Semi-supervised Learning by Entropy Minimization ⋆

Yves Grandvalet[1], Yoshua Bengio[2]

[1] Heudiasyc, UMR 6599 CNRS/UTC
60205 Compiègne cedex, France
grandval@utc.fr

[2] Dept. IRO, Université de Montréal
Montreal, Qc, H3C 3J7, Canada
bengioy@iro.umontreal.ca

Abstract : We consider the semi-supervised learning problem, where a decision rule is to be learned from labeled and unlabeled data. In this framework, we motivate minimum entropy regularization, which enables to incorporate unlabeled data in the standard supervised learning. This regularizer can be applied to any model of posterior probabilities. Our approach provides a new motivation for some existing semi-supervised learning algorithms which are particular or limiting instances of minimum entropy regularization. A series of experiments illustrates that the proposed solution benefits from unlabeled data. The method challenges mixture models when the data are sampled from the distribution class spanned by the generative model. The performances are definitely in favor of minimum entropy regularization when generative models are misspecified, and the weighting of unlabeled data provides robustness to the violation of the "cluster assumption". Finally, we also illustrate that the method can be far superior to manifold learning in high dimension spaces, and also when the manifolds are generated by moving examples along the discriminating directions.

Résumé : Nous considérons le problème de l'apprentissage semi-supervisé, où une règle de décision est induite sur la base d'exemples étiquetés et non-étiquetés. Dans ce cadre, nous motivons l'utilisation de la régularisation par minimum d'entropie, qui permet d'utiliser les données non-étiquetées dans l'apprentissage de modèles discriminants. Cette technique peut être appliquée à tout modèle discriminant estimant des probabilités *a posteriori*. Notre approche fournit un nouveau point de vue sur certains algorithmes d'apprentissage semi-supervisé existants, qui peuvent être interprétés comme cas particulier ou limite de la régularisation par minimum d'entropie. Une série d'expérience illustre que la solution proposée permet de bénéficier de données non-étiquetées. Les résultats

⋆This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778. This publication only reflects the authors' views.

concurrencent ceux des modèles génératifs quand ces derniers utilisent le modèle correct, correspondant à celui de la distribution des données. Les performances sont clairement en faveur de la régularisation par minimum d'entropie quand le modèle devient incorrect, et la pondération des exemples non-étiquetés permet d'atteindre des solutions robustes aux violations de l'hypothèse de départ, qui postule que les classes sont bien séparées. Finalement, nous illustrons que la méthode peut également être de loin supérieure aux techniques récentes d'apprentissage de variété, que ce soit dans les espaces de grande dimension, ou quand les variétés sont produites par des exemples transformés sur les directions discriminantes.

Keywords : semi-supervised learning, minimum entropy, transduction, self-training, EM algorithm, spectral methods, logistic regression.

1 Introduction

In the classical supervised learning classification framework, a decision rule is to be learned from a learning set $\mathcal{L}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$, where each example is described by a pattern $\mathbf{x}_i \in \mathcal{X}$ and by the supervisor's response $y_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$. Here, we consider semi-supervised learning, where the supervisor's responses are limited to a subset of $\mathcal{L}_n$.

In the terminology used here, semi-supervised learning refers to learning a decision rule on $\mathcal{X}$ from labeled and unlabeled data. However, the related problem of transductive learning, that is, of predicting labels on a set of predefined patterns, is addressed as a side issue. Semi-supervised problems occur in many applications where labeling is performed by human experts. They have been receiving much attention during the last few years, but some important issues are unresolved (Seeger, 2002).

In the probabilistic framework, semi-supervised learning can be modeled as a missing data problem, which can be addressed by generative models such as mixture models thanks to the EM algorithm and extensions thereof (McLachlan, 1992). Generative models apply to the joint density of patterns and class (X, Y). They have appealing features, but they also have major drawbacks. Their estimation is much more demanding than discriminative models, since the model of $P(X, Y)$ is exhaustive, hence necessarily more complex than the model of $P(Y|X)$. More parameters are to be estimated, resulting in more uncertainty in the estimation process. The generative model being more precise, it is also more likely to be misspecified. Finally, the fitness measure is not discriminative, so that better models are not necessarily better predictors of class labels. These difficulties have lead to proposals where unlabeled data are processed in supervised classification algorithms (Bennett & Demiriz, 1999; Joachims, 1999; Amini & Gallinari, 2002; Grandvalet, 2002; Szummer & Jaakkola, 2003). Here, we propose an estimation principle applicable to any probabilistic classifier, aiming at making the most of unlabeled data when they should be beneficial, while controling their contribution to provide robustness to the learning scheme.

2 Derivation of the Criterion

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 $\mathcal{L}_n = \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$, where $\mathbf{z}_i \in \{0,1\}^K$ denotes the dummy variable representing the actually available labels (while y represents the precise and complete class information): if $\mathbf{x}_i$ is labeled ω_k, then $z_{ik} = 1$ and $z_{i\ell} = 0$ for $\ell \neq k$; if $\mathbf{x}_i$ is unlabeled, then $z_{i\ell} = 1$ for $\ell = 1, \ldots, K$. More generally, $\mathbf{z}$ should be thought as the index of possible labels. Hence, an imprecise knowledge, such that "$\mathbf{x}_i$ belongs either to class ω_1 or ω_2" can be encoded by letting z_{i1} and z_{i2} to be one, and $\mathbf{z}_{ik} = 0$ for $k > 2$.

We assume that labels are missing at random, that is, the missingness mechanism is independent from the missing information: the label is missing because it is not observed, not because it is deliberately hidden. In the general setup where $\mathbf{z}$ can indicate any subset of Ω, some information is missing when two or more labels are possible. The missing-at-random assumption reads $P(\mathbf{z}|\mathbf{x}, Y = \omega_k) = P(\mathbf{z}|\mathbf{x}, Y = \omega_\ell)$ for all (k, ℓ) such that $z_k = z_\ell = 1$.

Assuming independent examples, and noting that the event "y_i belongs to the subset indicated by $\mathbf{z}_i$" follows a Bernoulli distribution of parameter $\sum_{k=1}^K z_{ik} P(Y = \omega_k|\mathbf{x}_i)$, the conditional log-likelihood of $(Z|X)$ on the observed sample is then

$$L(\boldsymbol{\theta}; \mathcal{L}_n) = \sum_{i=1}^n \log \left(\sum_{k=1}^K z_{ik} f_k(\mathbf{x}_i; \boldsymbol{\theta}) \right) + h(\mathbf{z}_i) \ , \tag{1}$$

where $h(\mathbf{z})$, which does not depend on $P(X, Y)$, is only affected by the missingness mechanism, and $f_k(\mathbf{x}; \boldsymbol{\theta})$ is the model of $P(Y = \omega_k|\mathbf{x})$ parameterized by $\boldsymbol{\theta}$.

This criterion is a concave function of $f_k(\mathbf{x}_i; \boldsymbol{\theta})$, and for simple models such as the ones provided by logistic regression, the semi-supervised objective function is also concave in $\boldsymbol{\theta}$, so that the global solution can be obtained by numerical optimization. Maximizing (1) corresponds to maximizing the complete likelihood if no assumption whatsoever is made on $P(X)$ (McLachlan, 1992).

Provided $f_k(\mathbf{x}_i; \boldsymbol{\theta})$ sum to one, the likelihood is not affected by unlabeled data: unlabeled data convey no information. In the maximum a posteriori (MAP) framework, Seeger (2002) remarks that unlabeled data are useless regarding discrimination when the priors on $P(X)$ and $P(Y|X)$ factorize: observing $\mathbf{x}$ does not inform about y, unless the modeler assumes so. Benefitting 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 is known to convey information.

2.2 When Are Unlabeled Examples Informative?

Theory provides little support to the numerous experimental evidences, such as (Joachims, 1999; Nigam & Ghani, 2000; Nigam *et al.*, 2000), 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 learning, in the terminology used here, 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 $P(\mathbf{x}) = pP(\mathbf{x}|Y = \omega_1) + (1 - p)P(\mathbf{x}|Y = \omega_2)$ where the estimation problem is essentially reduced to the one of estimating the mixture parameter p (Castelli & Cover, 1996). These studies conclude that the (asymptotic) information content of unlabeled examples decreases as classes overlap.[1] Thus, the assumption that classes are well apart, separated by a low density area, is sensible if we expect to take advantage of unlabeled examples.

2.3 A Measure of Class Overlap

The conditional entropy $H(Y|X)$ is a measure of class overlap, which is invariant to the parameterization of the model. The entropy may be related to the usefulness of unlabeled data only where labeling is indeed ambiguous. Hence, we propose to measure the conditional entropy of class labels conditioned on the observed variables

$$H(Y|X, Z) = -E_{XYZ}[\log P(Y|X, Z)] \ , \tag{2}$$

where E_X denotes the expectation with respect to X.

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] Let $(\boldsymbol{\theta}, \boldsymbol{\psi})$ denote the model parameters of $P(X, Y, Z)$; the maximum entropy prior verifying $E_{\Theta\Psi}[H(Y|X, Z)] = c$, where the constant c quantifies how small the entropy should be on average, takes the form

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto \exp\left(-\lambda H(Y|X, Z)\right) \ , \tag{3}$$

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

[1] This statement, given explicitly by O'Neill (1978), is also formalized, though not stressed, by Castelli & Cover (1996), where the Fisher information for unlabeled examples at the estimate $\hat{p}$ is clearly a measure of the overlap between class conditional densities: $I_u(\hat{p}) = \int \frac{(P(\mathbf{x}|Y=\omega_1) - P(\mathbf{x}|Y=\omega_2))^2}{\hat{p}P(\mathbf{x}|Y=\omega_1) + (1-\hat{p})P(\mathbf{x}|Y=\omega_2)} \, d\mathbf{x}$.

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

Computing $H(Y|X,Z)$ requires a model of $P(X,Y,Z)$ 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,Z) by the sample average. This substitution, which can be interpreted as "modeling" $P(X,Z)$ by its empirical distribution, yields

$$H_{\text{emp}}(Y|X,Z;\mathcal{L}_n) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} P(Y=\omega_k|\mathbf{x}_i,\mathbf{z}_i)\log P(Y=\omega_k|\mathbf{x}_i,\mathbf{z}_i) \quad . \quad (4)$$

This empirical functional is plugged in (3) to define an empirical prior on parameters $\boldsymbol{\theta}$, that is, a prior whose form is partly defined from data (Berger, 1985).

2.4 Entropy Regularization

As detailed in appendix A, the missing-at-random assumption implies

$$P(Y=\omega_k|\mathbf{x},\mathbf{z}) = \frac{z_k P(Y=\omega_k|\mathbf{x})}{\sum_{\ell=1}^{K} z_\ell P(Y=\omega_\ell|\mathbf{x})} \quad . \quad (5)$$

Recalling that $f_k(\mathbf{x};\boldsymbol{\theta})$ denotes the model of $P(Y=\omega_k|\mathbf{x})$, the model of $P(Y=\omega_k|\mathbf{x},\mathbf{z})$ is defined as follows:

$$g_k(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \frac{z_k f_k(\mathbf{x};\boldsymbol{\theta})}{\sum_{\ell=1}^{K} z_\ell f_\ell(\mathbf{x};\boldsymbol{\theta})} \quad .$$

For labeled data, $g_k(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = z_k$, and for unlabeled data, $g_k(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = f_k(\mathbf{x};\boldsymbol{\theta})$.

From now on, we drop the reference to parameter $\boldsymbol{\theta}$ in f_k and g_k to lighten notation. The MAP estimate is defined as the maximizer of the posterior distribution, that is, the maximizer of

$$\begin{aligned} C(\boldsymbol{\theta},\lambda;\mathcal{L}_n) &= L(\boldsymbol{\theta};\mathcal{L}_n) - \lambda H_{\text{emp}}(Y|X,Z;\mathcal{L}_n) \\ &= \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i)\right) + \lambda\sum_{i=1}^{n}\sum_{k=1}^{K} g_k(\mathbf{x}_i,\mathbf{z}_i)\log g_k(\mathbf{x}_i,\mathbf{z}_i) \quad , (6) \end{aligned}$$

where the constant terms in the log-likelihood (1) and log-prior (3) have been dropped. While $L(\boldsymbol{\theta};\mathcal{L}_n)$ is only sensitive to labeled data, $H_{\text{emp}}(Y|X,Z;\mathcal{L}_n)$ is only affected by the value of $f_k(\mathbf{x})$ on unlabeled data.

Note that the empirical approximation H_{emp} (4) of H (2) breaks down for wiggly functions $f_k(\cdot)$ with abrupt changes between data points (where $P(X)$ is bounded from below). As a result, it is important to constrain $f_k(\cdot)$ in order to enforce the closeness of the two functionals. In the following experimental section, we imposed such a constraint on $f_k(\cdot)$ by adding to the criterion C (6) a smoothness penalty.

3 Related Work

3.1 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 by Zhu *et al.* (2003) as a criterion to learn weight function parameters in the context of transduction in manifold learning.

3.2 Input-Dependent and Information Regularization

Input-dependent regularization aims at incorporating some knowledge about the density $P(X)$ in the modeling of $P(Y|X)$. In the framework of Bayesian inference, Seeger (2002) proposes to encode this knowledge by structural dependencies in the prior distributions. Information regularization, proposed by Szummer & Jaakkola (2003) and later developped by Corduneanu & Jaakkola (2003), is another approach where the density $P(X)$ is assumed to be known, and where the mutual information between variables X and Y is supposed to be low within predefined neighborhoods.

Entropy regularization differs from input-dependent regularization in that it is expressed only in terms of $P(Y|X,Z)$ and does not involve a model of $P(X)$. However, we stress that for unlabeled data, the MAP estimation is consistent with the maximum (complete) likelihood approach when $P(X)$ is small near the decision surface. Indeed, whereas the complete likelihood maximizes $\log P(X)$ 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.

3.3 Self-Training

Self-training (Nigam & Ghani, 2000) is an iterative process, where a learner imputes the labels of examples which have been classified with confidence in the previous step. Amini & Gallinari (2002) analyzed this technique and have shown that it is equivalent to a version of the classification EM algorithm, which minimizes the likelihood deprived of the entropy of the partition. In the context of conditional likelihood with labeled and unlabeled examples, the criterion is

$$\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i) \right) + \sum_{k=1}^{K} g_k(\mathbf{x}_i) \log g_k(\mathbf{x}_i) \ ,$$

which is recognized as an instance of the criterion (6) with $\lambda = 1$.

Self-confident logistic regression (Grandvalet, 2002), also proposed by Jin & Ghahramani (2003) as "the EM model", is another algorithm optimizing the criterion for $\lambda = 1$. Using smaller λ values 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, slowly increasing λ defines a scheme similar to the increase of the C^* parameter in the

transductive SVM algorithm of Joachims (1999). These schemes are somewhat similar to the deterministic annealing procedures, used for example in clustering (Rose *et al.*, 1990). They are expected to help the optimization process to avoid poor local minima of the criterion.

3.4 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 & 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 B, generalizes Theorem 5, Corollary 6 of Tong & Koller (2000) to the margin defined in transductive learning when using the proposed minimum entropy criterion.

Theorem 1
In the two-class linear separable case, the logistic regression model with bounded weights, fitted by the minimum entropy criterion, converges towards the maximum margin separator (with maximal distance from labeled and unlabeled examples) as the bound goes to infinity.

Hence, the minimum entropy solution can closely mimic semi-supervised SVM (Bennett & Demiriz, 1999), which partially solves the enumeration problem of the original solution proposed by Vapnik (1998).

Note however that our criterion is not concave in f_k, so that the convergence toward the global maximum cannot be guaranteed. To our knowledge, this apparent fault is shared by all inductive semi-supervised algorithms (learning a decision rule) 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 consistent with a decision rule. The enumeration of all possible configurations is only avoided thanks to an heuristic process which may fail. Most 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 predefined, hence non-discriminant, similarity measure. The experimental section below demonstrates that this may be a serious shortcomming in high dimensional spaces, or when *a priori* similar patterns should be discriminated.

4 Experiments

4.1 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 if minimum entropy can compete with generative methods which are usually advocated in this framework.

The minimum entropy criterion is applied to the 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 (6) for minimum entropy) are tuned by ten-fold cross-validation.

Minimum entropy logistic regression is also compared to the classic EM algorithm for Gaussian mixture models (two means and one common covariance matrix estimated by maximum likelihood on labeled and unlabeled examples (McLachlan, 1992). Bad local maxima of the likelihood function are avoided by initializing EM with the parameters of the true distribution when the latter is a Gaussian mixture, or with maximum likelihood parameters on the (fully labeled) test sample when the distribution departs from the model. This initialization advantages EM, since it is guaranteed to pick, among all local maxima of the likelihood, the one which is in the basin of attraction of the optimal value. Furthermore, this initialization prevents interferences that may result from the "pseudo-labels" given to unlabeled examples at the first E-step. In particular, "label switching" (badly labeled clusters) is avoided at this stage.

4.1.1 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 ω_1 is normal with mean $(aa \dots a)$ and unit covariance matrix. Class ω_2 is normal with mean $-(aa \dots a)$ 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 n_l labeled examples, ($n_l = 50, 100, 200$) and n_u unlabeled examples, ($n_u = n_l \times (1, 3, 10, 30, 100)$). 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 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, EM benefits 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 1 provides more informative summaries than these raw numbers. The plots represent the error rates (averaged over n_l) versus Bayes error rate and the n_u/n_l ratio. The first plot shows that, as asymptotic theory suggests (O'Neill, 1978; Castelli & Cover, 1996), unlabeled examples are more beneficial when the Bayes error is low. This observation supports the relevance of the minimum entropy assumption.

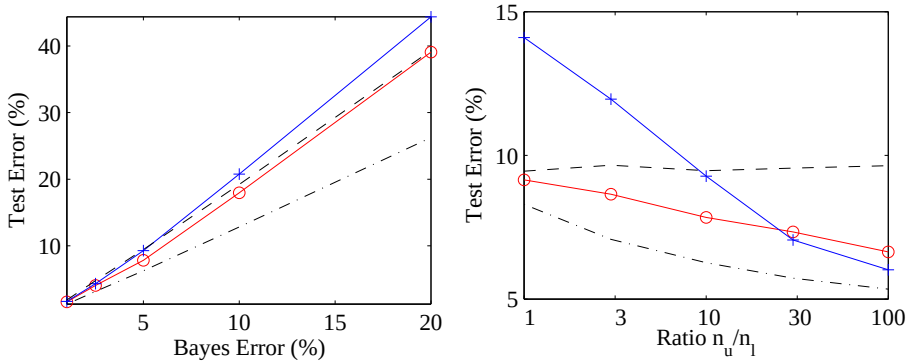

Figure 1: Left: test error *vs.* Bayes error rate for $n_u/n_l = 10$; right: test error *vs.* n_u/n_l ratio for 5 % Bayes error ($a = 0.23$). Test errors of minimum entropy logistic regression ($\circ$) and mixture models ($+$). The errors of logistic regression (dashed), and logistic regression with all labels known (dash-dotted) are shown for reference.

The second plot illustrates that the minimum entropy model takes quickly advantage of unlabeled data when classes are well separated. With $n_u = 3n_l$, the model considerably improves upon the one discarding unlabeled data. This graph also illustrates the consequence of the demanding parametrization of generative models. For very large sample sizes, with 100 times more unlabeled examples than labeled examples, the generative approach eventually becomes more accurate than the diagnosis approach. However, mixture models are outperformed by the simple logistic regression model when the sample size is low, because their number of parameters is quadratic (*vs.* linear) in the number of input features.

4.1.2 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 2, should be compared with the right-hand-side of Figure 1. 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 diagnosis models behave smoothly as in the previous case, and the minimum entropy criterion performance improves.

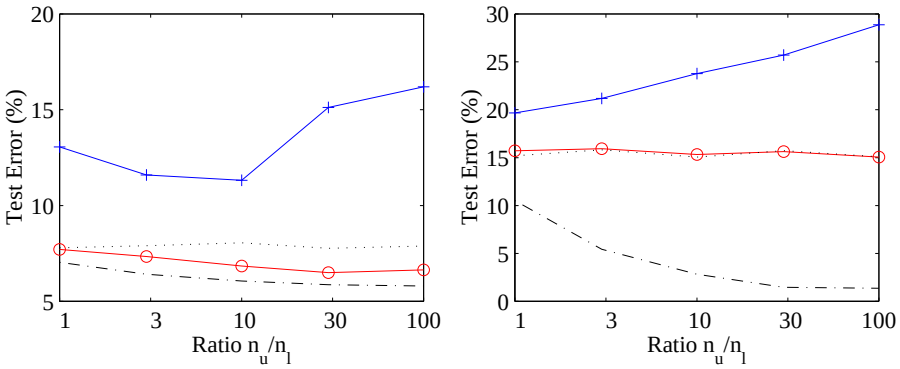

Figure 2: Test error *vs.* n_u/n_l ratio for $a = 0.23$. Average test errors for minimum entropy logistic regression ($\circ$) and mixture models ($+$). The test error rates of logistic regression (dotted), and logistic regression with all labels known (dash-dotted) are shown for reference. Left: experiment with outliers; right: experiment with uninformative unlabeled data.

The last series of experiments illustrate the robustness with respect to the cluster assumption, by testing it on distributions where unlabeled examples are not informative, and where a low density $P(X)$ does not indicate a boundary region. The data is drawn from two Gaussian clusters like in the first series of experiment, but the label is now independent of the clustering: an example $\mathbf{x}$ belongs to class ω_1 if $x_2 > x_1$ and belongs to class ω_2 otherwise: the Bayes decision boundary is 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 1 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. On the other hand, the diagnosis models behave well, and the minimum entropy 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 performance in this situation.

4.1.3 Comparison with manifold transduction

Although our primary goal is to infer a decision function, we also provide comparisons with a transduction algorithm of the "manifold family". We chose the consistency method of Zhou *et al.* (2004) for its simplicity. As suggested by the authors, we set $\alpha = 0.99$ and the scale parameter σ^2 was optimized on test results (Zhou *et al.*, 2004). The results are reported in Table 1. The experiments are limited due to the memory requirements of the consistency method in our naive MATLAB implementation.

Table 1: Error rates (%) of minimum entropy (ME) *vs.* consistency method (CM), for $a = 0.23$, $n_l = 50$, and a) pure Gaussian clusters b) Gaussian clusters corrupted by outliers c) class boundary passing through the Gaussian clusters.

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

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. Furthermore, when classes do not correspond to clusters, the consistency method performs random class assignments. In fact, our setup, which was designed for the comparison of global classifiers, is extremely defavorable to manifold methods, since the data is truly 50-dimensional. In this situation, local methods suffer from the "curse of dimensionality", and many more unlabeled examples would be required to get sensible results. Hence, these results mainly illustrate that manifold learning is not the best choice in semi-supervised learning for truly high dimensional data.

4.2 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 3. 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 dataset 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 %.

Anger Fear Disgust Joy Sadness Surprise Neutral

Figure 3: Examples from the facial expression recognition database.

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. Again, the local methods fail. This 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.

5 Discussion

We propose to tackle the semi-supervised learning problem in the supervised learning framework, by using a minimum entropy regularizer. This regularizer is motivated by theory, which shows that the information content of unlabeled examples is higher for classes with little overlap. Maximum a posteriori estimation enables to incorporate minimum entropy regularization in the learning process of any probabilistic classifier. In this framework, minimum entropy is interpreted as a "usefulness prior" for unlabeled examples, whose strength can be controlled.

Minimizing entropy gradually increases the confidence of the classifier output at unlabeled examples. Our proposal encompasses self-learning as a particular case, where, at the end of the learning process, entropy minimization converges to a solution assigning hard labels to unlabeled data. The transductive large margin classifier is another limiting case: minimizing entropy on the training sample is a means to drive the decision boundary away from these examples.

Both local and global classifiers can be fitted by the minimum entropy criterion. Global classifiers allow to improve over manifold learning when data do not lie on a low-dimensional manifold, or, as illustrated in the expression recognition experiment, when the classification task aims at differentiating examples transformed along some global direction of the manifold. Our experiments also suggest that supervised learning with minimum entropy regularization may be a serious contender to generative models. It compares favorably to 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.

A Detailed derivation of $P(Y|X,Z)$

Bayes' rule and total probability theorem yields:

$$P(Y=\omega_k|\mathbf{x},\mathbf{z}) = \frac{P(\mathbf{z}|\mathbf{x},Y=\omega_k)P(Y=\omega_k|\mathbf{x})}{\sum_{\ell=1}^{K} P(\mathbf{z}|\mathbf{x},Y=\omega_\ell)P(Y=\omega_\ell|\mathbf{x})} \tag{7}$$

From the definition of $\mathbf{z}$, we have $P(\mathbf{z}|Y=\omega_k)=0$ when $z_k=0$, which implies $P(\mathbf{z}|\mathbf{x},Y=\omega_k) = z_k P(\mathbf{z}|\mathbf{x},Y=\omega_k)$.

$$\begin{aligned} P(Y=\omega_k|\mathbf{x},\mathbf{z}) &= \frac{z_k P(\mathbf{z}|\mathbf{x},Y=\omega_k)P(Y=\omega_k|\mathbf{x})}{\sum_{\ell=1}^{K} z_\ell P(\mathbf{z}|\mathbf{x},Y=\omega_\ell)P(Y=\omega_\ell|\mathbf{x})} & (8)\\ &= \frac{z_k P(Y=\omega_k|\mathbf{x})}{\sum_{\ell=1}^{K} z_\ell P(Y=\omega_\ell|\mathbf{x})} \ , & (9) \end{aligned}$$

where the last line is derived from the missing-at-random assumption, $P(\mathbf{z}|\mathbf{x},Y=\omega_k) = P(\mathbf{z}|\mathbf{x},Y=\omega_\ell)$ for all (k,ℓ) such that $z_k=z_\ell=1$.

■

B Proof of theorem 1

Theorem 1

In the two-class linear separable case, the logistic regression model with bounded weights, fitted by the minimum entropy criterion, converges towards the maximum margin separator (with maximal distance from labeled and unlabeled examples) as the bound goes to infinity.

Proof.

Consider the logistic regression model parameterized by $\boldsymbol{\theta}=(\mathbf{w},b)$: $P(Y|\mathbf{x})$ is modeled by $f(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$ for the positive class, and by $1-f(\mathbf{x};\boldsymbol{\theta})$ for the negative class. Let $t_i \in \{-1,+1\}$ be a binary variable defined as follows: if $\mathbf{x_i}$ is a positive labeled example, $t_i=+1$; if $\mathbf{x_i}$ is a negative labeled example, $t_i=-1$; if $\mathbf{x_i}$ is an unlabeled example, $t_i = \mathrm{sign}(f(\mathbf{x}_i)-1/2)$. The *margin* for the labeled or unlabeled example i is defined as $m_i(\boldsymbol{\theta}) = t_i(\mathbf{w}^T\mathbf{x_i}+b)$.

The cost C (6) can then be written as a function of $m_i = m_i(\boldsymbol{\theta})$ as follows

$$C(\boldsymbol{\theta}) = -\sum_{i=1}^{n_l} \log(1+e^{-m_i}) - \lambda \sum_{i=n_l+1}^{n} \log(1+e^{-m_i}) + \frac{m_i e^{-m_i}}{1+e^{-m_i}} \ , \tag{10}$$

where the indices $[1,n_l]$ and $[n_l+1,n]$ correpond to labeled and unlabeled data, respectively. The bounded weight estimate $\widehat{\boldsymbol{\theta}}_B = (\widehat{\mathbf{w}}_B, \widehat{b}_B)$ is obtained by optimizing C under the constraint $\|\mathbf{w}\| \leq B$. In the sequel, $m_i(\widehat{\boldsymbol{\theta}}_B)$ will be denoted $\widehat{m}_i$

We first show that, as B goes to infinity, all margins $\widehat{m}_i$ go to infinity. Let $\boldsymbol{\theta}^* = (\mathbf{w}^*, b^*)$ be the parameters of the maximum margin separator with $\|\mathbf{w}^*\|=1$. Let

m_i^* be $m_i(\boldsymbol{\theta}^*)$. From the definition of $\boldsymbol{\theta}^*$, $m_i^* > 0$, $i = 1, \ldots, n$. Since $m_i(B\boldsymbol{\theta}^*) = Bm_i(\boldsymbol{\theta}^*)$, $\lim_{B\to\infty} m_i(B\boldsymbol{\theta}^*) = \infty$, and $C(B\boldsymbol{\theta}^*)$ goes to zero:

$$\begin{aligned}\lim_{B\to\infty} C(B\boldsymbol{\theta}^*) &= \lim_{B\to\infty} -\sum_{i=1}^{n_l} e^{-Bm_i^*} - \lambda \sum_{i=n_l+1}^{n} Bm_i^* e^{-Bm_i^*} \\ &= 0 \ .\end{aligned}$$

Suppose now that there is at least one example i, such that $\widehat{m}_i \leq M$, where M is a positive constant. Then, C (10) can trivialy be bounded from above: $C(\widehat{\boldsymbol{\theta}}_B) \leq -\log(1 + \exp(-M))$ if i is labeled and $C(\widehat{\boldsymbol{\theta}}_B) \leq -\lambda \log(1 + \exp(-M))$ if i is unlabeled. Since $B\boldsymbol{\theta}^*$ is an admissible solution with $\lim_{B\to\infty} C(B\boldsymbol{\theta}^*) = 0$, $\widehat{\boldsymbol{\theta}}_B$ cannot maximize C if M is finite. We thus conclude that $\lim_{B\to\infty} \widehat{m}_i = \infty$, $i = 1, \ldots, n$.

We now show that $\|\widehat{\mathbf{w}}_B\| = B$. For this, we write the gradient of $C(\alpha\boldsymbol{\theta})$ with respect to α:

$$\begin{aligned}\left.\frac{\partial C(\alpha\boldsymbol{\theta})}{\partial\alpha}\right|_{\alpha=1} &= \sum_{i=1}^{n_l} \frac{e^{-m_i}}{(1+e^{-m_i})} \frac{\partial m_i}{\partial\alpha} + \lambda \sum_{i=n_l+1}^{n} \frac{m_i e^{-m_i}}{(1+e^{-m_i})^2} \frac{\partial m_i}{\partial\alpha} \\ &= \sum_{i=1}^{n_l} \frac{m_i e^{-m_i}}{(1+e^{-m_i})} + \lambda \sum_{i=n_l+1}^{n} \frac{m_i^2 e^{-m_i}}{(1+e^{-m_i})^2} \ ,\end{aligned}$$

As $\lim_{B\to\infty} \widehat{m}_i > 0$, each term in the sum is strictly positive for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_B$, and $\lim_{B\to\infty} \left.\frac{\partial C(\alpha\widehat{\boldsymbol{\theta}}_B)}{\partial\alpha}\right|_{\alpha=1} > 0$. The constraint $\|\mathbf{w}\| \leq B$ is thus active at $\widehat{\boldsymbol{\theta}}_B = (\widehat{\mathbf{w}}_B, \widehat{b}_B)$, hence $\|\widehat{\mathbf{w}}_B\| = B$.

Finally, we derive that logistic regression asymptotically achieves maximum margin separation. Let $m_0^* = \min_{i\in[1,n]} m_i^*$ and $\widehat{m}_0 = \min_{i\in[1,n]} \widehat{m}_i$ denote the minimum margin among all labeled and unlabeled examples, and the minimum margin achieved by logistic regression, respectively. We show below that $\lim_{B\to\infty} \frac{\widehat{m}_0}{\|\widehat{\mathbf{w}}_B\|} = m_0^*$.

Let I_l^* and I_u^* denote the set of indices of respectively labeled and unlabeled examples with minimum margin $I_l^* = \{i \in [1, n_l] | m_i^* = m_0^*\}$ and $I_u^* = \{i \in [n_l + 1, n] | m_i^* = m_0^*\}$. Accordingly, we denote $\widehat{I}_l = \{i \in [1, n_l] | \widehat{m}_i = \widehat{m}_0\}$ and $\widehat{I}_u = \{i \in [n_l + 1, n] | \widehat{m}_i = \widehat{m}_0\}$. Finally, we define $c^* = \min(2, \arg\min_{m_i^* > m_0^*} \frac{m_i^* - m_0^*}{m_0^*})$, and $\widehat{c} = \min(2, \arg\min_{\widehat{m}_i > \widehat{m}_0} \frac{\widehat{m}i - \widehat{m}_0}{\widehat{m}_0})$.

$$\begin{aligned}C(B\boldsymbol{\theta}^*) &= -\left(|I_l^*| - \lambda \sum_{i\in I_u^*} (1 + Bm_i^*)\right) e^{-Bm_0^*} + O(Bm_0^* \, e^{-c^* B m_0^*}) \\ C(\widehat{\boldsymbol{\theta}}_B) &= -\left(|\widehat{I}_l| - \lambda \sum_{i\in \widehat{I}_u} (1 + \widehat{m}_i)\right) e^{-\widehat{m}_0} + O(\widehat{m}_0 \, e^{-\widehat{c}\widehat{m}_0})\end{aligned}$$

where $|I|$ denote the cardinal number of set I.

We now note that for any $\varepsilon < 1$, $\lim_{B\to\infty} Bm_0^* e^{-(c^*-\varepsilon)Bm_0^*} = 0$, hence

$$\begin{aligned}\lim_{B\to\infty} e^{\varepsilon Bm_0^*} C(B\boldsymbol{\theta}^*) &= \lim_{B\to\infty} -\left(|I_l^*| - \lambda \sum_{i\in I_u^*}(1+Bm_i^*)\right) e^{-(1-\varepsilon)Bm_0^*} \\ &= 0\end{aligned}$$

The optimality of $\widehat{\boldsymbol{\theta}}_B$ entails $C(\widehat{\boldsymbol{\theta}}_B) \geq C(B\boldsymbol{\theta}^*)$, which implies that for any $\varepsilon < 1$,

$$\begin{aligned}\lim_{B\to\infty} e^{\varepsilon Bm_0^*} C(\widehat{\boldsymbol{\theta}}_B) &= 0 \\ \lim_{B\to\infty} -\left(|\widehat{I}_l| - \lambda \sum_{i\in\widehat{I}_u}(1+\widehat{m}_i)\right) e^{\varepsilon Bm_0^* - \widehat{m}_0} &= 0\end{aligned}$$

Hence, for any $\varepsilon < 1$, $\lim_{B\to\infty} B(\frac{\widehat{m}_0}{B} - \varepsilon m_0^*) = \infty$. As by definition $\frac{\widehat{m}_0}{B} \leq m_0^*$, we conclude that $\lim_{B\to\infty} \frac{\widehat{m}_0}{B} = \lim_{B\to\infty} \frac{\widehat{m}_0}{\|\widehat{\mathbf{w}}_B\|} = m_0^*$.

■

Note that besides the linear separable case, this theorem can be easily extended to kernelized logistic regression using kernels ensuring linear separability (such as the Gaussian kernel).

References

ABBOUD B., DAVOINE F. & MO D. (2003). Expressive face recognition and synthesis. In *Computer Vision and Pattern Recognition Workshop*, volume 5, p.5̃4.

AMINI M. R. & GALLINARI P. (2002). Semi-supervised logistic regression. In *15th European Conference on Artificial Intelligence*, p. 390–394: IOS Press.

BENNETT K. P. & DEMIRIZ A. (1999). Semi-supervised support vector machines. In M. S. KEARNS, S. A. SOLLA & D. A. COHN, Eds., *Advances in Neural Information Processing Systems 11*, p. 368–374: MIT Press.

BERGER J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, 2 edition.

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

CASTELLI V. & COVER T. M. (1996). 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.

CORDUNEANU A. & JAAKKOLA T. (2003). On information regularization. In *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence (UAI)*.

FRIEDMAN J., HASTIE T. & TIBSHIRANI R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, **28**(2), 337–407.

GRANDVALET Y. (2002). Logistic regression for partial labels. In *9th Information Processing and Management of Uncertainty in Knowledge-based Systems – IPMU'02*, p. 1935–1941.

JIN R. & GHAHRAMANI Z. (2003). Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*: MIT Press.

JOACHIMS T. (1999). Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, p. 200–209.

KANADE T., COHN J. & TIAN Y. (2000). Comprehensive database for facial expression analysis. In *4th IEEE International Conference on Automatic Face and Gesture Recognition*.

MCLACHLAN G. J. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley.

NIGAM K. & GHANI R. (2000). Analyzing the effectiveness and applicability of co-training. In *Ninth International Conference on Information and Knowledge Management*, p. 86–93.

NIGAM K., MCCALLUM A. K., THRUN S. & MITCHELL T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, **39**(2/3), 103–134.

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

ROSE K., GUREWITZ E. & FOX G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, **11**(9), 589–594.

SEEGER M. (2002). *Learning with labeled and unlabeled data*. Rapport interne, Institute for Adaptive and Neural Computation, University of Edinburgh.

SZUMMER M. & JAAKKOLA T. S. (2003). Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems 15*: MIT Press.

TONG S. & KOLLER D. (2000). Restricted bayes optimal classifiers. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, p. 658–664.

VAPNIK V. N. (1998). *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. New York: Wiley.

ZHOU D., BOUSQUET O., NAVIN LAL T., WESTON J. & SCHÖLKOPF B. (2004). Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*.

ZHU X., GHAHRAMANI Z. & LAFFERTY J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *20th Int. Conf. on Machine Learning*, p. 912–919.