

In several application domains such as biology, computer vision, social network analysis and information retrieval, multi-class classification problems arise in which data instances not simply belong to one particular class, but exhibit a certain membership to all the classes. Often referred to as learning fuzzy, mixed or partial memberships, this type of problems has been recently studied in fields like fuzzy set theory, statistics and machine learning, mainly from an unsupervised learning perspective. Given the interpretation of fuzzy class memberships in real-world applications, we present a supervised probabilistic approach. To this end, we show in particular how kernel-based logistic regression models can learn fuzzy memberships in an adequate manner, just by replacing zero-one-coded class labels by fuzzy labels in the likelihood function. Empirical results on several real-world data sets show that our approach leads to quasi-identical results but a tremendous gain in computational complexity, when compared with a naive algorithm that transforms each fuzzy class label into many zero-one coded class labels.

**Keywords:** fuzzy/mixed/partial membership models, logistic regression, kernel methods, machine learning.

## 1 Introduction

As a general introduction to the concept of fuzzy membership models, let us start with a slightly controversial yet suitable example about the classification of humans into ethnic groups. Until the 18th century it has been claimed by scientists that all humans can be subdivided into five main classes, namely white, black, yellow, red and brown people. Nowadays such superseded views have been recognized as too restrictive: in which class do we put for example a man with a European mother and an Asian father? In some sense one can say that such a man obtains a membership to both classes white and yellow. So, hypothetically speaking, if we would like to construct a machine learning algorithm for classifying humans, for example based on

DNA data, then we would definitely need a classifier capable of handling fuzzy class labels.

Actually, it turns out that many real-world multi-class classification problems can be translated into a setting where non-crisp class labels are observed. Scene classification in computer vision is such an application, as images can simultaneously belong to let's say the classes "sunsets" and "historic buildings" [Woods et al., 1995, Boutell et al., 2004]. Here a strong connection with multi-label classification and multi-label ranking can be seen [Fürnkranz et al., 2008, Hüllermeier et al., 2008]. Similar arguments hold for related application domains such as text and scientific literature categorization [Erosheva et al., 2004] and social network analysis [Koutsourakis and Eliassi-Rad, 2008].

Other applications can be found in the life sciences. Fuzzy membership models have been applied to the classification of satellite images for crop-land suitability analysis [Nisar-Ahamad et al., 2000]. In microbiology as well, fatty-acid profiles describe bacterial species in terms of fuzzy memberships [Marttinen et al., 2008], and in biochemistry fuzzy memberships are observed when chemical compounds are produced in certain proportions, thereby depending on the chosen environmental conditions. Finally, biological experts in agriculture often screen a fixed number of plants to construct disease prediction models, such that for each field a fraction of these plants is classified in each of the disease classes, again resulting in fuzzy class memberships [Isebaert et al., 2009].

In the latter applications rather a link with probabilistic classifiers and multi-class probability estimation can be claimed. Indeed, from a probabilistic perspective, one can interpret fuzzy memberships as probability estimates for the class labels or as class proportions, so that in essence for each data object in the training data set a multinomial distribution over the class labels is observed. In this light fuzzy membership models and traditional probabilistic classifiers both produce probability estimates as outputs, but

they differ in the sense that the former models directly take fuzzy memberships as training labels, while the latter ones only accept crisp training labels as input. In some implementations of generalized linear models, fuzzy memberships can be indirectly processed as training labels for binary classification, but this trick is at least not widely established and rarely supported by statistical packages [Agresti, 2002].

Given the large number of potential applications in various domains, researchers in statistics, machine learning and fuzzy sets have shown interest in developing learning algorithms for fuzzy memberships. In statistics and machine learning the notions mixed membership or partial membership are mainly established, and here previous research was mainly focussed on unsupervised settings, such that in clustering algorithms a data instance can simultaneously exhibit a membership in several clusters. Such ideas have for example been incorporated in mixed models [Gormley and Murphy, 2008], probabilistic graphical models [Airoldi et al., 2008] and Bayesian clustering techniques [Heller et al., 2008].

In fuzzy set theory, unsupervised approaches have been popular too, consider for example the well-known fuzzy c-means clustering method, but here a few supervised algorithm have been proposed as well [Cai et al., 2007, Orsenigo and Vercellis, 2007]. Nevertheless, these algorithms tend to depart from a similar setup as probabilistic classifiers, in the sense that they also assume crisp class labels for training data, but fuzzy memberships are predicted instead of probability estimates.

While assuming a probabilistic interpretation of fuzzy class memberships, we will in this article discuss how the kernel logistic regression method can be generalized so that fuzzy class labels are accepted as input. We give in the following section a brief summary of the basic algorithm, followed by its extension for fuzzy memberships in Section 3.

## 2 Kernel logistic regression for multi-class problems

We start with introducing some notations. Let us in general assume a  $K$ -class classification problem in which the classes are formally denoted as  $\mathcal{C}_1, \dots, \mathcal{C}_K$ . For ease of notation, training data will consist of pairs of the form  $(\mathbf{x}, \mathbf{y})$  in which  $\mathbf{x}$  represents a feature vector and  $\mathbf{y}$  is a vector of length  $K$  such that the  $k$ -th entry is 1 when the corresponding instance has label  $\mathcal{C}_k$ , while all other entries are zero. This condition will be relaxed in the following section such that the vector  $\mathbf{y}$  will represent fuzzy memberships to all classes. Furthermore, let us assume that examples are identically and independently drawn according to an unknown joint distribution over an input space  $\mathcal{X}$  and a label space  $\mathcal{Y}$ . We define a data set of size  $N$  as  $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . We will use the notation  $\mathbf{y}_n = \{y_{n1}, \dots, y_{nK}\}$  so that  $y_{nk}$  denotes the  $k$ -th entry in the label vector that corresponds with the  $n$ -th data instance.

Given this notation, multi-class logistic regression (KLR) methods estimate the probability that a given data instance belongs to class  $\mathcal{C}_k$  as follows:

$$\begin{aligned} p_k(\mathbf{x}_n) &= \Pr\{y_{nk} = 1 \mid \mathbf{x}_n\} \\ &= \frac{\exp(f_k(\mathbf{x}_n))}{\sum_{l=1}^K \exp(f_l(\mathbf{x}_n))}, \end{aligned}$$

in which  $f_1, \dots, f_K : \mathcal{X} \rightarrow \mathbb{R}$  are scoring functions that assign a continuous value to data instances. In traditional logistic regression models these scoring functions are just linear models. In kernel logistic regression models [Zhu and Hastie, 2004], they can be generally represented in the following way:

$$f_k(\mathbf{x}) = \mathbf{w}_k \cdot \phi(\mathbf{x}) + b,$$

with  $\phi$  representing a feature mapping to a possibly high-dimensional feature space and  $\mathbf{w}_1, \dots, \mathbf{w}_K$  vectors of parameters that must be estimated based on training data. According to the representer theorem [Schölkopf and Smola, 2002], the scoring functions can

be equivalently expressed as follows:

$$f_k(\mathbf{x}) = \sum_{n=1}^N \alpha_{nk} K(\mathbf{x}_n, \mathbf{x}),$$

with  $K$  a kernel corresponding to  $\phi$  and  $\alpha_{nk}$  dual parameters.

In probabilistic models an estimate of the parameters is usually obtained by maximum likelihood estimation. The multinomial likelihood is given by:

$$\begin{aligned} L(\mathbf{w}_1, \dots, \mathbf{w}_K) &= \Pr\{\mathbf{y}_1, \dots, \mathbf{y}_N \mid \mathbf{w}_1, \dots, \mathbf{w}_K\} \\ &= \prod_{n=1}^N \prod_{k=1}^K (p_k(\mathbf{x}_n))^{y_{nk}}. \end{aligned}$$

Equivalently, we can minimize the negative log-likelihood:

$$\begin{aligned} \ln L &= \sum_{n=1}^N \sum_{k=1}^K y_{nk} \ln(p_k(\mathbf{x}_n)) \\ &= \sum_{n=1}^N \sum_{k=1}^K y_{nk} \left( f_k(\mathbf{x}_n) \right. \\ &\quad \left. - \ln \left( \sum_{l=1}^K \exp(f_l(\mathbf{x}_n)) \right) \right). \end{aligned}$$

In the two-class case, the minimum is usually found with gradient descent. For the multi-class case, one arrives at a constrained optimization problem, since it must hold that:

$$\sum_{k=1}^K p_k(\mathbf{x}) = 1, \quad \forall \mathbf{x}.$$

Variants of the sequential minimal optimization algorithm found in implementations of support vector machines have been proposed for the multi-class case [Keerthi et al., 2005, Zhu and Hastie, 2005].

### 3 Learning fuzzy memberships with KLR

In order to extend kernel logistic regression so that data objects can have fuzzy memberships to all of the classes, we will allow now that  $\mathbf{y} \in [0, 1]^K$  instead of  $\mathbf{y} \in \{0, 1\}^K$ . In addition,

we will impose the following constraint on the label vectors:

$$\sum_{k=1}^K y_{nk} = 1, \quad \forall n.$$

This constraint results in a probabilistic interpretation of fuzzy class memberships, and it also guarantees that we rather stay in a multi-class classification setting instead of a multi-label classification setting. A naive and computationally inefficient approach for modelling fuzzy memberships with existing statistical tools could consist of artificially creating a new (much larger) dataset  $D^* = \{(\mathbf{x}_1^*, \mathbf{y}_1^*), \dots, (\mathbf{x}_{N^*}^*, \mathbf{y}_{N^*}^*)\}$  by multiplying the original dataset size with a certain factor. In this new dataset we then assign crisp labels class to the duplicates of the original data instances, in accordance with the distribution generated by the fuzzy memberships, so that  $\mathbf{y}_m^* \in \{0, 1\}^K$  for all  $(\mathbf{x}_m^*, \mathbf{y}_m^*) \in D^*$ . Formally, let the multiplication factor be  $\tau$ , then  $|D^*| = \tau \times |D|$  and

$$\mathbf{y}_n \simeq \frac{1}{\tau} \sum_{m=1: \mathbf{x}_m = \mathbf{x}_n}^{N^*} \mathbf{y}_m, \quad n \in \{1, \dots, N\},$$

Basically, the larger  $\tau$  is, the better the approximation, but also the more intractable this approach becomes, because the size of  $D^*$  blows up very fast. To overcome this computational bottleneck, we can write down the log-likelihood for  $D^*$  as follows:

$$\begin{aligned} \ln L &= \sum_{m=1}^{N^*} \sum_{k=1}^K y_{mk}^* \ln(p_k(\mathbf{x}_m^*)) \\ &= \sum_{n=1}^N \sum_{t=1}^{\tau} \sum_{k=1}^K y_{u(n,t)k}^* \ln(p_k(\mathbf{x}_n)) \\ &\simeq \sum_{n=1}^N \sum_{k=1}^K y_{nk} \ln(p_k(\mathbf{x}_n)) \\ &= \sum_{n=1}^N \sum_{k=1}^K y_{nk} \left( f_k(\mathbf{x}_n) \right. \\ &\quad \left. - \ln \left( \sum_{l=1}^K \exp(f_l(\mathbf{x}_n)) \right) \right). \end{aligned}$$

in which an index function  $u : \mathbb{N}[1, N] \times \mathbb{N}[1, \tau] \rightarrow \mathbb{N}[1, N^*]$  is used in order to map

an object from  $D$  to its  $t$ -th duplicate in  $D^*$ . Thus, we can avoid the construction of  $D^*$ , while still a very similar log-likelihood function is minimized. Beside computational advantages, this approach is also conceptually preferred since no approximation is required any more.

## 4 Experiments

## 5 Discussion

In this paper we considered multi-class classification problems where uncertainty is observed in the class labels, so that data instances obtain a fuzzy membership to several classes. As an extension of kernel logistic regression methods, we presented a simple, yet effective supervised learning approach to model this type of data. To this end, it was shown that the naive idea of multiplying the original dataset and replacing the fuzzy labels by several crisp labels can be avoided, since fuzzy class labels can be included in the log-likelihood in an elegant way. Initial experimental results on synthetic and real-world data confirm that including the fuzzy memberships in the log-likelihood function leads to quasi-identical results as multiplying the dataset. Simultaneously, the computational burden of the latter approach can be avoided. To this end, an existing R implementation of kernel logistic regression was modified. During the talk we will explain more thoroughly the empirical results.

We would like to thank Ji Zhu from the University of Michigan for providing R code on kernel logistic regression.

## References

- A. Agresti. *Categorical Data Analysis, 2nd version*. John Wiley and Sons, 2002.
- E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9: 1981–2014, 2008.
- M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37:1751–1771, 2004.
- W. Cai, S. Chen, and D. Zhang. Robust fuzzy relational classifier incorporating the soft class labels. *Pattern Recognition Letters*, 28:2250–2263, 2007.
- E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101:5220–5227, 2004.

Figure 1: Example of figure title

- J. Fürnkranz, E. Hüllermeier, E. Mencia, and K. Brinker. Multilabel classification by calibrated label ranking. *Machine Learning*, 73:133–153, 2008.
- I. Gormley and T. Murphy. A mixture of experts models for rank data with applications in election studies. *Annals of Applied Statistics*, 2:1452–1477, 2008.
- K. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th Conference on Machine Learning, Helsinki, Finland*, pages 392–399, 2008.
- E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1916, 2008.
- S. Isebaert, S. De Saeger, R. Devreese, R. Verhoeven, P. Maene, B. Heremans, and G. Haesaert. Mycotoxin-producing fusarium species occurring in winter wheat in Belgium (Flanders) during 2002-2005. *Journal of Phytopathology*, 157:108–116, 2009.
- S. Keerthi, K. Duan, S. Shevade, and A. Poo. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61:151–165, 2005.
- P.S. Koutsourelakis and T. Eliassi-Rad. Finding mixed-memberships in social networks. In *Proceedings of the AAAI Spring Symposium on Social Information Processing, Stanford, CA, USA*, 2008.
- P. Marttinen, J. Tang, B. De Baets, P. Dawyndt, and J. Corander. Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:74–85, 2008.
- T. Nisar-Ahamad, K. Gropal-Rao, and J. Murthy. GIS-based fuzzy membership model for crop-land suitability analysis. *Agricultural Systems*, 63:75–95, 2000.
- C. Orsenigo and C. Vercellis. Evaluating membership functions for fuzzy discrete SVM. *Lecture Notes in Computer Science*, 4578:187–194, 2007.
- B. Schölkopf and A. Smola. *Learning with Kernels, Support Vector Machines, Regularisation, Optimization and Beyond*. The MIT Press, 2002.
- K. Woods, D. Cook, L. Hall, K. Bowyer, and L. Stark. Learning membership functions in a function-based object recognition system. *Journal of Artificial Intelligence Research*, 3:187–222, 1995.
- J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5:427–443, 2004.
- J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14:185–205, 2005.