Kernel expansions with unlabeled examples

Martin Szummer
MIT AI Lab & CBCL
Cambridge, MA

Tommi Jaakkola
MIT AI Lab
Cambridge, MA

Abstract

Modern classification applications necessitate supplementing the few available labeled examples with unlabeled examples to improve classification performance. We present a new tractable algorithm for exploiting unlabeled examples in discriminative classification. This is achieved essentially by expanding the input vectors into longer feature vectors via the unlabeled examples. The resulting classification method can be interpreted as a discriminative kernel density estimate and is readily trained via the EM algorithm, which in this case is both discriminative and achieves the optimal solution. We provide, in addition, a purely discriminative formulation of the estimation problem by appealing to the maximum entropy framework. We demonstrate that the proposed approach requires very few labeled examples for high classification accuracy.

1 Introduction

In many modern classification problems such as text categorization, very few labeled examples are available but a large number of unlabeled examples can be readily acquired. Various methods have recently been proposed to take advantage of unlabeled examples to improve classification performance. This includes the use of EM in combination with Naive Bayes models for text classification [5], the co-training framework[1], transduction [7, 4], and maximum entropy discrimination [3].

These approaches are divided primarily on the basis of whether they employ generative modeling or are motivated by robust classification. Unfortunately, the computational effort scales exponentially with the number of unlabeled examples for exact solutions in discriminative approaches such as transduction [7, 3]. Various approximations are available [4, 3] but their effect remains unclear.

In this paper, we formulate an alternative discriminative approach to exploiting unlabeled examples, effectively by using them to expand the representation of examples. This approach has several advantages including the ability to represent the

true Bayes optimal decision boundary and making explicit use of the density over the examples. It is also computationally feasible as stated.

The paper is organized as follows. We start by discussing the kernel density estimate and providing a smoothness condition. We subsequently define the expansion and formulate the EM algorithm for discriminative training. In addition, we provide a purely discriminative version of the parameter estimation problem and formalize it as a maximum entropy discrimination problem. We then demonstrate experimentally that various concerns about the approach are not warranted.

2 Kernel density estimation and classification

We start by assuming a large number of *labeled* examples $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $y_i \in \{-1, 1\}$ and $\mathbf{x}_i \in \mathcal{R}^d$. A joint kernel density estimate can be written as

$$P(\mathbf{x}, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(y, y_i)\, K(\mathbf{x}, \mathbf{x}_i) \tag{1}$$

where $\int K(\mathbf{x}, \mathbf{x}_i) d\mu(\mathbf{x}) = 1$ for each i. With an appropriately chosen kernel K, a function of n, $P(\mathbf{x}, y)$ will be *consistent* in the sense of converging to the joint density as $n \to \infty$.

Given a fixed number of examples the kernel functions $K(\mathbf{x}, \mathbf{x}_i)$ may be viewed as conditional probabilities $P(\mathbf{x}|i)$, where i indexes the observed points. For the purposes of this paper, we assume a Gaussian form $K(\mathbf{x}, \mathbf{x}_i) = N(\mathbf{x}; \mathbf{x}_i, \sigma^2 I)$. The labels y_i assigned to the sampled points $\mathbf{x}_i$ may themselves be noisy and we incorporate $P(y|i)$, a location specific probability of labels. The resulting joint density model is

$$P(\mathbf{x}, y) = \frac{1}{n} \sum_{i=1}^{n} P(y|i)\, P(\mathbf{x}|i) \tag{2}$$

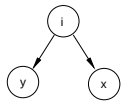

Figure 1:

Interpreting $1/n$ as a prior probability of the index variable $i = 1, \ldots, n$, the resulting model conforms to the graph depicted in Figure 1. This is reminiscent of the *aspect model* for clustering of dyadic data[8]. There are two main differences. First, the number of aspects here equals the number of examples and the model is not suitable for clustering. Second, we do not search for the probabilities $P(\mathbf{x}|i)$ (kernels), they are associated with each observed example and are merely adjusted in terms of scale (kernel width). This restriction carries a significant computational advantage in classification, which is the objective in this paper.

The posterior probability of the label y given an example $\mathbf{x}$ is given by $P(y|\mathbf{x}) = \sum_i P(y|i)P(i|\mathbf{x})$, where $P(i|\mathbf{x}) \propto P(\mathbf{x}|i)$ as $P(i)$ is assumed to be uniform. The quality of the posterior probability depends both on how accurately $P(y|i)$ are known as well as on the properties of the membership probabilities $P(i|\mathbf{x})$ (always known) that must be relatively smooth.

Here we provide a simple condition on the membership probabilities $P(i|\mathbf{x})$ so that any noise in the sampled labels for the available examples would not preclude accu-

rate decisions. Note, for example, that $P(y|\mathbf{x}_j) \neq \delta(y, y_j)$, i.e., the predicted label for $\mathbf{x}_j$ does not (and should not) necessarily equal the given noisy label.

The following simple lemma is readily obtained, e.g., via McDiarmid's theorem.

Lemma 1 *Let $I_n = \{1, \ldots, n\}$. Given a collection of distributions $p_{i|k} \geq 0$, $\sum_{i \in I_n} p_{i|k} = 1$ for $k \in I_n$, where $\|p_{\cdot|k}\|_2 \leq \epsilon/\sqrt{2\log(2n/\delta)}, \forall k \in I_n$, and independent samples $y_i \in \{-1, 1\}$ from some $P(y_i|i), i \in I_n$, then*

$$P(\exists k \in I_n : |\sum_{i=1}^{n} y_i p_{i|k} - \sum_{i=1}^{n} w_i p_{i|k}| > \epsilon) \leq \delta \tag{3}$$

where $w_i = P(y_i = 1|i) - P(y_i = -1|i)$ and the probability is taken over the independent samples.

The lemma applies to our case by setting $p_{i|k} = P(i|\mathbf{x}_k)$, $\{y_i\}$ represents the sampled labels for the examples, and by noting that the sign of $\sum w_i P(i|\mathbf{x})$ is the MAP decision rule from our model, $P(y = 1|\mathbf{x}) - P(y = -1|\mathbf{x})$. The lemma states that so long as the membership probabilities have appropriately bounded squared norm, the noise in the labeling is inconsequential as far as the classification decisions are concerned. Note, for example, that for $p_i = 1/n$, $\|p\|_2 = 1/\sqrt{n}$ implying that the conditions are achievable for large n. The squared norm of $P(i|\mathbf{x})$ is directly controlled by the kernel width σ^2. Algorithms for adjusting the kernel width(s) on the basis of this are given in the longer version of this paper.

3 Expansion and the EM estimation

A useful way to view the resulting kernel density estimate is that each example $\mathbf{x}$ is represented by a vector of membership probabilities $P(i|\mathbf{x})$, $i = 1, \ldots, n$. Such *mixture distance* representations have been used extensively; it can also be viewed as a Fisher score vector computed with respect to adjustable weighting $P(i)$. The examples in this new representation are classified by associating $P(y|i)$ with each component and computing $P(y|\mathbf{x}) = \sum_i P(y|i)P(i|\mathbf{x})$. An alternative approach to exploiting kernel density estimates in classification is given by [6].

We now dispense with the assumption that we have labels for all but few of the available examples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. In this case, we may continue to use the model defined above and estimate the free parameters, $P(y|i)$,$i = 1, \ldots, n$, from the few labeled examples. In other words, we can maximize the conditional log-likelihood

$$\sum_{t=1}^{T} \log P(y_t|\mathbf{x}_t) = \sum_{t=1}^{T} \log \sum_{i=1}^{n} P(y_t|i)P(i|\mathbf{x}_t) \tag{4}$$

where the first summation is over the labeled examples and $T \ll n$. Since $P(i|\mathbf{x}_t)$ are fixed, this objective function is jointly concave in the free parameters and lends itself to a unique maximum value. The concavity also guarantees that this optimization is easily performed via the EM algorithm:

Let $p_{i|t}$ be the soft assignment for aspect i given $(\mathbf{x}_t, y_t)$, i.e., $p_{i|t} = P(i|\mathbf{x}_t, y_t) \propto P(y_t|i)P(i|\mathbf{x}_t)$. The EM algorithm iterates between the E-step, where $p_{i|t}$ are recomputed from the current estimates of $P(y|i)$, and the M-step where we update $P(y|i) \leftarrow \sum_{t:y_t=y} p_{i|t} / \sum_t p_{i|t}$.

This procedure may have to be adjusted in cases where the overall frequency of different labels in the (labeled) training set deviates significantly from uniform. A simple rescaling $P(y|i) \leftarrow P(y|i)/T_y$ by the frequencies T_y and renormalization after each M-step would probably suffice.

The runtime of this algorithm is $\mathcal{O}(T\,n)$. The discriminative formulation suggests that EM will provide reasonable parameter estimates $P(y|i)$ for classification purposes. The quality of the solution, as well as the potential for overfitting, is contingent on the smoothness of the kernels or, equivalently, smoothness of the membership probabilities $P(i|\mathbf{x})$. Note, however, that whether or not $P(y|i)$ will converge to the extreme values 0 or 1 is not an indication of overfitting.

Exploiting unlabeled examples in the manner described above has the advantage that we can represent the Bayes optimal decisions provided that the kernel density estimate is sufficiently accurate. Note that such decisions for the unlabeled examples included in the expansion need to be made on the basis of $P(y|\mathbf{x}_i)$ and not on the basis of $P(y|i)$, which function as parameters.

4 Alternative discriminative formulation

An alternative discriminative formulation is also possible, one that is more sensitive to the decision boundary rather than probability values associated with the labels. To this end, consider the posterior probability $P(y|\mathbf{x}) = \sum_i P(y|i)P(i|\mathbf{x})$. The decisions are made on the basis of the sign of the discriminant function

$$\mathcal{L}(\mathbf{x}) = P(y=1|\mathbf{x}) - P(y=-1|\mathbf{x}) = \sum_{i=1}^{n} w_i P(i|\mathbf{x}) \tag{5}$$

where $w_i = P(y=1|i) - P(y=-1|i)$. This is a linear classifier and there are many ways of estimating the weights w_i discriminatively. The weights should remain bounded, however, i.e, $w_i \in [-1,1]$, so long as we wish to maintain the kernel density interpretation. Estimation algorithms with Euclidean norm regularization such as SVMs would not be appropriate in this sense. We rely here on the relation $w_i = E\{y|i\} = \sum_y yP(y|i)$ and estimate the distributions $\{P(y|i)\}$ within the maximum entropy discrimination (MED) framework [3].

The maximum entropy formalism encodes the principle that our choice of the labels for the unlabeled examples should remain uninformative to the extent possible given the classification objective. More formally, given a set of T training examples and labels $(y_1, \mathbf{x}_1), \ldots, (y_T, \mathbf{x}_T)$, we minimize $-\sum_i H(P_{y|i}) + C\sum_t \epsilon_t$ subject to the classification constraints

$$y_t \left[\sum_i \sum_y y\, P(y|i) P(i|\mathbf{x}_t) \right] + \epsilon_t \geq \epsilon \tag{6}$$

Here ϵ specifies the target separation ($\epsilon \in [0,1]$) and the slack variables $\epsilon_t \geq 0$ permit deviations from the target to ensure that a solution always exists. The advantage of this formulation is that effort is spent only on those training examples whose classification is uncertain. Examples already classified correctly with a margin larger than ϵ are effectively ignored. The optimization problem and algorithms are explained in the appendix.

5 Other expansions

If we give up the probabilistic interpretation of the kernels K as density estimators, we are free to choose them from a more general class of functions. For example, the kernels no longer have to integrate to 1. An expansion of $\mathbf{x}$ in terms of these kernels can still be meaningful; as a special case, when linear kernels are chosen, the expansion reduces to weighting distances between points by the covariance of the data. Distinctions along high variance directions then become easier to make, which is helpful when between-class scatter is greater than within-class scatter.

Thus, even though the probabilistic interpretation is missing, a simple preprocessing step can help support vector machines to take advantage of unlabeled data. Simply expand the inputs $\mathbf{x}$ in terms of kernels G from unlabeled points: $\phi(\mathbf{x}) = [G(\mathbf{x}, \mathbf{x}_1), \ldots, G(\mathbf{x}, \mathbf{x}_n)]$.

6 Results

An obvious concern with the proposed approach is that it may be too flexible resulting in a poor generalization. Figure 2a) demonstrates that this is not the case and, instead, the test classification error approaches the limiting asymptotic rate exponentially fast. The problem considered was a DNA splice site classification problem with 500 examples for which $d = 100$. Varying sizes of random subsets were labeled and all the examples were used in the expansion as unlabeled examples. The error rate was computed on the basis of the remaining $500 - n$ examples without labels, where n denotes the number of labeled examples. The results in the figure were averaged across 20 independent runs. The exponential rate of convergence towards the limiting rate is evidenced by the linear trend in the semilog figure 2a). The mean test errors shown in figure 2b) indicate that the purely discriminative training (MED) can contribute substantially to the accuracy. The kernel width in these experiments was simply fixed to a reasonable a priori value given the scaling of the examples and not optimized for the task.

Another concern is perhaps that the formulation is valid only in cases where we have a large number of unlabeled examples. In principle, the method could deteriorate rapidly after the kernel density estimate no longer can be assumed to give reasonable estimates. Figure 3a) illustrates that this is not a valid interpretation. The problem here is to classify high density DNA array experiments on the basis of the leukemia types that the tissues used in the array experiments corresponded to. Each input vector for the classifier consists of the expression levels of over 7000 genes that were included as probes in the arrays. The number of experiments or examples available was 38 for training and 34 for testing. We included all the examples as unlabeled examples in the expansion and randomly selected labels for the training examples, measuring the performance only on the test examples (which were of slightly different type and hence more appropriate for assessing generalization). Figure 3a) shows rapid convergence for EM and the discriminative MED formulation. The "asymptotic" level here corresponds to about one classification error among the 34 test examples. The results were averaged over 20 independent runs. Figure 3b) compares decision boundaries for EM with unlabeled data versus SVM using only labeled data, on a simple problem with only two labeled points but many unlabeled points.

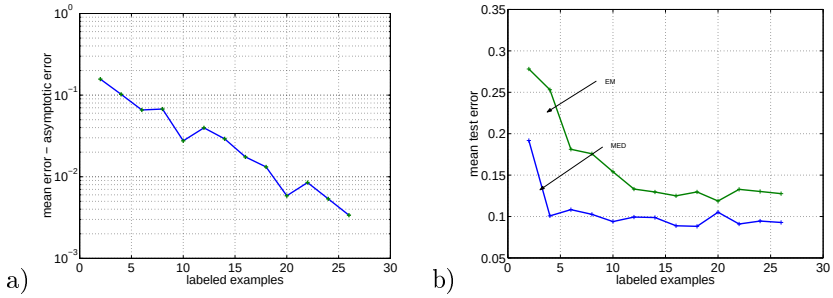

Figure 2: a) A semilog plot of the test error rate for the EM formulation less the asymptotic rate as a function of labeled examples. The linear trend in the figure implies that the error rate approaches the asymptotic error exponentially fast. b) The mean test errors for EM (green line) and MED (blue line) as a function of the number of labeled examples.

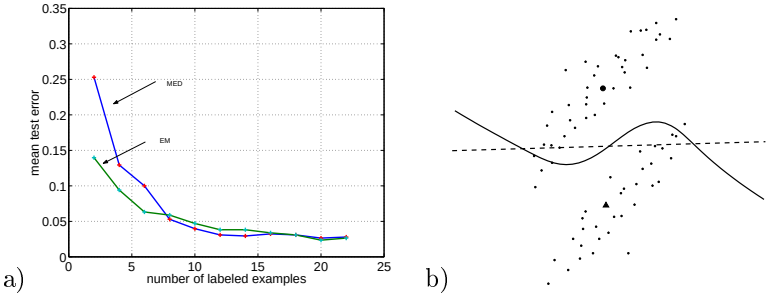

Figure 3: a) The mean test errors for the leukemia classification problem as a function of the number of randomly chosen labeled examples. Results are given for both EM (green line) and MED (blue line) formulations. b) Decision boundary of EM (solid) and labeled-data only SVM (dashed) using the same Gaussian kernel, labeled points (big circle, triangle) and unlabeled points (small dots).

7 Discussion

We have provided an alternative framework for exploiting unlabeled examples in discriminative classification problems. The framework involves a combination of the ideas of kernel density estimation and representational expansion of input vectors. A simple EM algorithm is sufficient for finding globally optimal parameter estimates. In addition, we have shown that a purely discriminative formulation can yield substantially better results within the framework. We plan to compare our approach to discriminative methods based on the labeled examples alone.

There are a number of possible extensions. These include formalizing the selection of the kernel width parameter (preliminary analysis provided in this paper), analysis of the framework similarly to [2], and exploring alternative expansions not based on the kernel density interpretation.

References

[1] Blum A., Mitchell T. (1998). Combining Labeled and Unlabeled Data with Co-Training . In *Proc. 11th Annual Conf. Computational Learning Theory.*

[2] Castelli V., Cover T. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on information theory* **42** (6): 2102–2117.

[3] Jaakkola T., Meila M., and Jebara T. (1999). Maximum entropy discrimination. In *Advances in Neural Information Processing Systems 12.*

[4] Joachims, T. (1999). Transductive inference for text classification using support vector machines. *International conference on Machine Learning.*

[5] Nigam K., McCallum A., Thrun S., and Mitchell T. (1999). Text classification from labeled and unlabeled examples. To appear in *Machine Learning.*

[6] Tong S., Koller D. (1999). Bayes optimal hyperplanes – maximal margin hyperplans.

[7] Vapnik V. (1998). *Statistical learning theory.* John Wiley & Sons.

[8] Hofmann T., Puzicha J. (1998) Unsupervised Learning from Dyadic Data. International Computer Science Institute, TR-98-042.

A Maximum entropy solution

The unique solution to the maximum entropy estimation problem is found via introducing Lagrange multipliers $\{\lambda_t\}$ for the classification constraints. $0 \leq \lambda_t \leq C$, where the lower bound comes from the inequality constraints and the upper bound from the linear margin penalties being minimized. To represent the solution and find the optimal setting of λ_t we must evaluate the partition function

$$Z(\lambda) = e^{-\sum_t \lambda_t \epsilon} \sum_{y_1,\ldots,y_n} \prod_{i=1}^{n} e^{\sum_t \hat{y}_t \lambda_t y_i P(i|\mathbf{x})} = e^{-\sum_t \lambda_t \epsilon} \prod_{i=1}^{n} \left(e^{\sum_t \hat{y}_t \lambda_t P(i|\mathbf{x})} + e^{-\sum_t \hat{y}_t \lambda_t P(i|\mathbf{x})} \right) \quad (7)$$

that normalizes the maximum entropy distribution. Here $\hat{y}$ denote the observed labels. Minimizing the jointly convex log-partition function $\log Z(\lambda)$ with respect to the Lagrange multipliers leads to the optimal setting $\{\lambda_t^*\}$. This optimization is readily done via an axis parallel line search. The required gradients are given by

$$\frac{\partial \log Z(\lambda)}{\partial \lambda_k} = -\epsilon + \sum_i \tanh\left(\sum_t \hat{y}_t \lambda_t^* P(i|\mathbf{x}_t) \right) \hat{y}_k P(i|\mathbf{x}_k) = -\epsilon + \hat{y}_k \sum_i E_{P^*}\{ y_i \} P(i|\mathbf{x}_k) \quad (8)$$

(this is essentially the classification constraint). The expectation is taken with respect to the maximum entropy distribution $P^*(y_1, \ldots, y_n) = P^*(y_1|1) \cdots P^*(y_n|n)$ where the components are $P^*(y|i) \propto \exp\{\sum_t \hat{y}_t \lambda_t y_i P(i|\mathbf{x})\}$ The label averages $w_i^* = E_{P^*}\{ y_i \} = \sum_y y\, P^*(y|i)$ are needed for the decision rule as well as in the optimization. We can identify these from above $w_i^* = \tanh(\sum_t \hat{y}_t \lambda_t^* P(i|\mathbf{x}_t))$ and they are readily evaluated. Finding the solution involves $\mathcal{O}(T^2\, n)$ operations.