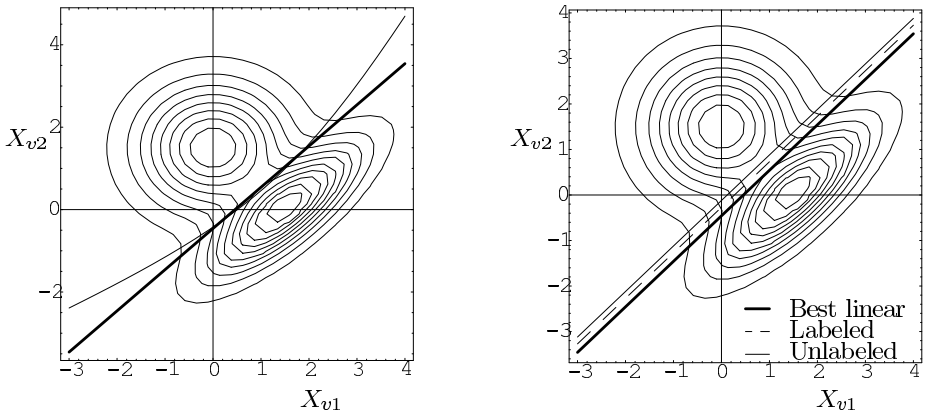

Figure 5.2 Graphs for the Gaussian example. On the left, contour plots of the mixture $p(X_{v1}, X_{v2})$, the optimal classification boundary (quadratic curve) and the best possible classification boundary of the form $x_{v2} = x_{v1} + \gamma$. On the right, the same contour plots, and the best linear boundary (lower line), the linear boundary obtained from labeled data (middle line) and the linear boundary obtained from unlabeled data (upper line).

This example suggests the following situation. Suppose we collect a large number l of labeled samples from $\mathrm{P}(Y_v, X_{v1}, X_{v2})$, with $\eta = 3/5$ and $\rho = 4/5$. The labeled estimates form a sequence of Bernoulli trials with probability 3/5, so the estimates quickly approach η_l^* (the variance of $\hat{\eta}$ decreases as $6/(25l)$). If we then add a very large amount of unlabeled data to our data, $\hat{\eta}$ approaches η_u^* and the classification error increases.

changing η and ρ

By changing the values of η and ρ, we can produce other interesting situations. For example, if $\eta = 3/5$ and $\rho = -4/5$, the best linear boundary is $x_{v2} = x_{v1} - 0.37199$, the boundary from labeled data is $x_{v2} = x_{v1} - 0.27031$, and the boundary from unlabeled data is $x_{v2} = x_{v1} - 0.34532$; the latter boundary is "between" the other two — additional unlabeled data lead to improvement in classification performance! As another example, if $\eta = 3/5$ and $\rho = -1/5$, the best linear boundary is $x_{v2} = x_{v1} - 0.29044$, the boundary from labeled data is $x_{v2} = x_{v1} - 0.27031$, and the boundary from unlabeled data is $x_{v2} = x_{v1} - 0.29371$. The best linear boundary is "between" the other two. In this case we attain the best possible linear boundary by mixing labeled and unlabeled data with $\lambda = 0.08075$.

We have so far found that taking larger and larger amounts of unlabeled data changes not only the variance of estimates but also their average behavior. The Gaussian example shows that we cannot always expect labeled data to produce a better classifier than the unlabeled data. Still, one would intuitively expect labeled data to provide more guidance to a learning procedure than unlabeled data. Is there anything that can be said about the (intuitively plausible and empirically visible) more valuable status of labeled data?

"labeled" limit better than the "unlabeled" one?

One informal argument is this. Suppose we have an estimate $\hat{\theta}$. It is typically the case that the smaller the value of the expected Kullback- Leibler divergence between

$p(Y_v|X_v)$ and $p(Y_v|X_v,\hat{\theta})$, the smaller the classification error, where the Kullback-Leibler divergence is $EKL(\theta) = \mathbf{E}[\log(p(Y_v|X_v)/p(Y_v|X_v,\theta)]$ [Garg and Roth, 2001, Cover and Thomas, 1991]. Direct minimization of expected Kullback-Leibler divergence yields $EKL(\theta_t^*)$ where $\theta_t^* = \arg\max_\theta \mathbf{E}[\log p(Y_v|X_v,\theta)]$. Now unlabeled data asymptotically yields $EKL(\theta_u^*)$ where $\theta_u^* = \arg\max_\theta \mathbf{E}[\log p(X_v|\theta)]$, and labeled data asymptotically yields $EKL(\theta_l^*)$ where $\theta_l^* = \arg\max_\theta \mathbf{E}[\log p(Y_v|X_v,\theta)]+ \mathbf{E}[\log p(X_v|\theta)]$. Note the following pattern. We are interested in minimizing $\mathbf{E}[\log p(Y_v|X_v,\theta)]$. While labeled data allows us to minimize a combination of this quantity plus $\mathbf{E}[\log p(X_v|\theta)]$, unlabeled data only allows us to minimize $\mathbf{E}[\log p(X_v|\theta)]$. When the "model is incorrect," this last quantity may in fact be far from the "true" $\mathbf{E}[\log p(X_v)]$, and we may be getting less help from unlabeled data than we might get from labeled data. This informal argument seems to be at the core of the perception that labeled data should be more valuable than unlabeled data when "model is incorrect." The analysis presented in this chapter adds to this perception the following comment: by trying to (asymptotically) minimize an expected value $\mathbf{E}[\log p(X_v)|\theta]$ that may even be unrelated to the "true" $\mathbf{E}[\log p(X_v)]$, we may in fact be *led astray* by the unlabeled data.

5.5 Finite sample effects

Asymptotic analysis can provide insight into complex phenomena, but finite sample effects are also important. In practice one may have very little labeled data, and the estimates $\hat{\theta}$ from labeled data may be so poor that the addition of unlabeled data is a positive move. This can be explained as follows. A small number of labeled samples may lead to estimators with high variance, thus likely to yield high classification error [Friedman, 1997]. In those circumstances the inclusion of unlabeled data may lead to a substantial decrease in variance and a decrease in classification error, even as the bias is negatively affected by the unlabeled data.

In general, the more parameters one has to estimate, the larger the variance of estimators for the same amount of data. If we have a classifier with a large number of attributes and we have only a few labeled samples, the variance of estimators is likely to be large, and classification performance is likely to be poor — the addition of unlabeled data is then a reasonable action to take. Consider again Figure 5.1.c. Here we have a Naive Bayes classifier with 49 attributes. If we have a relatively large amount of labeled data, we start close to the "labeled" limit $e(\theta_l^*)$, and then we observe performance degradation as we move towards $e(\theta_u^*)$. However, if we have few labeled samples, we start with very poor performance, and we decrease classification error by moving towards $e(\theta_u^*)$.

many attributes

We note that text classification is an important problem where many attributes are often available (often thousands of attributes), and where generative semi-supervised learning has been successful [Nigam et al., 2000].

text classification

5.6 Model search and robustness

looking for correct models

In semi-supervised learning we must always consider the possibility that a more accurate statistical model will lead to significant gains from unlabeled data. That is, we should look for the "correct" model whenever possible. In fact, the literature has described situations where a fixed-structure classifier, like the Naive Bayes, performs poorly, while model search schemes can lead to excellent classifiers [Bruce, 2001, Cohen et al., 2003, 2004]. In particular, Cohen et al. [2004] discuss and compare different model search strategies with labeled and unlabeled data for Bayesian network classifiers. Results show that TAN classifiers, learned with the EM algorithm [Meila, 1999], can sometimes improve classification and eliminate performance degradation with unlabeled data compared to the simpler Naive Bayes. In contrast, structure learning algorithms that maximize the likelihood of class and attributes, such as those proposed by Friedman [1998] and van Allen and Greiner [2000], are not likely to find structures yielding good classifiers in a semi-supervised manner, because of their focus on fitting the joint distribution rather than the a posteriori distribution (as also argued by Friedman et al. [1997] for the purely supervised case). The class of independence-based methods for structure learning, also known as constraint-based or test-based methods, is another alternative for attempting to learn the correct model. However, these methods do not easily adapt to use unlabeled data. Such a modification of algorithms by Cheng et al. [1997] is presented in Cohen et al. [2004], showing either none or marginal improvement compared to the EM version of TAN, while requiring much greater computational complexity. A third alternative is to perform structure search, attempting to maximize classification accuracy directly. Cohen et al. [2004] proposed to use a stochastic structure search algorithm (Markov chain Monte Carlo), accepting or rejecting models based on their classification accuracy (estimated using the labeled training data), while learning the parameters of each model using maximum likelihood estimation with both labeled and unlabeled data. This strategy yielded very good results for datasets with moderate number of labeled samples (and much larger number of unlabeled samples), but did not work well for datsets with very small number of labeled samples, because of its dependence on estimation of the classification error during the search.

detecting incorrect models

Given the results in this chapter, unlabeled data can also be useful in testing modeling assumptions. If the addition of unlabeled data to an existing pool of labeled data degrades performance, then there is clear indication that modeling assumptions are incorrect. In fact one can test whether differences in performance are statistically significant, using results by O'Neill [1978]; once one finds that a particular set of modeling assumptions is flawed, a healthy process of model revision may be started. In fact, one might argue that model search/revision should always be an important component in the toolset of semi-supervised learning [Cozman et al., 2003b].

5.7 Conclusion

Given the possibility of performance degradation, it seems that some care must be taken in generative semi-supervised learning. Statements that are intuitively and provably true when models are "correct" may fail (sometimes miserably!) when models are "incorrect." Apparently mild modeling errors may cause unlabeled data to degrade performance, even in the absence of numerical errors, and even in situations where more labeled data would be beneficial. Examples of performance degradation from outliers and other common modeling errors can be easily concocted [Cozman et al., 2003a].

In the absence of modeling errors, labeled data differ from unlabeled data only on the "information they carry about the decisions associated with the decision regions" [Castelli and Cover, 1995]. However as we consider the possibility of modeling errors, labeled data and unlabeled data also differ in the biases they induce on estimates. The analysis in Sections 5.2, 5.3, and 5.4 focused on asymptotic bias, a strategy that avoids distractions from finite sample effects and numerical errors. However we note that finite sample effects may be important in practice, as we discuss in Section 5.5.

methodology

At this point is is perhaps useful to add a few comments of methodological character. Given a pool of labeled and unlabeled data, generative semi-supervised learning is an attractive strategy. However one should always start by learning a *supervised* classifier with the labeled data. This "baseline" classifier can then be compared to other semi-supervised classifiers through cross-validation or similar techniques. Whenever modeling assumptions seem inaccurate, unlabeled data can be used to test modeling assumptions. If time and resources are available, model search should be conducted, attempting to reach a "correct" model — that is, a model where unlabeled data will be truly beneficial. Techniques discussed in Section 5.6 can be employed in this setting. An additional step is to compare the baseline classifier to non-generative methods. There are many semi-supervised non-generative classifiers, as discussed in other chapters of this book. There is also a significant number of methods that use labeled and unlabeled data for different purposes — for example, methods where the unlabeled data are used only to conduct dimensionality reduction (Chapter 12). However we should warn that a few empirical results in the literature suggest the possibility of performance degradation in non-generative semi-supervised learning paradigms, such as transductive SVM [Zhang and Oles, 2000] and co-training [Ghani, 2002].

active learning

A final methodological comment concerns *active* learning — that is, the option of labeling selected samples among the unlabeled data. This option should be seriously considered whenever possible. It may be that the most profitable use of unlabeled data in a particular problem is exactly as a pool of samples from which some samples can be carefully selected and labeled. In general, we should take the value of a labeled sample to be considerably higher than the value of an unlabeled sample.