Logistic regression for partial labels

Yves Grandvalet
Heudiasyc, UMR CNRS 6599,
Université de Technologie de Compiègne,
BP 20.529, 60205 Compiègne cedex, France
Yves.Grandvalet@hds.utc.fr

Abstract

This paper discusses learning from partially labeled data in the framework of probabilistic supervised classification. Minimum commitment logistic regression is a conservative solution to the problem of imprecise labels, which should be appropriate if the faithful estimation of posterior probabilities is an issue. Semi-supervised learning is among the problems considered, and a series of experiments shows that our second proposal, self-consistent logistic regression is a serious contender to more classical solutions involving generative models.

Keywords: partial labels, logistic regression semi-supervised learning.

1 Introduction

In the classical supervised learning classification framework, a decision rule is to be build from a learning set $\mathcal{L}_n = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where each example is described by a pattern $\boldsymbol{x}_i \in \mathcal{X}$ and by the response of a supervisor $y_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$. This response variable is a supposedly *the* correct class among the finite set of exclusive classes Ω.

This paper aims at providing means to construct probabilistic classification models when the learning set includes examples whose class is not precisely known. Instead of answering the correct class, the supervisor is only supposed to return a subset of possible classes which should include the correct solution. This kind of information is sometimes a more faithful description of the true state of knowledge when labeling is performed by an expert. For example, in medical diagnosis, a physician is sometimes able to discard some diseases, but not to pinpoint the precise illness of his patient. The problem may also arise because the information required for specifying a single label is not available, since differentiating between two or more classes requires tests which are not systematically performed on all examples. Last but not least, some examples may not be labeled at all: in particular, semi-supervised learning is a special case of partially labeled problem, where all examples are either precisely labeled or unlabeled, *i.e.* with labels belonging to Ω.

Partial labelling has been investigated in the frameworks of probability and Dempster-Shafer theories [1]. Dempster-Shafer theory enables to reason on beliefs expressed on subsets of Ω without distributing them to singletons. Its description is out of the scope of this paper which focuses on the probabilistic framework. The reader is referred to [1] and references therein. Ambroise *et al.* [1] also propose a probabilistic solution based on an extension of the EM algorithm for fitting mixture models. The algorithms presented here differ in the respect that they do not model the joint distribution of data $(\boldsymbol{x}, y)$, but only the conditional probability of $(y|\boldsymbol{x})$.

The particular case of semi-supervised learning problem has recently received much attention and several solutions have already been proposed, *e.g.* [3, 5, 8]. Most of them rely on some explicit or implicit model of the joint distribution of data [5, 8]. One exception is provided by Bennett and Demiriz [3], where solutions having many examples near the decision boundary are penalized. Another one is given by Anderson [2] who generalized logistic regression to semi-supervised learning for discrete (or discretized) explicative variables. The algorithms presented here also apply to logistic regression, but they may be used either for discrete or continuous explicative variables.

2 Algorithms

In this paper, we focus on logistic regression, which is a generalized linear model providing linear discriminant rules. Its simplicity makes it an ideal guinea-pig for testing maximum likelihood type criteria devoted to learning in the presence of partial labels. The principles discussed here are however easily generalized to any discriminant method based on maximum likelihood estimation of posterior probabilities, such as generalized additive models [6] or neural networks [4].

2.1 Logistic regression

Logistic regression fits the log-ratio of posterior probabilities by a linear model. The corresponding estimate of the posterior probability $P(y = \omega_k|\boldsymbol{x})$ is given by

$$f_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_k^T \boldsymbol{x})}{\sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \boldsymbol{x})} \ , \qquad (1)$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_k\}_{k=1}^{K}$ is the set of parameters of the model, which is determined by maximizing the log-likelihood

$$L(\boldsymbol{\beta}; \mathcal{L}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log(f_k(\boldsymbol{x}_i)) \ , \qquad (2)$$

where t_{ik} are the so-called dummy variables coding class membership: if $y_i = \omega_k$, then , $t_{ik} = 1$ and $t_{ij} = 0$, for $j \neq k$. The log-likelihood (2) assumes a multinomial distribution[1] for $(y|\boldsymbol{x})$ whose parameters are constrained by the linear relationship on log-ratio of posterior probabilities (1). It thus encompasses several models of joint distribution $(\boldsymbol{x}, y)$, which may be characterized either by discrete, continuous, or partially discrete and continuous models on $(\boldsymbol{x}|y)$. This ubiquity renders the estimate less sensitive to the distributional form postulated [2].

The criterion (2) is a convex function of the model parameters $\boldsymbol{\beta}$ (1). Provided precautions are taken to avoid redundant parameterization, the Newton-Raphson algorithm efficiently determines the global maximizer. This algorithm also applies to maximum a posteriori criteria with Gaussian priors on parameters $\boldsymbol{\beta}$.

2.2 Generalizing logistic regression to partial labels

Let $\boldsymbol{z}_i \in \{0, 1\}^K$ denote the dummy variables representing partial labels, where $z_{ik} = 1$ means that ω_k is a possible label, whereas $z_{ik} = 0$ means that the label is not ω_k. In other words, $\boldsymbol{z}_i$ is the indicator of the subset $\Omega_i \subseteq \Omega$ corresponding to the partial label. A first solution to the problem of learning from partial labels consists in maximizing the likelihood of the observed sample $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^{n}$, assuming model (1) (with $2^K - 1$ modalities) for $P(\boldsymbol{z}|\boldsymbol{x})$. This solution does not address the right problem: we are not interested in modeling the imprecise responses of the teacher; our goal is to uncover the distribution of the correct label y knowing $\boldsymbol{x}$.

Thus, the logistic model (1) should represent the posterior probability of the correct label, and the generalization to partial labels should only affect the criterion (2). Three criteria are considered in the following sections.

[1] For notational convenience, we consider (without loss of generality) that only one observation is observed at each pattern $\boldsymbol{x}_i$.

2.3 Maximum entropy logistic regression

Let $g_k(\boldsymbol{z}_i, \boldsymbol{x}_i)$ model $P(y_i = \omega_k|\boldsymbol{z}_i, \boldsymbol{x}_i)$, the log-likelihood is

$$L_{\boldsymbol{g}}(\boldsymbol{\beta}; \mathcal{L}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) \log(f_k(\boldsymbol{x}_i)) \ . \tag{3}$$

Several models $\boldsymbol{g}$ may be proposed. Our interpretation of partial labels (restricting the set of possible labels) only implies the constraint $g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = z_{ik} g_k(\boldsymbol{z}_i, \boldsymbol{x}_i)$.

One of the most notorious principle of probabilistic inference is to assume the equirepartition in the distribution of unknown variables. In the present context, this principle amounts to consider that, in absence of other pieces of information, the distribution of the true label should be modelled by

$$g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = \frac{z_{ik}}{\sum_{j=1}^{K} z_{ij}} \ . \tag{4}$$

Once this model is assumed, the log-likelihood $L_{\boldsymbol{g}}$ (3) can be computed. It is concave and can be maximized with the algorithm used in standard logistic regression.

The resulting criterion

- is the log-likelihood on precisely labeled examples ($z_{ik}/\sum_{j=1}^{K} z_{ij} = t_{ik}$);
- sustains equirepartition of estimated posterior probabilities within the set of possible labels;
- in particular, unlabeled data favor identical values of $f_k(\boldsymbol{x}_i)$, $k = 1, \ldots, K$.

This last point tells us that unlabeled examples do not convey information. On the contrary, they increase the entropy of posterior probabilities. They are thus considered as a source of uncertainty decreasing the information content of the training sample. This is clearly a defect of the method. It stems from the model of $P(y|\boldsymbol{z}, \boldsymbol{x})$ (4), where $\boldsymbol{x}_i$ is not considered as a piece of information regarding y_i. The models of $P(y|\boldsymbol{z}, \boldsymbol{x})$ and $P(y|\boldsymbol{x})$ used in the criterion are thus inconsistent. A suitable model for $P(y|\boldsymbol{z}, \boldsymbol{x})$ will be considered in section 2.5.

2.4 Minimum commitment logistic regression

The events observed in the learning set $\mathcal{L}_n = \{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^{n}$ pertain to the membership of y_i to subsets of Ω. If applied directly to these events, the maximum likelihood principle does not require to model $P(y|\boldsymbol{x}, \boldsymbol{z})$, so that no additional assumption is necessary. A minimum commitment[2] solution is thus obtained, where, contrary to the maximum entropy principle, information supporting equirepartition of probabilities is differentiated from the absence of information.

Noting that the event "y_i belongs to the subset Ω_i indicated by $\boldsymbol{z}_i$" follows a Bernoulli distribution of parameter $P(y_i \in \Omega_i|\boldsymbol{x}_i) = \sum_{k=1}^{K} z_{ik} P(y_i = \omega_k|\boldsymbol{x}_i)$, the log-likelihood is

$$L(\boldsymbol{\beta}; \mathcal{L}_n) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} z_{ik} f_k(\boldsymbol{x}_i)\right) \ . \tag{5}$$

This criterion is concave, and the global solution can be obtained by the Newton-Raphson algorithm. Interestingly, another algorithm provides the global optimum of minimum commitment logistic regression. It consists in an EM algorithm alternating updates of the two sets of variables $\boldsymbol{f}$ and $\boldsymbol{g}$ of $L_{\boldsymbol{g}}$ (3). The M-step maximizes $L_{\boldsymbol{g}}$ with respect to $\boldsymbol{f}$ for a fixed value of $\boldsymbol{g}$; the E-step defines $\boldsymbol{g}(\boldsymbol{z}_i, \boldsymbol{x}_i)$ as the projection of $\boldsymbol{f}(\boldsymbol{x}_i)$ on the set of distributions compatible with $\boldsymbol{z}_i$:

$$g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = \frac{z_{ik} f_k(\boldsymbol{x}_i)}{\sum_{j=1}^{K} z_{ij} f_j(\boldsymbol{x}_i)} \ . \tag{6}$$

The criterion of minimum commitment logistic regression only penalizes evidence given to classes which are inconsistent with the partial label, without setting any preferences among the compatible distributions:

[2] This term was coined in the framework of Dempster-Schafer theory of evidence [1].

- it is the log-likelihood on precisely labeled examples ($g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = t_{ik}$);
- it is vacuous within possible labels (any distribution with constant mass $m_i = \sum_{k=1}^{K} z_{ik} f_k(\boldsymbol{x}_i)$ achieves the same value of the criterion) ;
- in particular, it is vacuous for unlabeled examples (for which the gradient $\partial L_{\boldsymbol{g}} / \partial \boldsymbol{f}$ is orthogonal to the constraint $\sum_{k=1}^{K} f_k(\boldsymbol{x}) = 1$).

The last point is in agreement with the precaution principle of minimum commitment: in the absence of any assumptions on $P(y|\boldsymbol{z}, \boldsymbol{x})$, an unlabeled example conveys no information. Of course, it is neither considered as a source of uncertainty.

This algorithm provides thus a solution to partially labeled data, but it can not be used to enhance the performances of a classifier on a semi-supervised task. When the learning set comprises only precisely labeled and unlabeled examples, minimum commitment logistic regression provides the same solution as standard logistic regression based only on labeled examples. We derive below another algorithm, which attempts to benefit from unlabeled examples by assuming that the model predictions may be a valuable indicator of the true class among possible ones.

2.5 Self-consistent logistic regression

Our last proposal can be viewed as an extension of minimum commitment logistic regression. As for the latter, this algorithm consists in maximizing $L_{\boldsymbol{g}}$ (3), where $\boldsymbol{g}$ is defined as the projection of the model of $P(y|\boldsymbol{x})$ on the set of distributions compatible with $\boldsymbol{z}$. The difference is that $L_{\boldsymbol{g}}$ (3) is now maximized with respect to both variables $\boldsymbol{f}$ and $\boldsymbol{g}$. As $\boldsymbol{g}$ is given by (6), the criterion is a function of $\boldsymbol{f}$ alone:

$$L(\boldsymbol{\beta}; \mathcal{L}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{z_{ik} f_k(\boldsymbol{x}_i)}{\sum_{j=1}^{K} z_{ij} f_j(\boldsymbol{x}_i)} \log(f_k(\boldsymbol{x}_i)) \ . \tag{7}$$

This cost function is not concave, but a local maximum can be computed by quasi-Newton algorithm. Our implementation uses the BFGS algorithm with the initial solution provided by minimum commitment logistic regression.

The maximization of criterion (7) penalizes class assignments which are inconsistent with the partial label, but also the equirepartition of posterior probabilities:

- it maximizes log-likelihood on precisely labeled examples ($g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = t_{ik}$);
- it penalizes the mass given to incorrect labels while minimizing entropy within the subset of possible labels for partially labeled examples;
- it minimizes entropy for unlabeled examples ($g_k(\boldsymbol{z}_i, \boldsymbol{x}_i) = f_k(\boldsymbol{x}_i)$).

Thus, the algorithm outputs precise predictions compatible with the possible labels. By "betting" on its ability to uncover the posterior distribution $P(y|\boldsymbol{x})$, the algorithm may benefit from unlabeled examples. The strategy is acknowledgedly risky and may turn a bad minimum commitment solution into a worse one.

Note that, compared to the maximum entropy solution, ensuring the highest consistency between the two models $\boldsymbol{f}$ and $\boldsymbol{g}$ results in a solution where unlabeled examples are considered to be informative. As in [3], solutions having many examples near the decision boundary are penalized. This kind of penalization can be motivated by theoretical arguments showing that unlabeled examples convey information about discrimination when classes are separated [7].

Experiments

For lack of space, we do not report experimental results where examples are labeled by subsets of Ω. Instead, we focus on the semi-supervised learning task which is more frequently encountered. The experimental setup is simple in order to avoid artifacts stemming from optimization problems. Our goal is to check to what extent supervised learning can be improved by unlabeled examples, and if our solution can compete with the one provided by generative models which are usually advocated in this framework.

Self-consistent logistic regression is thus compared to standard logistic regression and the EM algorithm for mixture models. The benchmark is a series of two-classes problems, where each class is generated with equal probability from a multivariate normal distribution. Class 1 is multivariate normal with mean $(00\dots0)$ and covariance matrix identity. Class 2 is multivariate normal with mean $(aa\dots a)$ and covariance matrix identity. The number of features is varied from 10 to 50 ($d = 10, 20, 30, 40, 50$) a tunes the Bayes error which varies from 2.5 % to 40 % (2.5 %, 5 %, 10 %, 20 %, 40 %). The learning set sizes range from 200 to 1700 ($n = 200, 300, 500, 800, 1700$) and the rate of missing label from 0 % to 95 % (0 %, 50 %, 75 %, 90 %, 95 %). Overall, 750 different setups are evaluated, and for each of them, 10 different training samples are generated, resulting in a total of 7500 experiments. Generalization performances are estimated on a test set of size 5000. The mixture models fitted by EM correspond to the model that truly generated data: two Gaussian subpopulations, with covariance matrices restricted to be equal. The logistic regression model is also compatible with the data distribution since the log-ratio of posterior probabilities is truly linear, but this compatibility is less demanding than the one on the joint distribution. In this respect, the benchmark is highly favorable to the generative modeling approach.

As there is no model bias, differences in prediction error rates are only due to differences in estimation efficiency. The average Bayes recognition rate is 84.5 %. The overall recognition rates are in favor of self-consistent logistic regression (81.2 %), followed by the EM algorithm (79.5 %) and logistic regression (77.1 %). Figure 1 provides a more complete summary of results. The plots represent the recognition rates for the three methods versus the rate of missing label and the Bayes error rate. For each curve, the reported results are averaged over all other setup parameters. The first plot shows that our benchmark is fair to mixture models: the performances of the three methods are about identical when no label are missing, and the advantage of self-consistent logistic discrimination is clearly shown to be due to its ability to handle learning sets with few labeled data.

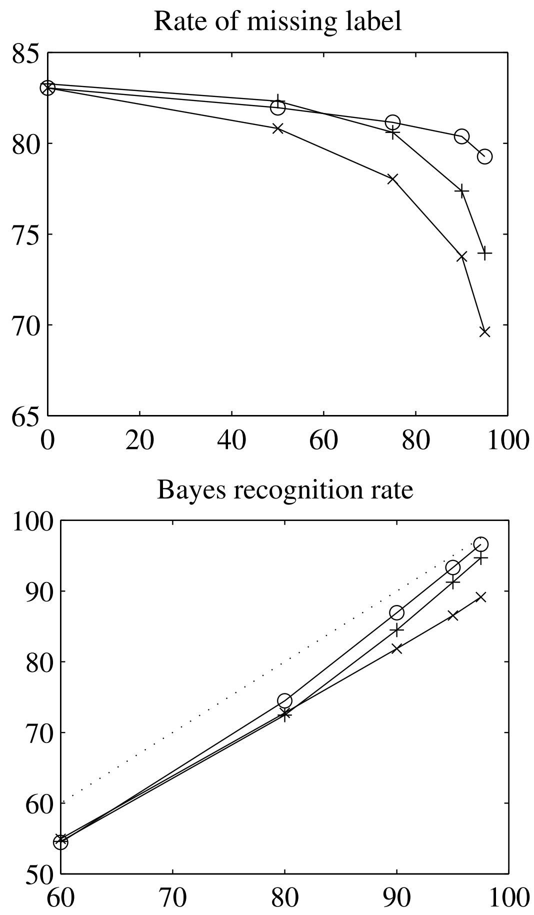

Figure 1: Average recognition rate (*vs.*) missing label and Bayes recognition rates for self-consistent logistic regression ($\circ$), mixture models ($+$) and logistic regression ($\times$).

An interesting, though rather intuitive conclusion results from the plot of recognition rates versus Bayes error. Either for EM or self-consistent logistic regression, unlabeled examples are mostly bene ficial when the Bayes error is low, *i.e.* when clusters are well separated. Experiments meet theory: O'Neill [7] shows that the asymptotical information content of unlabeled examples decreases as classes overlap. Figure 2 illustrates how the ratio of sample size to input dimension n/d affects recognition rates. Mixture models perform best when n/d is high, but their number of free parameters grow as the square of the number of features, while it grows linearly for the two logistic models. Hence self-consistent logistic regression is by far the best when the ratio n/d is low. When the Bayes error decreases, the cross-over point decreases ($n/d \simeq 20$ for a 90 % recognition rate, $n/d \simeq 10$ for a 97.5 % recognition rate); the advantage of self-consistent logistic regression before this point is higher and the one of mixture models past this point is lower. Overall, the decision rule provided by self-consistent logistic regression seems thus preferable.

3 Discussion

In this paper, we proposed two criteria to handle partial labels in supervised classification techniques. When applicable, the conservative solution provided by minimum commitment should improve classification and the estimation of posterior probabilities. The more radical self-consistent method promotes classifiers with high confidence. It should be less reliable regarding posterior probabilities, but the estimation of decision rule can benefit from unlabeled examples.

Our preliminary experiments suggest that self-consistent logistic discrimination may be a serious contender to generative models in semi-supervised learning. This result is encouraging research devoted to semi-supervised learning or transduction within the so-called diagnosis paradigm. In particular, the algorithms presented here could readily be generalized to many classifiers. For example, logistic regression has been generalized to additive models [6], and feed-forward neural networks with softmax activation function at the output layer [4]. The generalized criteria leading to minimum commitment and self-consistent logistic regression are thus directly applicable to these classifiers.

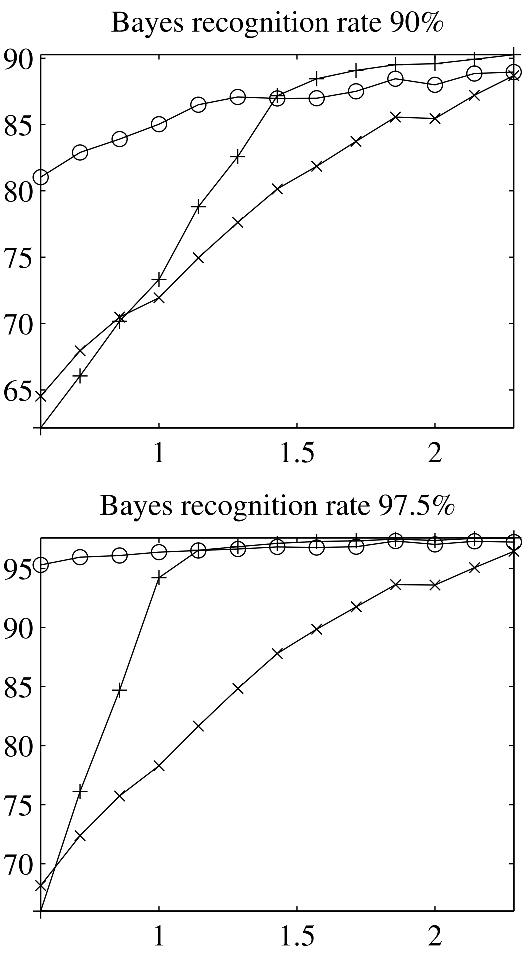

Figure 2: Average recogition rates for two Bayes recognition rates (and 90 % of missing label) *vs.* ratio of sample size to input dimension (log-scale), for self-consistent logistic regression ($\circ$), mixture models ($+$) and logistic regression ($\times$).

References

[1] C. Ambroise, T. Denœux, G. Govaert, and P. Smets. Learning from an imprecise teacher: probabilistic and evidential approaches. In *10th International Symposium on Applied Stochastic Models and Data Analysis*, volume 1, pages 101–105, June 2001.

[2] J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

[3] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In M. S. Kearns, S. A. Solla, and D. A.

Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 368–374. MIT Press, 1999.

[4] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, relationships to statistical pattern recognition. In F. Fogelman Soulié and J. Hérault, editors, *Neuro-computing: Algorithms, Architectures and Applications*, pages 227–236. Springer, 1990.

[5] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 416–422. MIT Press, 2001.

[6] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1990.

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

[8] M. Szummer and T. S. Jaakkola. Kernel expansions with unlabeled examples. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 626–632. MIT Press, 2001.