9 Entropy Regularization

Yves Grandvalet
Yoshua Bengio

The problem of semi-supervised induction consists in learning a decision rule from labeled and unlabeled data. This task can be undertaken by discriminative methods, provided that learning criteria are adapted consequently. In this chapter, we motivate the use of entropy regularization as a means to benefit from unlabeled data in the framework of maximum a posteriori estimation. The learning criterion is derived from clearly stated assumptions and can be applied to any smoothly parametrized model of posterior probabilities. The regularization scheme favors low density separation, without any modeling of the density of input features. The contribution of unlabeled data to the learning criterion induces local optima, but this problem can be alleviated by deterministic annealing. For well-behaved models of posterior probabilities, deterministic annealing EM provides a decomposition of the learning problem in a series of concave subproblems. Other approaches to the semi-supervised problem are shown to be close relatives or limiting cases of entropy regularization. A series of experiments illustrates the good behavior of the algorithm in terms of performance and robustness with respect to the violation of the postulated low density separation assumption. The minimum entropy solution benefits from unlabeled data and is able to challenge mixture models and manifold learning in a number of situations.

9.1 Introduction

semi-supervised induction

This chapter addresses semi-supervised induction, which refers to the learning of a decision rule, on the entire input domain $\mathcal{X}$, from labeled and unlabeled data. The objective is identical to the one of supervised classification: generalize from examples. The problem differs in the respect that the supervisor's responses are missing for some training examples. This characteristic is shared with transduction, which has however a different goal, that is, of predicting labels on a set of predefined

patterns.

In the probabilistic framework, semi-supervised induction is a missing data problem, which can be addressed by generative methods such as mixture models thanks to the EM algorithm and extensions thereof [McLachlan, 1992]. Generative models apply to the joint density of patterns x and class y. They have appealing features, but they also have major drawbacks. First, the modeling effort is much more demanding than for discriminative methods, since the model of $p(x, y)$ is necessarily more complex than the model of $P(y|x)$. Being more precise, the generative model is also more likely to be misspecified. Second, the fitness measure is not discriminative, so that better models are not necessarily better predictors of class labels (this issue is addressed in chapter 2).

These difficulties have lead to proposals where unlabeled data are processed by supervised classification algorithms. Here, we describe an estimation principle applicable to any probabilistic classifier, aiming at making the most of unlabeled data when they should be beneficial to the learning process, that is, when classes are well apart. The method enables to control the contribution of unlabeled examples, thus providing robustness with respect to the violation of the postulated low density separation assumption.

Section 9.2 motivates the estimation criterion. It is followed by the description of the optimization algorithms in section 9.3. The connections with some other principles or algorithms are then detailed in section 9.4. Finally, the experiments of section 9.5 offer a test bed to evaluate the behavior of entropy regularization, with comparisons to generative models and manifold learning.

9.2 Derivation of the Criterion

In this section, we first show that unlabeled data do not contribute to the maximum likelihood estimation of discriminative models. The belief that "unlabeled data should be informative" should then be encoded as a prior to modify the estimation process. We argue that assuming high entropy for $P(y|x)$ is a sensible encoding of this belief, and finally describe the learning criterion derived from this assumption.

9.2.1 Likelihood

The maximum likelihood principle is one of the main estimation technique in supervised learning, which is closely related to the more recent margin maximization techniques such as boosting and support vector machines [Friedman et al., 2000]. We start here by looking at the contribution of unlabeled examples to the (conditional) likelihood.

The learning set is denoted $\mathcal{L}_n = \{(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}, \ldots, x_n\}$, where the l first examples are labeled, and the $u = n - l$ last ones are unlabeled. We assume that labels are missing at random, that is, the missingness mechanism is independent from the missing class information. Let h be the random variable

missing value mechanism

missing at random

encoding missingness: $h = 1$ if y is hidden and $h = 0$ if y is observed. The missing at random assumption reads

$$P(h|x, y) = P(h|x) \ . \tag{9.1}$$

This assumption excludes cases where missingness may indicate a preference for a particular class (this can happen for example in opinion polls where the "refuse to answer" option may hide an inclination towards a shameful answer). Assuming independent examples, the conditional log-likelihood is then

$$L(\theta; \mathcal{L}_n) = \sum_{i=1}^{l} \ln P(y_i|x_i; \theta) + \sum_{i=1}^{n} \ln P(h_i|x_i) \ , \tag{9.2}$$

Maximizing (9.2) with respect to θ can be performed by dropping the second term of the right-hand side. It corresponds to maximizing the complete likelihood when no assumption whatsoever is made on $p(x)$ [McLachlan, 1992]. As unlabeled data are not processed by the model of posterior probabilities, they do not convey information regarding $P(y|x)$. In the maximum a posteriori (MAP) framework, unlabeled data are useless regarding discrimination when the priors on $p(x)$ and $P(y|x)$ factorize and are not tied (see chapter 2): observing x does not inform about y, unless the modeler assumes so. Benefiting from unlabeled data requires assumptions of some sort on the relationship between x and y. In the MAP framework, this will be encoded by a prior distribution. As there is no such thing like a universally relevant prior, we should look for an induction bias allowing to process unlabeled data when the latter are known to convey information.

9.2.2 When Are Unlabeled Examples Informative?

Theory provides little support to the numerous experimental evidences showing that unlabeled examples can help the learning process. Learning theory is mostly developed at the two extremes of the statistical paradigm: in parametric statistics where examples are known to be generated from a known class of distribution, and in the distribution-free Structural Risk Minimization (SRM) or Probably Approximately Correct (PAC) frameworks. Semi-supervised induction does not fit the distribution-free frameworks: no positive statement can be made without distributional assumptions, as for some distributions $p(x, y)$, unlabeled data are non-informative while supervised learning is an easy task. In this regard, generalizing from labeled and unlabeled data may differ from transductive inference.

In parametric statistics, theory has shown the benefit of unlabeled examples, either for specific distributions [O'Neill, 1978], or for mixtures of the form $p(x) = \pi p(x|y = 1) + (1-\pi)p(x|y = 2)$, where the estimation problem is essentially reduced to the one of estimating the mixture parameter π [Castelli and Cover, 1996]. These studies conclude that the (asymptotic) information content of unlabeled examples decreases as classes overlap.[1] Hence, in the absence of general results, postulating

information content of unlabeled examples

that classes are well apart, separated by a low density area, is sensible when one expects to take advantage of unlabeled examples.

9.2.3 A Measure of Class Overlap

There are many possible measures of class overlap. We chose Shannon's conditional entropy, which is invariant to the parameterization of the model, but the framework developped below could be applied to other measures of class overlap, such as Renyi entropies. Note however that the algorithms detailed in section 9.3.1 are specific to this choice. Obviously, the conditional entropy may only be related to the usefulness of unlabeled data where labeling is indeed ambiguous. Hence, the measure of class overlap should be conditioned on missingness

conditional entropy

$$\begin{aligned} H(y|x,h=1) &= -\mathbf{E}_{xy}\left[\ln P(y|x,h=1)\right] \\ &= -\int \sum_{m=1}^{M} \ln P(y=m|x,h=1) p(x,y=m|h=1)\, dx \ . \end{aligned} \tag{9.3}$$

In the MAP framework, assumptions are encoded by means of a prior on the model parameters. Stating that we expect a high conditional entropy does not uniquely define the form of the prior distribution, but the latter can be derived by resorting to the maximum entropy principle.[2]

The maximum entropy prior verifying $\mathbf{E}_\theta\left[H(y|x,h=1)\right] = c$, where the constant c quantifies how small the entropy should be on average, takes the form

$$p(\theta) \propto \exp\left(-\lambda H(y|x,h=1)\right) \ , \tag{9.4}$$

where λ is the positive Lagrange multiplier corresponding to the constant c.

Computing $H(y|x,h=1)$ requires a model of $p(x,y|h=1)$ whereas the choice of supervised classification is motivated by the possibility to limit modeling to conditional probabilities. We circumvent the need of additional modeling by applying the plug-in principle, which consists in replacing the expectation with respect to $(x|h=1)$ by the sample average. This substitution, which can be interpreted as "modeling" $p(x|h=1)$ by its empirical distribution, yields

plug-in principle

$$H_{\text{emp}}(y|x,h=1;\mathcal{L}_n) = -\frac{1}{u}\sum_{i=l+1}^{n}\sum_{m=1}^{M} P(m|x_i,t_i=\mathbf{1}) \ln P(m|x_i,t_i=\mathbf{1}) \ . \tag{9.5}$$

---

1. This statement, given explicitly by O'Neill [1978], is also formalized, though not stressed, by Castelli and Cover [1996], where the Fisher information for unlabeled examples at the estimate $\hat{\pi}$ is clearly a measure of the overlap between class conditional densities: $I_u(\hat{\pi}) = \int \frac{(p(x|y=1)-p(x|y=2))^2}{\hat{\pi}p(x|y=1)+(1-\hat{\pi})p(x|y=2)}\, dx$.

2. Here, maximum entropy refers to the construction principle which enables to derive distributions from constraints, not to the content of priors regarding entropy.

The missing at random assumption (9.1) yields $P(y|x, h = 1) = P(y|x)$, hence

$$H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n) = -\frac{1}{u} \sum_{i=l+1}^{n} \sum_{m=1}^{M} P(m|x_i) \ln P(m|x_i) \ . \tag{9.6}$$

This empirical functional is plugged in (9.4) to define an empirical prior on parameters θ, that is, a prior whose form is partly defined from data [Berger, 1985].

9.2.4 Entropy Regularization

The MAP estimate is defined as the maximizer of the posterior distribution, that is, the maximizer of

$$\begin{aligned} C(\theta, \lambda; \mathcal{L}_n) &= L(\theta; \mathcal{L}_n) - \lambda H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n) \\ &= \sum_{i=1}^{l} \ln P(y_i|x_i; \theta) + \lambda \sum_{i=l+1}^{n} \sum_{m=1}^{M} P(m|x_i; \theta) \ln P(m|x_i; \theta) \ , \end{aligned} \tag{9.7}$$

where the constant terms in the log-likelihood (9.2) and log-prior (9.4) have been dropped.

While $L(\theta; \mathcal{L}_n)$ is only sensitive to labeled data, $H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n)$ is only affected by the value of $P(m|x; \theta)$ on unlabeled data. Since these two components of the learning criterion are concave in $P(m|x; \theta)$, their weighted difference is usually not concave, except for $\lambda = 0$. Hence, the optimization surface is expected to possess local maxima, which are likely to be more numerous as u and λ grow. Semi-supervised induction is half-way between classification and clustering, hence, the progressive loss of concavity in the shift from supervised to unsupervised learning is not surprising, as most clustering optimization problems are nonconvex [Rose et al., 1990].

The empirical approximation H_{emp} (9.5) of H (9.3) breaks down for wiggly functions $P(m|\cdot)$ with abrupt changes between data points (where $p(x)$ is bounded from below). As a result, it is important to constrain $P(m|\cdot)$ in order to enforce the closeness of the two functionals. In the following experimental section, we imposed such a constraint on $P(m|\cdot)$ by adding a smoothness penalty to the criterion C (9.7). Note that this penalty also provides a means to control the capacity of the classifier.

9.3 Optimization Algorithms

9.3.1 Deterministic Annealing EM and IRLS

In its application to semi-supervised learning, the EM algorithm is generally used to maximize the joint likelihood from labeled and unlabeled data. This iterative algorithm increases the likelihood at each step and converges towards a stationary

point of the likelihood surface.

The criterion $C(\theta, \lambda; \mathcal{L}_n)$ (9.7) departs from the conditional likelihood by its entropy term. It is in fact formulated as each intermediate optimization subproblem solved in the deterministic annealing EM algorithm. This scheme was originally proposed to alleviate the difficulties raised by local maxima in joint likelihood for some clustering problems [Rose et al., 1990, Yuille et al., 1994]. It consists in optimizing the likelihood subject to a constraint on the level of randomness, measured by the entropy of the model of $P(y|x)$. The Lagrangian formulation of this optimization problem is precisely (9.7), where $T = 1 - \lambda$ is the analogue of a temperature. Deterministic annealing is the cooling process defining the continuous path of solutions indexed by the temperature. Following this path is expected to lead to a final solution with lower free energy, that is, higher likelihood.

Deterministic annealing

If the optimization criteria are identical, the goals, and the hyper-parameters used are different. On the one hand, in deterministic annealing EM, one aims at reaching the global maximum (or at least a good local optimum) of the joint likelihood. For this purpose, one starts from a concave function ($T \to \infty$) and the temperature is gradually lowered down to $T = 1$, in order to reach a state with high likelihood. On the other hand, the goal of entropy regularization is to alter the maximum likelihood solution, by biasing it towards low entropy. One starts from a possibly concave conditional likelihood ($\lambda = 0$, *i.e.*, $T = 1$) and the temperature is gradually lowered until it reaches some predetermined value $1 - \lambda_0 = T_0 \geq 0$, to return a good local maximum of $C(\theta, \lambda_0; \mathcal{L}_n)$.

Despite these differences, the analogy with deterministic annealing EM is useful because it provides an optimization algorithm for maximizing $C(\theta, \lambda; \mathcal{L}_n)$ (9.7). Deterministic annealing EM [Yuille et al., 1994] is a simple generalization of the standard EM algorithm. Starting from the solution obtained at the highest temperature, the path of solution is computed by gradually increasing λ. For each trial value of λ, the corresponding solution is computed by a two-step iterative procedure, where the expected log-likelihood is maximized at the M-step, and where soft (expected) assignments are imputed at the E-step for unlabeled data. The only difference with standard EM takes place at the E-step, where the expected value of labels is computed using the Gibbs distribution

Deterministic annealing EM

$$g_m(x_i; \theta) = \frac{P(m|x_i; \theta)^{\frac{1}{1-\lambda}}}{\sum_{\ell=1}^{M} P(\ell|x_i; \theta)^{\frac{1}{1-\lambda}}} \ ,$$

which distributes the probability mass according to the current estimated posterior $P(m|\cdot)$ (for labeled examples, the assignment is clamped at the original label $g_m(x_i; \theta) = \delta_{m y_i}$). For $0 < \lambda \leq 1$, the Gibbs distribution is more peaked than the estimated posterior. One recovers EM for $\lambda = 0$, and the hard assignments of classification EM (CEM) [Celeux and Govaert, 1992] correspond to $\lambda = 1$.

The M-step then consists in maximizing the expected log-likelihood with respect

to θ,

$$\theta^{s+1} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{m=1}^{M} g_m(x_i; \theta^s) \ln P(m|x_i; \theta) \ , \tag{9.8}$$

where the expectation is taken with respect to the distribution $(g_1(\cdot; \theta^s), \ldots, g_M(\cdot; \theta^s))$, and θ^s is the current estimate of θ.

The optimization problem (9.8) is concave in $P(m|x; \theta)$ and also in θ for logistic regression models. Hence it can be solved by second-order optimization algorithm, such as the Newton-Raphson algorithm, which is often referred to as iteratively reweighted least squares or IRLS in statistical textbooks [Friedman et al., 2000].

IRLS

We omit the detailed derivation of IRLS, and provide only the update equation for θ in the standard logistic regression model for binary classification problems. [3] The model of posterior distribution is defined as

$$P(1|x; \theta) = \frac{1}{1 + e^{-(\mathbf{w}^\top x + b)}} \ , \tag{9.9}$$

where $\theta = (\mathbf{w}, b)$. In the binary classification problem, the M-step (9.8) reduces to

$$\theta^{s+1} = \arg\max_{\theta} \sum_{i=1}^{n} g_1(x_i; \theta^s) \ln P(1|x_i; \theta) + (1 - g_1(x_i; \theta^s)) \ln(1 - P(1|x_i; \theta)) \ ,$$

where

$$g_1(x_i; \theta) = \frac{P(1|x_i; \theta)^{\frac{1}{1-\lambda}}}{P(1|x_i; \theta)^{\frac{1}{1-\lambda}} + (1 - P(1|x_i; \theta))^{\frac{1}{1-\lambda}}}$$

for unlabeled data and $g_1(x_i; \theta) = \delta_{1y_i}$ for labeled examples. Let $\mathbf{p}_\theta$ and $\mathbf{g}$ denote the vector of $P(1|x_i; \theta)$ and $g_1(x_i; \theta^s)$ values respectively, $\mathbf{X}$ the $(n \times (d+1))$ matrix of x_i values concatenated with the vector $\mathbf{1}$, and $\mathbf{W}_\theta$ the $(n \times n)$ diagonal matrix with ith diagonal entry $P(1|x_i; \theta)(1 - P(1|x_i; \theta))$. The Newton-Raphson update is

$$\theta \leftarrow \theta + \left(\mathbf{X}^\top \mathbf{W}_\theta \mathbf{X}\right)^{-1} \mathbf{X}^\top (\mathbf{g} - \mathbf{p}_\theta) \ . \tag{9.10}$$

Each Newton-Raphson update can be interpreted as solving a weighted least squares problem, and the scheme is iteratively reweighted by updating $\mathbf{p}_\theta$ (and hence $\mathbf{W}_\theta$) and applying (9.10) until convergence.

9.3.2 Conjugate Gradient

Depending on how $P(y|x)$ is modeled, the M-step (9.8) may not be concave, and other gradient-based optimization algorithms should be used. Even in the case

3. The generalization to kernelized logistic regression is straightforward, and the generalization to more than two classes results in similar expressions, but they would require numerous extra notations.

where a logistic regression model is used, conjugate gradient may turn out being computationally more efficient than the IRLS procedure. Indeed, even if each M-step of the deterministic annealing EM algorithm consists in solving a convex problem, this problem is non-quadratic. IRLS solves exactly each quadratic subproblem, a strategy which becomes computationally expensive for high dimensional data or kernelized logistic regression. The approximate solution provided by a few steps of conjugate gradient may turn out to be more efficient, especially since the solution θ^{s+1} returned at the sth M-step is not required to be accurate.

Depending on whether memory is an issue or not, conjugate gradient updates may use the optimal steps computed from the Hessian, or approximations returned by a line search. Theses alternatives have experimentally been shown to be much more efficient than IRLS on large problems [Komarek and Moore, 2003].

Finally, when EM does not provide a useful decomposition of the learning task, one can directly address the minimization of the learning criterion (9.7) with conjugate gradient, or other gradient-based algorithms. Here also, it is useful to define an annealing scheme, where λ is gradually increased from 0 to 1, in order to avoid poor local maxima of the optimization surface.

9.4 Related Methods

9.4.1 Minimum Entropy in Pattern Recognition

Minimum entropy regularizers have been used in other contexts to encode learnability priors [Brand, 1999]. In a sense, H_{emp} can be seen as a poor's man way to generalize this approach to continuous input spaces. This empirical functional was also used as a criterion to learn scale parameters in the context of transductive manifold learning [Zhu et al., 2003]. During learning, discrepancies between H (9.3) and H_{emp} (9.5) are prevented to avoid hard unstable solutions by smoothing the estimate of posterior probabilities.

9.4.2 Input-Dependent and Information Regularization

Input-dependent regularization, introduced by Seeger [2001] and detailed in chapter 2, aims at incorporating some knowledge about the density $p(x)$ in the modeling of $P(y|x)$. In the framework of Bayesian inference, this knowledge is encoded by structural dependencies in the prior distributions.

Information regularization introduced by Szummer and Jaakkola [2002], and extended as detailed in chapter **??** is another approach, where the density $p(x)$ is assumed to be known, and where the mutual information between variables x and y is penalized within predefined small neighborhoods. As the mutual information $I(x;y)$ is related to the conditional entropy by $I(x;y) = H(y) - H(y|x)$, low entropy and low mutual information are nearly opposite quantities. However, penalizing mutual information locally, subject to the class constraints provided by labeled

examples, highly penalizes the variations of $P(y|x)$ in the high density regions. Hence, like entropy regularization, information regularization favors solution where label switching is restricted to low density areas between disagreeing labels.

Entropy regularization differs from these schemes in that it is expressed only in terms of $P(y|x)$ and does not involve a model of $p(x)$. However, we stress that for unlabeled data, the MAP minimum entropy estimation is consistent with the maximum (complete) likelihood approach when $p(x)$ is small near the decision surface. Indeed, whereas the complete likelihood maximizes $\ln p(x)$ on unlabeled data, the regularizer minimizes the conditional entropy on the same points. Hence, the two criteria agree provided the class assignments are confident in high density regions, or conversely, when label switching occurs in a low density area.

9.4.3 Self-Training

Self-training is an iterative process, where a learner imputes the labels of examples which have been classified with confidence in the previous step. This idea, which predates EM, was independently proposed by several authors (see chapter 1). Amini and Gallinari [2002] analyzed this technique and have shown that it is equivalent to a version of the classification EM algorithm [Celeux and Govaert, 1992], which minimizes the likelihood deprived of the entropy of the partition.

In the context of conditional likelihood estimation from labeled and unlabeled examples, self-training minimizes C (9.7) with $\lambda = 1$. The optimization process itself is identical to the generalized EM described in section 9.3.1 with hard assignments [Grandvalet, 2002, Jin and Ghahramani, 2003].

Minimum entropy regularization is expected to have two benefits. First, the influence of unlabeled examples can be controlled, in the spirit of EM-λ [Nigam et al., 2000] Second, the deterministic annealing process, by which λ is slowly increased, is expected to help the optimization process to avoid poor local minima of the criterion. This scheme bears some similarity with the increase of the C^* parameter in the transductive SVM algorithm of Joachims [1999].

9.4.4 Maximal Margin Separators

Maximal margin separators are theoretically well founded models which have shown great success in supervised classification. For linearly separable data, they have been shown to be a limiting case of probabilistic hyperplane separators [Tong and Koller, 2000].

In the framework of transductive learning, Vapnik [1998] proposed to broaden the margin definition to unlabeled examples, by taking the smallest Euclidean distance between any (labeled and unlabeled) training point to the classification boundary. The following theorem, whose proof is given in Appendix 9.7, generalizes Theorem 5,

Corollary 6 of Tong and Koller [2000] to the margin defined in transductive learning[4] when using the proposed minimum entropy criterion.

Theorem 9.1 *Consider the two-class linear classification problem with linearly separable labeled examples, where the classifier is obtained by optimizing* $P(1|x;(\mathbf{w},b)) = 1/(1+e^{-(\mathbf{w}^\top x+b)})$ *with the semi-supervised minimum entropy criterion (9.7), under the constraint that* $||w|| \leq B$*. The margin of that linear classifier converges towards the maximum possible margin among all such linear classifiers, as the bound* B *goes to infinity.*

Hence, the minimum entropy solution can approach semi-supervised SVM [Vapnik, 1998, Bennett and Demiriz, 1998]. We however recall that the MAP criterion is not concave in $P(m|x;\theta)$, so that the convergence toward the global maximum cannot be guaranteed with the algorithms presented in section 9.3. This problem is shared by all inductive semi-supervised algorithms dealing with a large number of unlabeled data in reasonable time, such as mixture models or the transductive SVM of Joachims [1999]. Explicitly or implicitly, inductive semi-supervised algorithms impute labels which are somehow consistent with the decision rule returned by the learning algorithm. The enumeration of all possible configurations is only avoided thanks to a heuristic process, such as deterministic annealing, which may fail.

Most graph-based transduction algorithms avoid this enumeration problem because their labeling process is not required to comply with a parameterized decision rule. This clear computational advantage has however its counterpart: label propagation is performed via a user-defined similarity measure. The selection of a discriminant similarity measure is thus left to the user, or to an outer loop, in which case the overall optimization process is not convex anymore. The experimental section below illustrates that the choice of discriminant similarity measures is difficult in high dimensional spaces, and when *a priori* similar patterns should be discriminated.

9.5 Experiments

9.5.1 Artificial Data

In this section, we chose a simple experimental setup in order to avoid artifacts stemming from optimization problems. This setting enables to check to what extent supervised learning can be improved by unlabeled examples, and when minimum entropy can compete with generative methods which are traditionally advocated in this framework. The minimum entropy criterion is applied to the

4. That is, the margin on an unlabeled example is defined as the absolute value of the margin on a labeled example at the same location.

logistic regression model. It is compared to logistic regression fitted by maximum likelihood (ignoring unlabeled data) and logistic regression with all labels known. The former shows what has been gained by handling unlabeled data, and the latter provides the "crystal ball" ultimate performance obtained by guessing correctly all labels. All hyper-parameters (weight-decay for all logistic regression models plus the λ parameter (9.7) for minimum entropy) are tuned by ten-fold cross-validation.

These discriminative methods are compared to generative models. Throughout all experiments, a two-components Gaussian mixture model was fitted by the EM algorithm (two means and one common covariance matrix estimated by maximum likelihood on labeled and unlabeled examples [McLachlan, 1992]). The problem of local maxima in the likelihood surface is artificially avoided by initializing EM with the parameters of the true distribution when the latter is truly a two-component Gaussian mixture, or with maximum likelihood parameters on the (fully labeled) test sample when the distribution departs from the model. This initialization advantages EM, which is guaranteed to pick, among all local maxima of the likelihood, the one which is in the basin of attraction of the optimal value. In particular, this initialization prevents interferences that may result from the "pseudo-labels" given to unlabeled examples at the first E-step. The "label switching" problem (badly labeled clusters) is prevented at this stage.

Correct joint density model In the first series of experiments, we consider two-class problems in a 50-dimensional input space. Each class is generated with equal probability from a normal distribution. Class 1 is normal with mean $(a\ a \dots a)$ and unit covariance matrix. Class 2 is normal with mean $-(a\ a \dots a)$ and unit covariance matrix. Parameter a tunes the Bayes error which varies from 1 % to 20 % (1 %, 2.5 %, 5 %, 10 %, 20 %). The learning sets comprise l labeled examples, $(l = 50, 100, 200)$ and u unlabeled examples, $(u = l \times (1, 3, 10, 30, 100))$. Overall, 75 different setups are evaluated, and for each one, 10 different training samples are generated. Generalization performances are estimated on a test set of size 10 000.

This first benchmark provides a comparison for the algorithms in a situation where unlabeled data are known to convey information. Besides the favorable initialization of the EM algorithm to the optimal parameters, the generative models benefit from the *correctness* of the model: data were generated according to the model, that is, two Gaussian subpopulations with identical covariances. The logistic regression model is only *compatible* with the joint distribution, which is a weaker fulfillment than correctness.

As there is no modeling bias, differences in error rates are only due to differences in estimation efficiency. The overall error rates (averaged over all settings) are in favor of minimum entropy logistic regression (14.1 ± 0.3 %). EM (15.7 ± 0.3 %) does worse on average than logistic regression (14.9 ± 0.3 %). For reference, the average Bayes error rate is 7.7 % and logistic regression reaches 10.4 ± 0.1 % when all examples are labeled.

Figure 9.1 provides more informative summaries than these raw numbers. The first plot represents the error rates (averaged over l) versus Bayes error rate and

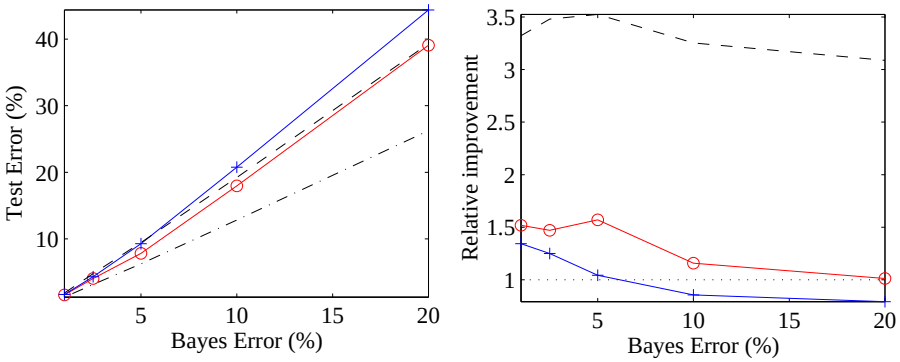

Figure 9.1 Left: test error of minimum entropy logistic regression (∘) and mixture models (+) *vs.* Bayes error rate for $u/l = 10$. The errors of logistic regression (dashed), and logistic regression with all labels known (dash-dotted) are shown for reference. right: relative improvement to logistic regression *vs.* Bayes error rate.

the u/l ratio. The second plot represents the same performances on a common scale along the abscissa, by showing the relative improvement of using unlabeled examples when compared to logistic regression ignoring unlabeled examples. The relative improvement is defined here as the ratio of the gap between test error and Bayes error for the considered method to the same gap for logistic regression. This plot shows that, as asymptotic theory suggests [O'Neill, 1978, Castelli and Cover, 1996], unlabeled examples are more beneficial when the Bayes error is low. This observation supports the relevance of the minimum entropy assumption.

Figure 9.2 illustrates the consequence of the demanding parametrization of generative models. Mixture models are outperformed by the simple logistic regression model when the sample size is low, since their number of parameters grows quadratically (*vs.* linearly) with the number of input features. This graph also shows that the minimum entropy model takes quickly advantage of unlabeled data when classes are well separated. With $u = 3l$, the model considerably improves upon the one discarding unlabeled data. At this stage, the generative models do not perform well, as the number of available examples is low compared to the number of parameters in the model. However, for very large sample sizes, with 100 times more unlabeled examples than labeled examples, the generative method eventually becomes more accurate than the discriminative one.

These results are reminiscent of those of Efron [1975], in the respect that the generative method is asymptotically slighly more efficient than the discriminative one, mainly because logistic regression makes little use of examples far from the decision surface. In the same respect, our observations differ from the comparison of Ng and Jordan [2001], which shows that naive Bayes can be competitive in terms of test error for small training sample sizes. This may be explained by the more general generative model used here, which does not assume feature independance.

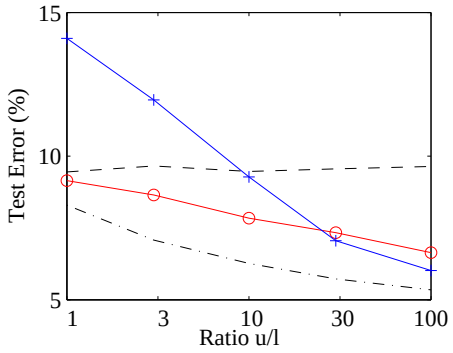

Figure 9.2 Test error *vs.* u/l ratio for 5 % Bayes error ($a = 0.23$). Test errors of minimum entropy logistic regression ($\circ$) and mixture models ($+$). The errors of logistic regression (dashed), and logistic regression with all labels known (dash-dotted) are shown for reference.

Misspecified joint density model In a second series of experiments, the setup is slightly modified by letting the class-conditional densities be corrupted by outliers. For each class, the examples are generated from a mixture of two Gaussians centered on the same mean: a unit variance component gathers 98 % of examples, while the remaining 2 % are generated from a large variance component, where each variable has a standard deviation of 10. The mixture model used by EM is now slightly misspecified since the whole distribution is still modeled by a simple two-components Gaussian mixture. The results, displayed in the left-hand-side of Figure 9.3, should be compared with Figure 9.2. The generative model dramatically suffers from the misspecification and behaves worse than logistic regression for all sample sizes. The unlabeled examples have first a beneficial effect on test error, then have a detrimental effect when they overwhelm the number of labeled examples. On the other hand, the discriminative models behave smoothly as in the previous case, and the minimum entropy criterion performance steadily improves with the addition of unlabeled examples.

The last series of experiments illustrate the robustness with respect to the cluster assumption, by which the decision boundary should be placed in low density regions. The samples are drawn from a distribution such that unlabeled data do not convey information, and where a low density $p(x)$ does not indicate class separation. This distribution is modeled by two Gaussian clusters, like in the first series of experiment, but labeling is now independent from clustering: example x_i belongs to class 1 if $x_{i2} > x_{i1}$ and belongs to class 2 otherwise: the Bayes decision boundary now separates each cluster in its middle. The mixture model is unchanged. It is now far from the model used to generate data. The right-hand-side plot of Figure 9.3 shows that the favorable initialization of EM does not prevent the model to be fooled by unlabeled data: its test error steadily increases with the amount of unlabeled data. Conversely, the discriminative models behave well, and the minimum entropy

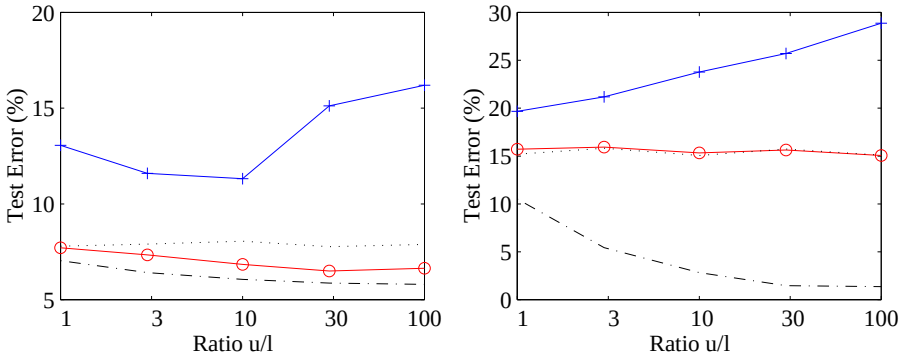

Figure 9.3 Test error *vs.* u/l ratio for $a = 0.23$. Average test errors for minimum entropy logistic regression ($\circ$) and mixture models ($+$). The test error rates of logistic regression (dotted), and logistic regression with all labels known (dash-dotted) are shown for reference. Left: experiment with outliers; right: experiment with uninformative unlabeled data.

Table 9.1 Error rates (%) of minimum entropy (ME) *vs.* consistency method (CM), for $a = 0.23$, $l = 50$, and a) pure Gaussian clusters b) Gaussian clusters corrupted by outliers c) class boundary separating one Gaussian cluster

n_u	50	150	500	1500
a) ME	10.8 ± 1.5	9.8 ± 1.9	8.8 ± 2.0	8.3 ± 2.6
a) CM	21.4 ± 7.2	25.5 ± 8.1	29.6 ± 9.0	26.8 ± 7.2
b) ME	8.5 ± 0.9	8.3 ± 1.5	7.5 ± 1.5	6.6 ± 1.5
b) CM	22.0 ± 6.7	25.6 ± 7.4	29.8 ± 9.7	27.7 ± 6.8
c) ME	8.7 ± 0.8	8.3 ± 1.1	7.2 ± 1.0	7.2 ± 1.7
c) CM	51.6 ± 7.9	50.5 ± 4.0	49.3 ± 2.6	50.2 ± 2.2

algorithm is not distracted by the two clusters; its performance is nearly identical to the one of training with labeled data only (cross-validation provides λ values close to zero), which can be regarded as the ultimate achievable performance in this situation.

Comparison with manifold transduction Although this chapter focuses on inductive classification, we also provide comparisons with a transduction algorithm relying on the manifold assumption. The consistency method [Zhou et al., 2004] is a very simple label propagation algorithm with only two tuning parameters. As suggested by Zhou et al. [2004], we set $\alpha = 0.99$ and the scale parameter σ^2 was optimized on test results. The results are reported in Table 9.1. The experiments are limited due to the memory requirements of the consistency method in our naive implementation.

Anger Fear Disgust Joy Sadness Surprise Neutral

Figure 9.4 Examples from the facial expression recognition database.

The results are extremely poor for the consistency method, whose error is way above minimum entropy, and which does not show any sign of improvement as the sample of unlabeled data grows. In particular, when classes do not correspond to clusters, the consistency method performs random class assignments.

In fact, the experimental setup, which was designed for the comparison of global classifiers, is not favorable to manifold methods, since the input data are truly 50-dimensional. In this situation, finding a discriminant similarity measure may require numerous degrees of freedom, and the consistency method provides only one tuning parameter: the scale parameter σ^2. Hence, these results illustrate that manifold learning requires more tuning efforts for truly high dimensional data, and some recent techniques may respond to this need [Sindhwani et al., 2005].

9.5.2 Facial Expression Recognition

We now consider an image recognition problem, consisting in recognizing seven (balanced) classes corresponding to the universal emotions (anger, fear, disgust, joy, sadness, surprise and neutral). The patterns are gray level images of frontal faces, with standardized positions, as displayed in figure 9.4. The data set comprises 375 such pictures made of 140×100 pixels [Abboud et al., 2003, Kanade et al., 2000]

We tested kernelized logistic regression (Gaussian kernel), its minimum entropy version, nearest neigbor and the consistency method. We repeatedly (10 times) sampled 1/10 of the data set for providing the labeled part, and the remainder for testing. Although (α, σ^2) were chosen to minimize the test error, the consistency method performed poorly with 63.8 ± 1.3 % test error (compared to 86 % error for random assignments). Nearest-neighbor get similar results with 63.1 ± 1.3 % test error, and Kernelized logistic regression (ignoring unlabeled examples) improved to reach 53.6 ± 1.3 %. Minimum entropy kernelized logistic regression regression achieves 52.0 ± 1.9 % error (compared to about 20 % errors for human on this database). The scale parameter chosen for kernelized logistic regression (by ten-fold cross-validation) amount to use a global classifier.

The failure of local methods may be explained by the fact that the database contains several pictures of each person, with different facial expressions. Hence, local methods are likely to pick the same identity instead of the same expression, while global methods are able to learn the discriminating directions.

9.6 Conclusion

Although discriminative methods do not benefit from unlabeled data in the maximum likelihoood framework, *maximum a posteriori* estimation enables to address the semi-supervised induction problem. The information content of unlabeled data decreases with class overlap, which can be measured by the conditional entropy of labels given patterns. Hence, the minimum entropy prior encodes a premise of semi-supervised induction, that is, the belief that unlabeled data may be useful. The postulate is optimistic in some problems where unlabeled data do not convey information regarding labels, but the strength of the prior is controlled by a tuning parameter, so that the contribution of unlabeled examples to the estimate may vanish.

Minimum entropy regularization is related to self-training in general and to transductive SVMs in particular. It promotes classifiers with high confidence on the unlabeled examples. A deterministic annealing process smoothly drives the decision boundary away from unlabeled examples, favoring low density separation.

The regularizer can be applied to local and global model of posterior probabilities. As a result, it can improve over local models when they suffer from the curse of dimensionality. Minimum entropy regularization may also be a serious contender to generative methods. It compares favorably to these mixture models in three situations: for small sample sizes, where the generative model cannot completely benefit from the knowledge of the correct joint model; when the joint distribution is (even slightly) misspecified; when the unlabeled examples turn out to be non-informative regarding class probabilities.

Finally, the algorithms presented in this chapter can be applied to a generalized version of the semi-supervised induction problem, where the examples may be labeled by any subset of labels, representing the set of plausible classes. This kind of information is sometimes a more faithful description of the true state of knowledge when labeling is performed by an expert.

9.7 Proof of theorem 9.1

Theorem 9.1 *Consider the two-class linear classification problem with linearly separable labeled examples, where the classifier is obtained by optimizing* $P(1|x;(\mathbf{w},b)) = 1/(1+e^{-(\mathbf{w}^\top x+b)})$ *with the semi-supervised minimum entropy criterion (9.7), under the constraint that* $||\mathbf{w}|| \leq B$*. The margin of that linear classifier converges towards the maximum possible margin among all such linear classifiers, as the bound* B *goes to infinity.*

Proof.

Consider the logistic regression model $P(1|x;\theta)$ parameterized by $\theta = (\mathbf{w}, b)$. Let $z_i \in \{-1, +1\}$ be a binary variable defined as follows: if x_i is a positive labeled example, $z_i = +1$; if x_i is a negative labeled example, $z_i = -1$; if x_i is an unlabeled

example, $z_i = \text{sign}(P(1|x;\theta) - 1/2)$. The *margin* for the ith labeled or unlabeled example is defined as $m_i(\theta) = z_i(\mathbf{w}^\top x_i + b)$.

The criterion C (9.7) can be written as a function of $m_i = m_i(\theta)$ as follows:

$$C(\theta) = -\sum_{i=1}^{l} \ln(1 + e^{-m_i}) - \lambda \sum_{i=l+1}^{n} \left(\ln(1 + e^{-m_i}) + \frac{m_i e^{-m_i}}{1 + e^{-m_i}} \right) \ , \qquad (9.11)$$

where the indices $[1, l]$ and $[l + 1, n]$ correspond to labeled and unlabeled data, respectively.

On the one hand, for all θ such that there exists an example with non-negative margin, the cost (9.11) is trivially upper-bounded by $-\ln(2)$ if the example is labeled and $-\lambda \ln(2)$ otherwise. On the other hand, by the linear separability assumption, there exists $\theta = (\mathbf{w}, b)$ with, say, $||\mathbf{w}|| = 1$ such that $m_i > 0$. Consider now the cost obtained with the admissible solution $B\theta$ as $B \to +\infty$. In this limit, since $m_i(B\theta) = Bm_i(\theta)$, all the terms of the finite sum (9.11) converge to zero, so that the value of the cost converges to its maximum value ($\lim_{B\to+\infty} C(B\theta) = 0$). Hence, in the limit of $B \to +\infty$ all margins of the maximizer of C are positive.

We now show that the maximizer of C achieves the largest minimal margin. The cost (9.11) is simplified by using the following equivalence relations when $B \to +\infty$:

$$\begin{aligned} \ln(1 + e^{-Bm_i}) &\sim e^{-Bm_i} \\ \frac{Bm_i e^{-Bm_i}}{1 + e^{-Bm_i}} &\sim Bm_i e^{-Bm_i} \ , \end{aligned}$$

which yields

$$C(B\theta) = -\sum_{i=1}^{l} e^{-Bm_i} + o(e^{-Bm_i}) - \lambda \sum_{i=l+1}^{n} Bm_i e^{-Bm_i} + o(Bm_i e^{-Bm_i}) \ .$$

Let us write $m^* > 0$ the minimum margin among the labeled examples and $m_* > 0$ the minimum margin among the unlabeled examples, N^* the number of minimum margin labeled examples (with $m_i = m^*$) and N_* the number of minimum margin unlabeled examples (with $m_i = m_*$). As $e^{-Bm_i} = o(e^{-Bm^*})$ when $m_i > m^*$, we obtain

$$C(B\theta) = -N^* e^{-Bm^*} + o(e^{-Bm^*}) - \lambda N_* Bm_* e^{-Bm_*} + o(Bm_* e^{-Bm_*}) \ .$$

Now we note that if $m^* < m_*$, then $Bm_* e^{-Bm_*} = o(e^{-Bm^*})$, and that if $m^* \geq m_*$ then $e^{-Bm^*} = o(Bm_* e^{-Bm^*})$. Hence, depending on whether $m^* < m_*$ or $m^* \geq m_*$ we either obtain

$$C(B\theta) = -N^* e^{-Bm^*} + o(e^{-Bm^*}) \qquad (9.12)$$

or

$$C(B\theta) = -\lambda N_* Bm_* e^{-Bm_*} + o(Bm_* e^{-Bm^*}) \ . \qquad (9.13)$$

Now, consider two different values of θ, θ_1 and θ_2, giving rise to minimum margins M_1 and M_2 respectively, with $M_1 > M_2$. The solution $B\theta_1$ will be prefered to $B\theta_2$ if $C(B\theta_1)/C(B\theta_2) < 1$. From (9.12) and (9.13), we see that it does not matter whether M_i is among the labels or the unlabeled, but only whether $M_1 > M_2$ or $M_2 > M_1$. In all cases $C(B\theta_1)/C(B\theta_2) \to 0$ when $M_1 > M_2$. This allows to conclude that as $B \to \infty$, the global maximum of $C(B\theta)$ over θ tends to a maximum margin solution, where the minimum margin M (over both labeled and unlabeled examples) is maximized.

References

B. Abboud, F. Davoine, and M. Dang. Expressive face recognition and synthesis. In *Computer Vision and Pattern Recognition Workshop*, volume 5, page 54, 2003.

Y. S. Abu-Mostafa. Machines that learn from hints. *Scientific American*, 272(4):64–69, 1995.

A. K. Agrawala. Learning with a probabilistic teacher. *IEEE Transactions on Information Theory*, 16:373–379, 1970.

S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

M. Amini and P. Gallinari. Semi-supervised logistic regression. In *ECAI*, 2002.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, 2004.

M. Belkin and P. Niyogi. Using manifold structure for partially labeled classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.

R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, New Jersey, 1961.

Y. Bengio, O. Delalleau, and N. Le Roux. The curse of dimensionality for local kernel machines. Technical Report 1258, Département d'informatique et recherche opérationnelle, Université de Montréal, 2005.

Y. Bengio, O. Delalleau, and N. Le Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006a.

Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004.

Y. Bengio, H. Larochelle, and P. Vincent. Non-local manifold parzen windows. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006b.

Y. Bengio and M. Monperrus. Non-local manifold tangent learning. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

K. Bennett and A. Demiriz. Semi-supervised support vector machines. In *NIPS*, volume 12, 1998.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In *NIPS*, Cambridge, MA, USA, 2004. MIT Press.

M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, 1999.

V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. on Information Theory*, 42(6): 2102–2117, 1996.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

T. F. Cox and M. A. Cox. *Multidimensional Scaling*. Chapman & Hall, 1994.

O. Delalleau, Y. Bengio, and N. Le Roux. Efficient non-parametric function induction in semi-supervised learning. In *Proceedings of the Tenth International Workshop on Artificial*

Intelligence and Statistics, 2005.

P. G. Doyle and J. L. Snell. Random walks and electric networks. *Mathematical Association of America*, 1984.

B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

B. Fischer, V. Roth, and J. M. Buhmann. Clustering with the connectivity kernel. In *NIPS*, volume 16, 2004.

S. C. Fralick. Learning to recognize patterns wothout a teacher. *IEEE Transactions on Information Theory*, 13:57–64, 1967.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Y. Grandvalet. Logistic regression for partial labels. In *th Information Processing and Management of Uncertainty in Knowledge-based System*, pages 1935–1941, 2002.

B. Haasdonk. Feature space interpretation of SVMs with indefinite kernels. *IEEE TPAMI*, 2004. In press.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, http://www.xplore-stat.de/ebooks/ebooks.html, 2004.

R. Jin and Z. Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

T. Joachims. Transductive learning via spectral graph partitioning. In *ICML*, 2003.

T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

P. Komarek and A. Moore. Fast robust logistic regression for large sparse datasets with binary outputs. In *Artificial Intelligence and Statistics*, 2003.

G. Lebanon. Learning riemannian metrics. In *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley, 1992.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

A. Y. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *NIPS*, volume 14, pages 841–848. MIT Press, 2001.

K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchel. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

T. J. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978.

C. S. Ong, X. Mary, S. Canu, and A. J. Smola. Learning with non-positive kernels. In *ICML*, pages 639–646, 2004.

M. Ouimet and Y. Bengio. Greedy spectral embedding. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

S. Rosenberg. *The Laplacian on a Riemannian Manifold*. Cambridge University Press, Cambridge, UK, 1997.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, MA, 1996.

L. K. Saul and M. I. Jordan. A variational model for model-based interpolation. In *NIPS*, volume 9, 1997.

H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11:363–371, 1965.

M. Seeger. Learning with labeled and unlabeled data. Technical report, Edinburgh University, 2001.

V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, 2005.

M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *NIPS*, volume 14, 2001.

M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

S. Tong and D. Koller. Restricted bayes optimal classifiers. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 658–664, 2000.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

P. Vincent and Y. Bengio. Density-sensitive metrics and kernels. Presented at the Snowbird Learning Workshop, 2003.

G. S. Watson. Smooth regression analysis. *Sankhya - The Indian Journal of Statistics*, 26:359–372, 1964.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA, 2001. MIT Press.

A. L. Yuille, P. Stolorz, and J. Utans. Statistical physics, mixtures of distributions, and the EM algorithm. *Neural Computation*, 6(2):334–340, 1994.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, volume 16, 2004.

X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.