Semi-Supervised Logistic Regression

Massih-Reza Amini and **Patrick Gallinari**[1]

Abstract. Semi-supervised learning has recently emerged as a new paradigm in the machine learning community. It aims at exploiting simultaneously labeled and unlabeled data for classification. We introduce here a new semi-supervised algorithm. Its originality is that it relies on a discriminative approach to semi-supervised learning rather than a generative approach, as it is usually the case. We present in details this algorithm for a logistic classifier and show that it can be interpreted as an instance of the Classification Expectation Maximization algorithm.

We also provide empirical results on two data sets for sentence classification tasks and analyze the behavior of our methods.

Keywords. Machine learning, Semi-supervised learning.

1 INTRODUCTION

In the classical supervised learning classification framework, a decision rule is to be build from a labeled data set $D_l = \{(x_i, t_i) \mid i=1,\ldots,n\}$ where each example is described by a pattern x_i and by the response of a supervisor $t_i=(t_{1i},\ldots,t_{Ci})$. In statistical machine learning, data are supposed to be drawn independently from a joint distribution $p(x, t)$ and the learned decision rule is supposed to capture the relation between these two variables.

In practice, labeling large amounts of data may sometimes require considerable human resources or expertise. This is for example the case for many information retrieval tasks where the relevance of retrieved information has to be evaluated by a human. For this type of application, although data are usually widely available, the development of labeled datasets is a long and resource consuming process. For other applications like medical diagnosis, labeling datasets may require expensive tests and be therefore very costly. For rapidly evolving domains or databases there is simply no time to process by hand large datasets.

The semi-supervised learning paradigm has been proposed as a solution to this type of problem when large corpora of unlabeled data are available together with a small amount of labeled data. Labeling a few data is usually affordable and does not take much time. Since there is a belief that unlabeled data contain relevant information about the class, it is a natural idea to try to extract this information so as to provide a classifier more evidence. Classification methods either rely on the estimation of class conditional densities and make use of Bayes rule to take a decision, or directly attempt to estimate the posterior class probabilities. The former are called generative methods and the latter discriminant methods. In the same way, semi-supervised techniques are classified as generative and discriminant. The former usually start from an unsupervised paradigm (e.g. density estimation) and extend it so as to incorporate labeled data, the latter attempt to extend supervised techniques (e.g. linear classifiers) to cope with additional unlabeled data. In this paper, we introduce a new semi-supervised algorithm. Its originality is that it relies on a discriminative approach to semi-supervised learning rather than a generative approach, as it is usually the case. The advantage is that the algorithm is generic - it can be used with many different discriminant classifiers - it also leads to cheap and efficient implementations. We describe here the algorithm for the case of a logistic classifier and show how it can be interpreted as an instance of the Classification Expectation Maximization algorithm (CEM). This general framework provides insights into the proposed method and allow for a simple convergence proof of the algorithm.

The paper is organized as follows, we first make a brief review of recent work in machine learning for semi-supervised techniques (section 2). We then present the formal framework of our model and its interpretation as a CEM instance (section 3). Finally we present a series of experiments on Reuters news-wire and on the Computation and Language (cmp_lg) of TIPSTER SUMMAC collections (section 4), and carry on a set of comparisons using different strategies.

2 PREVIOUS WORK

2.1 Generative approaches

First attempts to consider simultaneously labeled and unlabeled data came from the statistician and the pattern recognition community. A review of the work prior to 92 in the context of discriminant analysis may be found in [9]. Most approaches are generative, they start from a mixture density model where mixture components are identified to classes and attempt at maximizing the joint likelihood of labeled and unlabeled data. Since direct optimization is usually unfeasible, the EM algorithm is used to perform maximum likelihood estimation. Usually, for continuous variables, density components are assumed to be gaussian especially when deriving asymptotic analysis. Practical algorithms may be used for more general settings, as soon as the different statistics needed for EM may be estimated, e.g. for discrete variables, non parametric techniques (e.g. histograms) are often used in practice. The semi-supervised paradigm has been recently rediscovered by the machine learning community and many papers now deal with this subject. Most papers propose mixture density models similar to the one described above.

[10] consider a mixture of experts i.e. several mixture component may be associated to one class, when it is usually assumed that there is a one to one correspondence between classes and components. They then propose different models and EM implementation. [11] propose an algorithm which is a particular case of the general semi-supervised EM described in [9], and present an empirical evaluation for text classification, they also extend their model to multiple components per class. [12] propose a Kernel Discriminant Analysis which can be used for semi-supervised classification.

[1] Computer Science Laboratory of Paris 6 (LIP6), University of Pierre et Marie Curie, 8 rue du capitaine Scott, 75015 Paris, France, email: {amini, gallinari}@poleia.lip6.fr

2.2 *Discriminant approaches*

Anderson [1] suggests to modify logistic regression, a well known classifier to incorporate unlabeled data. To do so, he maximizes the joint likelihood of labeled and unlabeled data.

The co-training paradigm [2] which has been proposed independently is also related to semi supervised training. In this approach it is supposed that data x may be described by two modalities which are conditionally independent given the class of x. Two classifiers are used, one for each modality, each classifier operates alternatively as teacher and learner.

[3] present an interesting extension of a boosting algorithm which incorporate co-training. The work of [4] also bears similarities with this technique.

3 ALGORITHMS

We now introduce an iterative discriminant algorithm for semi-supervised learning. This algorithm is generic in the sense that it can be used with any discriminant classifier. For simplifying the presentation, we will consider only two-class classification and detail the algorithm for the case of a simple logistic classifier. This is not restrictive since the algorithm and analysis can be easily extended for any discriminant classifier and for multi-class problems. We briefly present logistic regression in 3.2.

We then proceed to describe our algorithm under the general framework of the Classification maximum likelihood (CML) approach [5, 9]. For this we first introduce CML and the Classification EM algorithm in 3.3, we then show how this framework can be adapted to handle labeled-unlabeled data (3.4) and leads to natural discriminant formulation (3.5). Casting our method in this framework ensures that all properties of CEM (e.g. convergence) hold for our method. This algorithm is detailed in the particular case of logistic regression in section (3.5).

3.1 Framework

We consider a binary decision problem where there are available a set of labeled data D_l and a set of unlabeled data D_u. We will denote, $D_l=\{(x_i, t_i) \mid i=1,\dots,n\}$ where $x_i \in \mathrm{R}^d$, $t_i=(t_{1i},t_{2i})$ is the class indicator vector for x_i and $D_u=\{x_i \mid i= n+1,\dots,n+m\}$. The latter are assumed to have been drawn from a mixture of densities with two components C_1, C_2 in some unknown proportions π_1,π_2. We will consider that unlabeled data have an associated missing indicator vector $t_i = (t_{1i},t_{2i})$, $(i=n+1, \dots, n+m)$ which is a class or cluster indicator vector. The algorithms we consider attempt to iteratively partition the data into the two components C_1 and C_2. We will denote $(P_1^{(j)}, P_2^{(j)})$ the partition into two clusters computed by an algorithm at iteration j.

3.2 Logistic regression

Logistic regression is a well known technique for classification. [1, 6]. The only distributional assumption with this method is that the log likelihood ratio of class distributions is linear in the observations (1), this assumption is verified by a large range of exponential density families, e.g. normal, beta, gamma, etc.

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = \beta_0 + \beta^t.x \qquad (1)$$

Where the f_k, $k=\{1,2\}$ are class conditional parametric densities and $\beta = \{\beta\}_{k=0}^{d}$ is the set of parameters of the model. An advantage of such a model is that it gives the posterior probabilities a simple form:

$$p(P_1 / x) = \frac{1}{1+\exp(-(\beta_0 + \beta^t.x))} \text{ and } p(P_2 / x) = 1 - p(P_1 / x) \qquad (2)$$

The βs are trained to optimize the following log-likelihood [1]:

$$L(P,\beta) = \sum_{k=1}^{2} \sum_{x_i \in P_k} \log(p(P_k / x_i;\beta)) \qquad (3)$$

Where, $\sum_{x_i \in P_k}$ is a summation over all examples x_i in the partition P_k. Criterion (3) is a convex function of the model parameters (1). The latter are estimated in order to maximize (3), gradient techniques are generally used to this end.

This model could be implemented using a simple logistic unit G whose parameters are (β_0, β), i.e. $G(x) = \frac{1}{1+\exp(-(\beta_0+\beta^t.x))}$. After the estimation of β, $G(x)$ and 1 - G(x) are used to estimate $p(P_1/x)$ and $p(P_2/x)$.

3.3 Classification Maximum Likelihood approach

Let us now introduce the classification maximum likelihood (CML) approach to clustering [14]. In this unsupervised approach there are m samples generated via a mixture density:

$$f(x,\Theta) = \sum_{k=1}^{2} \pi_k.f_k(x,\theta_k) \qquad (4)$$

Where the f_k are parametric densities with unknown parameters θ_k and π_k the mixture proportions. The goal here is to cluster the samples into *2* components P_1, P_2. Under the mixture sampling scheme, samples x_i are taken from the mixture density f, and the CML criterion is [5]:

$$\log L_{CML}(P,\pi,\theta) = \sum_{k=1}^{2} \sum_{i=1}^{m} t_{ki} \log\left\{\pi_k.f_k(x_i,\theta_k)\right\} \qquad (5)$$

Note that this is different from the classical mixture maximum likelihood (MML) approach that has been used for most semi-supervised algorithms, where the log joint likelihood reads:

$$\log L_{MML}(P,\pi,\theta) = \sum_{i=1}^{m} \log\left\{\sum_{k=1}^{2} \pi_k.f_k(x_i,\theta_k)\right\} \qquad (6)$$

For MML the goal is to model the data distribution, whereas in the CML approach, we want to cluster data. For CML the mixture indicator t_{ki} for a given data x_i is treated as an unknown parameter of the model and has to be estimated together with the θs. Eq. (5) corresponds to the complete data likelihood of the variables (x,t), the t indicators correspond to a hard decision on the mixture component identity. Many clustering algorithms can be considered as particular cases of CML [5]. The classification EM algorithm (CEM) [5] is an iterative technique, which has been proposed for maximizing (5), it is similar to the classical EM except for an additional ***C***-step where each x_i is assigned to one and only one component of the mixture. The algorithm is sketched below.

CEM

Initialization: start from an initial partition $P^{(0)}$

j^{th} iteration, $j \geq 0$:

E –step. Estimate the posterior probability that x_i belongs to P_k ($i=1,\dots, m$; $k=1,2$):

$$E[t_{ki}^{(j)} / x_i; P^{(j)}, \pi^{(j)}, \theta^{(j)}] = \frac{\pi_k^{(j)} . f_k(x_i; \theta_k^{(j)})}{\sum_{k=1}^{c} \pi_k^{(j)} . f_k(x_i; \theta_k^{(j)})} \qquad (7)$$

C – step. Assign each x_i to the cluster $P_k^{(j+1)}$ with maximal posterior probability according to $E[t/x]$.

M–step. Estimate the new parameters ($\pi^{(j+1)}$, $\theta^{(j+1)}$) which maximize log $L_{CML}(P^{(j+1)}, \pi^{(j)}, \theta^{(j)})$.

3.4 Semi-supervised generative-CEM

CML has been proposed for unlabeled data, but it can be easily modified to handle both *labeled* and *unlabeled* data [9], the only difference is that the t_{ki} for labeled data are known, so the log likelihood criterion becomes:

$$\log L_C = \sum_{k=1}^{2} \left\{ \sum_{x_i \in P_k} \log\{\pi_k . f_k(x_i, \theta_k)\} + \sum_{i=n+1}^{n+m} t_{ki} \log\{\pi_k . f_k(x_i, \theta_k)\} \right\} \quad (8)$$

In this expression, the first summation inside the brackets is over the labeled samples, and the second one over unlabeled samples.

CEM can then be easily adapted to the case of semi supervised learning for maximizing (8) instead of (5); the t_{ki} for the unlabeled data are estimated as in the classical CEM (E and C steps) and are kept fixed for the labeled data. In this generative approach, any density estimation technique can be used for the f_k. In our experiments, we have been using normal distributions. Once the densities have been estimated, classification decision on unknown data can be made according to the Bayes decision rule.

3.5 Generalizing logistic regression to unlabeled data : Semi-supervised discriminant-CEM

The above generative approach indirectly computes posteriors $p(P_k/x)$ via conditional density estimation. This could lead to poor estimates for high dimensions or when only few data are labeled which is usually the case for semi-supervised learning.

A more straightforward approach would be to use a discriminant model to directly estimate the posteriors without spending resources on the more complex task of density estimation. In this section, we first rewrite the semi-supervised CML criterion (8) in a suitable form which puts in evidence the role of posterior probabilities.

We then show how it is possible to maximize this likelihood with discriminant classifiers. This leads to a modified CEM algorithm. For simplification, we detail this algorithm for a logistic model.

Using Bayes rule, CML criterion (8) can be rewritten:

$$\log L_C = \log \widetilde{L}_C(P, \beta) + \sum_{i=1}^{n+m} \log p(x_i) \qquad (9)$$

Where:

$$\log \widetilde{L}_C(P, \beta) = \sum_{k=1}^{2} \left\{ \sum_{x_i \in P_k} \log\{p(P_k / x_i)\} + \sum_{i=n+1}^{n+m} t_{ki} \log\{p(P_k / x_i)\} \right\} (10)$$

$p(x)$ is the marginal distribution of data and β denotes the "parameters" of the classifiers, e.g. the weights of a linear classifier, the β coefficients of the logistic classifier (1), the weights of a neural network, the local decision functions of a classification tree, etc. (9) is the classification likelihood of the β. When using a discriminant classifier, we make no assumption about $p(x)$, therefore the maximum likelihood estimate of β is the same for (10). See [1, 9] in the case of logistic classifiers.

These maximum likelihood estimates can be obtained via a modified CEM, we detail this algorithm for the particular case of the logistic classifier (1).

Logistic-CEM

Initialization: Train a discriminant logistic model $G^{(0)}(x)$ over D_l, let $P^{(0)}$ be the initial partition obtained from this model on $D_l \cup D_u$.

j^{th} iteration, $j \geq 0$:

E –step. Estimate the posterior probability that x_i belongs to P_k on D_u ($i=n+1,\dots, n+m$; $k=1,2$) using the output of the logistic classifier $G^{(j)}(x)$:

$$p(P_1^{(j)} / x) = \frac{1}{1 + \exp(-(\beta_0^{(j)} + \beta^{t(j)} . x))} \quad \text{and} \quad p(P_2^{(j)} / x) = 1 - p(P_1^{(j)} / x)$$

C – step. Assign each x_i to the cluster $P_k^{(j+1)}$ with maximal posterior probability according to $p(P_k^{(j)} / x)$:

$$\forall i, \forall k \in \{1,2\}, t_{ki}^{(j+1)} = (\text{sgn}(p(P_k^{(j)} / x_i; \beta^{(j)}) - 0.5) \qquad (11)$$

M–step. Find new parameters ($\beta_0^{(j+1)}$, $\beta^{(j+1)}$) which maximize log $\widetilde{L}_C$ ($P^{(j+1)}$, $\beta_0^{(j)}$, $\beta^{(j)}$).

For the logistic classifier, there is no analytical method for maximizing $\widetilde{L}_C$ ($P^{(j+1)}$, $\beta_0^{(j)}$, $\beta^{(j)}$) in the M-step. We have used the quasi-Newton procedure for that. An advantage of this method is that they require only the first order derivatives at each iteration while giving an estimate of the matrix of second order derivatives at the maximum point.

The main difference here with the generative method is that no assumption is made on the conditional densities f_1 and f_2 except that their log ratio is linear in x. The algorithm directly attempts to estimate the $p(P_1/x)$, which is the quantity we are interested in, instead of the conditional densities.

It will be shown later to outperform significantly the generative approach, especially when there are few labeled data available. It does so by using fewer parameters than the generative approach and is faster to run.

4 Numerical experiments

We have performed experiments to evaluate CEM-generative and CEM-logistic for two large real data sets. The task we have used is the selection of relevant sentences from documents for constructing an extract summary.

This is a classical approach to text summarization. From a machine learning point of view, this is a 2-class classification task: sentences have to be classified into relevant or irrelevant. The usual approach is to label document from a corpus at the sentence level and to train classifiers using a supervised classification approach as described in the seminal work of [8].

This is particularly well suited for semi-supervised learning since the construction of such datasets is particularly tedious. We have used two datasets *a*) the Reuters data set consisting of news-wire summaries: this corpus is composed of 1000 documents and their associated extracted sentence summaries and b) the Computation and Language (cmp_lg) collection of TIPSTER SUMMAC. This corpus is composed of 183 scientific articles.

In both cases, the data set was split into a training and a test set whose size was respectively 1/3 and 2/3 of the available data. Sentences are encoded into five features as described in [8].

4.1 Experimental Results

For evaluation measure, we have used the *average precision* which is a classical measure in information retrieval. This is more relevant here than the classical correct classification percentage and is computed as follows.

For each document, sentences are ranked according to the classifier score. The *k* top sentences are then selected for *k* = 1, 2, …. For each *k* value, we compute the *precision* at *k* :

$$\text{Precision} = \frac{\#\text{ of sentences extracted by the system which are in the target summaries}}{\text{total}\,\#\text{ of sentences extracted by the system}}$$

The average precision is the mean of all these values for all documents.

Table 1 compares a baseline naive Bayes classifier with the generative and logistic CEM classifiers, all trained in a fully supervised way. The two CEM classifiers allow for approximately 10% increase both in average precision and in accuracy over naive Bayes for both databases. This is not surprising, but this shows that they behave reasonably on this dataset. Logistic CEM is slightly above generative CEM.

Figure 1 shows the average precision on the test sets of Reuters and SUMMAC cmp-lg for different ratio of labeled-unlabeled documents in the training set, and for the generative and logistic semi-supervised algorithms On the *x*-axis, 10% means that 10% of the documents in the training sets were labeled for training, the 90% remaining being used as unlabeled training documents.

For comparison, we have also performed test with a logistic classifier trained only on the labeled sentences without using the unlabeled sentences in the training set.

The logistic classifier trained only on labeled sentences performs well but is clearly below the CEM-algorithms in the regions of interest where the ratio of labeled data is small (less than 50% in these experiments). This shows that unlabeled data indeed contains relevant information and that semi-supervised algorithms allow to extract part of this information.

Logistic CEM uniformly outperforms the other systems in all regions. This is particularly clear for SUMMAC cmp-lg, which is a small document set. In this case, the discriminant approach is clearly superior to the generative approach which suffers from estimation problems. With only 10% of labeled documents in the training set, the logistic CEM approach is over the baseline naive Bayes system and using about 20% of labeled documents allows to reach half the performance increase of the fully supervised approach. Another interesting result is that both logistic and generative CEM trained on semi-supervised learning scheme (using always 10% of labeled documents together with 90% of unlabeled documents on the training set) gave similar performances than the baseline naive Bayes classifier trained with all documents on the training set. Tests with more sophisticated classifiers did not lead to improvements compared to logistic classification. These results have still to be confirmed and refined by experiments on other datasets, but they rise hope that the proposed methods perform well for some real world tasks.

Table 1. Comparison between a baseline naive Bayes system and different learning classifiers on Reuters and cmp_lg test sets. All classifiers are trained in a fully supervised way. Accuracy is the ratio of correct classification for both relevant and irrelevant sentences.

	Reuters data set		Cmp_lg collection	
System	**Average Precision (%)**	**Accuracy (%)**	**Average Precision (%)**	**Accuracy (%)**
Naive Bayes	61,02	63,03	61,83	63,48
Generative-CEM	72,86	73,06	74,12	74,79
Logistic-CEM	73,84	74,22	75,26	76,92

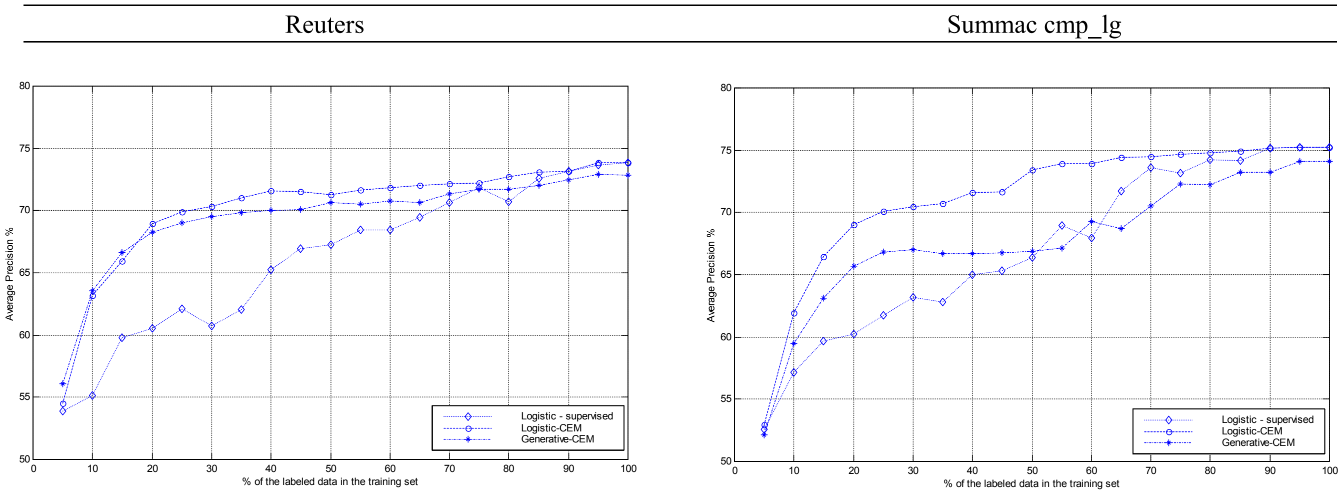

Figure 1. Average precision of 3 trainable summarizers with respect to the ratio of labeled documents in the training set. The systems are the logistic and generative CEM algorithms and a logistic unit trained only on labeled data.

5 CONCLUSION

We have introduced new generative and discriminant algorithms for training classifiers in presence of labeled and unlabeled data. These algorithms have been derived in the framework of CEM algorithms and are pretty general in the sense that they can be used respectively with any density estimation method and discriminant classifier. We have detailed the discriminant technique in the case of a logistic classifier. We have provided an experimental analysis of the proposed methods with regard to the ratio of labeled data in the training set, and we have shown that using only 10 to 20% of labeled allows to reach half of the performance increase provided by a fully supervised approach.

We have also compared discriminant and generative approaches to semi-supervised learning and the former has been found clearly superior to the latter especially for small collections.

6 REFERNCES

[1] J.A. Anderson, S.C. Richardson. ‘Logistic Discrimination and Bias correction in maximum likelihood estimation’. *Technometrics*, **21**, 71-78 (1979).

[2] A. Blum, T. Mitchell *Combining Labeled and Unlabeled Data with Co-Training*. Proceedings of the 1998 Conference on Computational Learning Theory. (1998).

[3] M. Collins and Y. Singer. *Unsupervised models for named entity classification*. In Proceedings of EMNLP, (1999)

[4] V. De Sa, *Learning classification with unlabeled data.* Advances in Neural Information Processing Systems. **6,** (1993).

[5] G. Celeux, G. Govaert ‘A classification EM algorithm for clustering and two stochastic versions’. *Computational Statistics & Data Analysis*, **14**, 315-332 (1992).

[6] D.R. Cox, *Some procedures associated with the logistic qualitative response curve*. Research Papers in Statistics: Festschrift for J. Neyman. Wiley edn., 55-71, (1966).

[7] R. O. Duda, P. T. Hart Pattern Recognition and Scene Analysis. Edn. Wiley (1973).

[8] J. Kupiec, J. Pedersen, F. Chen, *A Trainable Document Summarizer*. Proceedings of the 18th ACM SIGIR, 68-73, (1995).

[9] G.J. McLachlan *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons edn., New-York 1992.

[10] D. Miller, H. Uyar, *A Mixture of Experts classifier with learning based on both labeled and unlabeled data.* Advances in Neural Information Processing Systems. **9,** 571-577, (1996).

[11] K. Nigam, A. McCallum, A. Thrun, T. Mitchell, *Text Classification from labeled and unlabeled documents using EM*. In proceedings of National Conference on Artificial Intelligence. (1998).

[12] V. Roth, V. Steinhage, *Nonlinear Discriminant Analysis using Kernel Functions*. Advances in Neural Information Processing Systems, **12**, (1999).

[13] A.J. Scott, M.J. Symons, *Clustering Methods based on Likelihood Ratio Criteria*. Biometrics. **27,** 387-397 , (1991).

[14] M. .J. Symons, Clustering criteria and multivariate normal mixtures. Biometrics. **37**, 35-43, (1981).