Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. 2005. On discriminative joint density modeling. In: João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo (editors). Proceedings of the 16th European Conference on Machine Learning (ECML 2005). Porto, Portugal. 3-7 October 2005. Berlin, Germany. Springer. Lecture Notes in Artificial Intelligence, volume 3720, pages 341-352.

On Discriminative Joint Density Modeling

Jarkko Salojärvi[1], Kai Puolamäki[1], and Samuel Kaski[1,2]

[1] Laboratory of Computer and Information Science, Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
{forename.surname}@hut.fi
[2] Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014 University of Helsinki, Finland

Abstract. We study discriminative joint density models, that is, generative models for the joint density $p(c, \mathbf{x})$ learned by maximizing a discriminative cost function, the conditional likelihood. We use the framework to derive generative models for generalized linear models, including logistic regression, linear discriminant analysis, and discriminative mixture of unigrams. The benefits of deriving the discriminative models from joint density models are that it is easy to extend the models and interpret the results, and missing data can be treated using justified standard methods.

1 Introduction

We study a classification task where a learning set, consisting of paired data $(\mathbf{x}, c)$, is given. The c is the value of a categorical variable, associated with observations $\mathbf{x}$. The observations may be collected from several different kinds of data sources; some may be real-valued measurements from sensors, whereas some may be probabilistic predictions. What all the values $\mathbf{x}$ have in common is that the c are assumed to depend on them. The task is to predict c for a test set where only the values of $\mathbf{x}$ are known. The c are often referred to as the (values of the) dependent variable, and the $\mathbf{x}$ the values of the independent variable or covariate.

There are two traditional modeling approaches for predicting c, discriminative and generative. Discriminative models optimize the conditional probability $p(c|\mathbf{x})$ (or some other discriminative criterion) directly. The models are good classifiers, since they do not waste resources on modeling those properties of the data that do not affect the value of c, that is, the distribution of $\mathbf{x}$. A classic example of a discriminative model is logistic regression, which is a special case of Generalized Linear Models (GLMs) [1]. In GLMs, functions of linear combinations $\beta^T\mathbf{x}$ of the independent variables are sought in order to predict $p(c|\mathbf{x}, \beta)$.

The other traditional approach is generative modeling of the joint distribution $p(c, \mathbf{x})$. The benefit of generative models is that compared to purely discriminative models, they add prior knowledge of the distribution of $\mathbf{x}$ into the task. This facilitates for example inferring missing values, since the model is

assumed to generate also the covariates $\mathbf{x}$. The models are often additionally simpler to construct, and their parameters offer simple explanations in terms of expected sufficient statistics. A classic example of generative models is the linear discriminant analysis (LDA).

Several publications have been devoted to comparing the discriminative and generative approaches [2–4]. A common model pair in the comparisons has been Linear Discriminant Analysis (or Naive Bayes) vs. logistic regression. With infinite amount of data, generative modeling by maximizing the joint likelihood produces optimal parameters for classification, assuming that the true data distribution is contained in the model family. However, with real-world data this is unlikely [5], and better predictions for c can be achieved by maximizing the conditional likelihood.[3] In practice, with large amounts of data, generative models are inferior to discriminative models, since the assumed model is always incorrect, but with small sample sizes generative models may show better performance [4].

The two modeling approaches are related. A discriminative classifier can be obtained by simply changing the objective function from the joint likelihood $p(c, \mathbf{x}|\theta)$ to the conditional likelihood $p(c|\mathbf{x}, \theta)$ by use of the Bayes formula, and then optimizing the model parameters. The method has been put to extensive use in speech processing applications, where good results have been obtained using discriminative hidden Markov models [6]. What is often neglected is that even after converting a joint density model to a discriminative model, the model still constructs a density estimate for $\mathbf{x}$. In this paper we show that this information may be useful, even if the model is inaccurate, for example in predicting missing values of $\mathbf{x}$. We also show that the discriminative joint density models are very close to so-called generalized linear models with random effects. The models operate in the same parameter space, but the generative formulation restricts the space.

Discriminative joint density models allow straightforward generalization to combining different types of measured data: continuous, categorical, or probabilities. In this paper we introduce, as an example, a discriminative joint density model for multinomial data, a discriminative version of the mixture of unigrams model.

---

[3] Joint density modeling minimizes the Kullback-Leibler divergence between the model $p(c, \mathbf{x}|\theta)$ and the "true" model $p(c, x)$,

$$\mathcal{D}_{KL} = \sum p(c, \mathbf{x}) \log \frac{p(c, \mathbf{x})}{p(c, \mathbf{x}|\theta)} = \sum p(c, x) \log \frac{p(c|x)}{p(c|x, \theta)} + \sum p(x) \log \frac{p(x)}{p(x|\theta)} \quad ,$$

where the first term is the conditional likelihood. If the true model is included in the model family, the latter term can be made to vanish, but otherwise, in the case of an incorrect model, it is always nonzero for joint likelihood models. When the true model is not within the model family, the joint likelihood model is thus asymptotically always worse than the conditional likelihood model.

2 Background

2.1 Exponential Family Distributions

An exponential family distribution can always be written in the canonical form

$$p(\mathbf{x}|\theta) = \exp\left(T(\mathbf{x})^T\theta - \log Z(\theta) - \log Y(\mathbf{x})\right) \quad , \tag{1}$$

where the $T(\mathbf{x})$ are the (observed) sufficient statistics, θ the natural parameters, and $\log Z(\theta)$ is the convex normalization term (partition function).

The key definition [7] needed here is the dual parameter μ, [4]

$$\mu = \langle T(\mathbf{x})\rangle_{p(\mathbf{x}|\theta)} = \frac{\partial \log Z}{\partial \theta} \quad . \tag{2}$$

The natural parameters do not in general (with Gaussian being the exception) lie within the same space as the sufficient statistics [8], which complicates their use and interpretation. This is why exponential distributions are usually expressed in terms of dual parameters μ which lie in the same space as the mean of the sufficient statistics (and sometimes they are referred to as expected sufficient statistics, for obvious reasons). The mapping Eq. (2) constrains the allowed values of dual parameters to a plane tangential to the partition function $\log Z(\theta)$. This means that it is always possible to find a θ^* corresponding to the sufficient statistics $T(x)$ (see [7, 8] for more details).

2.2 Generalized Linear Models

In GLMs [1] the dependent variable c is modelled with an exponential family distribution of the form

$$p(c|\mathbf{x}, \mathbf{B}) = \exp\{T(c)^T(\mathbf{B}^T\mathbf{x}) - F(\mathbf{B}^T\mathbf{x}) - \log Y(c)\} \quad . \tag{3}$$

The GLM thus assumes a mapping $\theta = \mathbf{B}^T\mathbf{x}$ to natural parameters. The function $\mu = f(\theta) = \frac{\partial}{\partial\theta}F(\theta)$ then provides a mapping to dual parameters. Here $f(\theta)$ is the inverse of a *link function*. The most often used is the *canonical link* function which is obtained if we select the partition function $f(\theta) = \frac{\partial}{\partial\theta}\log Z(\theta)$.

Generalized Linear Model with Random Effects. It is realistic to assume that there is uncertainty associated with the measured values of $\mathbf{x}$, that is, they contain noise. In statistical modeling the most common assumption is additive noise, $\theta = \mathbf{B}^T\mathbf{x} + \mathbf{Z}\mathbf{u}$, where $\mathbf{Z}$ is assumed to be known and $\mathbf{u}$ is an exponential family noise term [9]. The approach thus makes a probabilistic mapping to natural parameters. Here $\mathbf{B}^T\mathbf{x}$ provides the sufficient statistics for θ. Notice that the approach is still fully discriminative; the distribution of $\mathbf{x}$ is not modelled.

In GLMs with random effects the log-likelihood $\log p(c|\mathbf{x}, \mathbf{B}, \mathbf{u}) + \log p(\mathbf{u}|\mathbf{A})$ is then optimized with respect to β and $\mathbf{u}$ [9], with known values of the noise variance $\mathbf{A}$ (it is determined by $\mathbf{Z}$). See [1, 9] for more detailed descriptions.

[4] For compactness of our formulas, we will denote $\langle T(x)\rangle_{p(x|\theta)} = E_{p(x|\theta)}\{T(x)\}$.

3 Discriminative Joint Density Modeling

In a *discriminative joint density model* the set of variables Y is divided into two classes, $Y = C \cup X$, where the C are the dependent variables over which we want to discriminate and the X are the independent variables. The log-likelihood of the discriminative model is $\log p(C|X,\theta) = \log p(Y|\theta) - \log p(X|\theta)$. Optimization of discriminative generative mixture models is usually done using gradient ascent-based methods (as in this paper). Various EM-type algorithms have also been proposed (see [10] and references therein).

We concentrate here on mixture models, where the identity of the mixture component is a hidden variable. The theory is more general, however. Each value of the hidden variable is associated with a deterministic mapping to a value of the dependent variable c. We will next illustrate the differences between discriminative and ordinary joint density modeling with Linear Discriminant Analysis (LDA).

3.1 Linear Discriminant Analysis

As usually expressed in terms of dual parameters, the a posteriori decision rule of LDA is [11]

$$p(C = j|\mathbf{x}_i) = \frac{\pi(j)p(\mathbf{x}_i|\bar{\mathbf{m}}_j, \mathbf{S})}{\sum_{j'} \pi(j')p(\mathbf{x}_i|\bar{\mathbf{m}}_{j'}, \mathbf{S})} \quad , \tag{4}$$

where $\pi(j)$ is the prior class probability, and $\bar{\mathbf{m}}_j$ denotes the mean of the distribution of $\mathbf{x}$ for the class j. Index i runs over data items, $i \in [1 \ldots N]$. LDA assumes that data from each class is generated from a Gaussian distribution, all of the classes having the same within-class covariance S.

The decision rule (4) is a direct formulation of a discriminative joint density model cost function, with each class being modeled by one Gaussian. Usually, the above equation is not optimized directly. Instead, an asymptotically optimal classifier that models the joint likelihood is obtained by estimating μ_i by class centroids, and $\mathbf{S}$ by the within-class covariance. The joint likelihood solution and the discriminative solution obtained by optimizing Eq. (4) are asymptotically the same if the "true" data distribution follows the assumptions of the LDA model. Otherwise the solutions differ (see Fig.1 for a toy example).

3.2 Log-Linear Regression

As illustrated in the toy example of Figure 1, the best model for classification optimizes $p(c|\mathbf{x}, \beta)$, which in the case of LDA is the a posteriori decision rule. A classic example of a case where $p(c|\mathbf{x}, \beta)$ is optimized directly is the log-linear regression.

In log-linear regression the probability of a class j for a data item $\mathbf{x}_i$ is computed by

$$p(C = j|\mathbf{x}_i, \mathbf{B}) \equiv p_{ji} = \frac{e^{\beta_j^T \mathbf{x}_i}}{\sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i}} \quad , \tag{5}$$

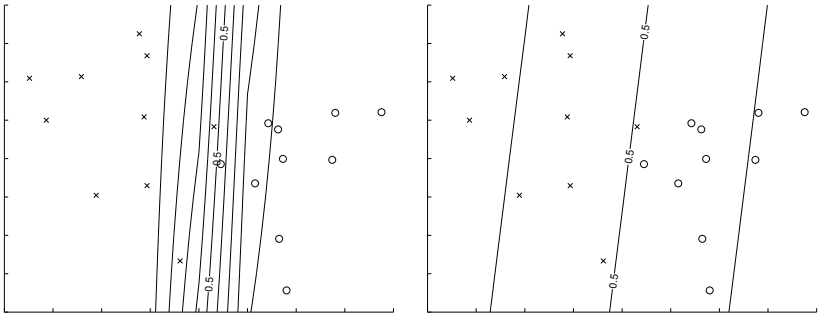

Fig. 1. Difference of class distributions of discriminative and joint density models. Discriminative modeling is optimal for predicting c (Left). In a joint likelihood model the class difference is optimized only implicitly, resulting in softer class borders (Right). In this toy example both models have the same covariance matrix, the within-class covariance, and only the cluster centroids are optimized. The contour plot shows the probability $p(c|x)$ in 0.1 intervals. "X" and "O" denote samples from different classes.

where $\mathbf{x}$ is the vector of independent variables and β_j the vector of coefficients for a given class j. The β_j is constructed to incorporate also a constant term β_{j0} by having one component of $\mathbf{x}$ to be always 1. The β_j form the columns of matrix $\mathbf{B}$. Each observation i can be considered as a draw from a multinomial, and hence the log-likelihood will be

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta(c_i, j) \log p_{ji} \quad , \tag{6}$$

where $\delta(c_i, j)$ picks the class index j corresponding to the class of sample i.

We will next show the relationship between LDA and loglinear models. By inserting Eq. (5) into the log-likelihood (6), we get

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta(c_i, j) \beta_j^T \mathbf{x}_i - \log \left(\sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i} \right) \quad . \tag{7}$$

We may take the constant term β_{j0} out from $\beta_j = [\beta_{j0}\ \beta_{j,1\ldots d}]$,

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta(c_i, j) \left(\beta_{j0} + \beta_{j,1\ldots d}^T \mathbf{x}_i \right) - \log \left(\sum_{j'} e^{\beta_{j'0}} e^{\beta_{j',1\ldots d}^T \mathbf{x}_i} \right) \quad . \tag{8}$$

At this point we insert prior information into the model family: we require that the β_{j0} and $\mathbf{x}$ come from exponential family distributions. We first require that β_{j0} comes from a multinomial distribution by reparameterizing $\beta_{j0} \rightarrow \log \pi(j) - \log \sum_j \pi(j)$ (here we in effect add a constraint that $\sum_j \pi(j) = 1$). The term $\beta_{j,1\ldots d}^T \mathbf{x}_i$ can be interpreted to be $\log p(\mathbf{x}_i | \beta_{j,1\ldots d})$ without the normalization

term. We can restrict $\mathbf{x}$ to an exponential family model by reparameterizing $\beta_{j,1\ldots d}^T \mathbf{x}_i \rightarrow \beta_{j,1\ldots d}^T \mathbf{x}_i - \log Z(\beta_{j,1\ldots d}) - \log Y(\mathbf{x}_i)$. The $\beta_{j,1\ldots d}$ then form the natural parameters and $\mathbf{x}_i$ the sufficient statistics of the model.

Using Equation (1), we get

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N}\sum_{j=1}^{C} \delta(c_i, j)\left(\log \pi(j) + \log p(\mathbf{x}_i|\beta_{j,1\ldots d})\right) - \log\left(\sum_{j'} \pi(j') p(\mathbf{x}_i|\beta_{j',1\ldots d})\right) \\ &= \sum_{i=1}^{N} \log \frac{\prod_j \left(\pi(j) p(\mathbf{x}_i|\beta_{j,1\ldots d})\right)^{\delta(c_i,j)}}{\sum_{j'} \pi(j') p(\mathbf{x}_i|\beta_{j',1\ldots d})} \quad .\end{aligned}$$

This is the same as LDA in Eq. (4) if the $p(\mathbf{x}_i|\beta_{j,1\ldots d})$ are Gaussian.

Notice that the constraint that $\mathbf{x}$ can be modelled by an exponential family distribution restricts the parameter space of $\beta_{j,1\ldots d}$ through $\log Z(\beta_{j,1\ldots d})$.[5] As an example, for multinomial distributions this effectively removes one degree of freedom, since $\sum \mu = 1$. Note additionally that the discriminative joint density model prefers values of β which are close to the θ^* corresponding to the mean of the observed sufficient statistics of $\mathbf{x}$.

4 General Description of Discriminative Joint Density Models

We will now formalize a general description of the discriminative joint density model. We define a model that generates the observed (categorical) values c and the associated measurements $\mathbf{x}$. Each measurement $\mathbf{x}_i$ consists of S different kinds of data sources indexed by s, each modelled with an appropriate exponential family distribution. Our goal is to optimize $P(c|X, \theta)$, where $\theta = \{\pi, \beta\}$ denote all parameters of the model. We assume that X can be modelled using an exponential family distribution, given a mixture component l. The information $\mathbf{x}$ carries about c is therefore visible also in the sufficient statistics of X, and thus the parameters of the generative distributions. The model can be optimized for discriminating the classes by maximizing the conditional likelihood

$$p(c_i|\mathbf{x}_i, \theta) = \frac{\prod_k \left(\sum_{l \in \mathcal{C}_k} p(l, \mathbf{x}_i|\beta_l, \pi(l))\right)^{\delta(c_i,k)}}{\sum_{l'} p(l', \mathbf{x}_i|\beta_{l'}, \pi(l'))} \quad , \tag{9}$$

where l indexes the mixture component, and $\mathcal{C}_k$ is the set of components associated with class k. $\pi(l)$ is the probability that the data was generated from mixture component l, and β_l are the parameters of the component l. See also Figure 2.

The observed variables of our model are the classification C and the associated independent variables X_s. The parameters of the model are given by

[5] Logistic regression, on the other hand, assumes that the β are independent with values allowed to vary over the whole real-valued space.

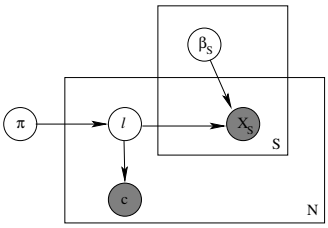

Fig. 2. A graphical model of the discriminative joint density model. Here l is the index of the distribution that is used to predict class c. The grey circles indicate observed values. S is the number of data sources, and N the number of data items.

$\theta = \{\pi, \beta_{1\ldots S}^{1\cdots L}\}$. Notice that the generative models are the same for the discriminative and joint likelihood models. The difference is in the optimization.

A benefit of the discriminative joint density formulation, compared to alternative discriminative models, is that the model (and thus logistic regression) is easy to extend into cases where $\mathbf{x}$ is better modelled by a mixture of exponential family models. The generative formulation also makes it simple to model several independent variables and different forms of data, such as multinomial or probability distributions [12]. Besides giving class predictions, the parameters of the discriminative joint density models are directly interpretable in terms of sufficient statistics for $\mathbf{x}$.

4.1 Generative Model for Generalized Linear Models

The generative formulation can be easily extended to the GLM model class, of which the log-linear model (see Section 3.2) is a special case. For simplicity, we will assume that exactly one mixture component j corresponds to each class label c. For convenience we will drop out the index i from $\mathbf{x}_i, c_i$ in the following.

We begin with the objective function of discriminative joint density models, Eq. (9), which can also be written as

$$p(c|\mathbf{x}, \beta) = \exp\{\delta(c, j) \log p(\mathbf{x}, j|\beta_j, \pi(j)) - \log \sum_{j'} p(\mathbf{x}, j'|\beta_{j'}, \pi(j'))\} \quad . \quad (10)$$

By comparing this form with (1), we notice that the form corresponds to a multinomial distribution with natural parameters $\theta_j = \log p(\mathbf{x}, j|\beta_j, \pi(j))$, sufficient statistics $T(c) = \delta(c, j)$, and with $\log Z(\theta) = \log \sum_{j'} p(\mathbf{x}, j'|\beta_{j'}, \pi(j'))\} = \log Z(p(\mathbf{x}|\beta, \pi))$. Since we pick one class for each $\mathbf{x}$, the $\log Y(c)$ is zero.

By writing θ in an exponential family notation, we get

$$\theta_j = \log p(\mathbf{x}, j|\beta_j, \pi(j)) = T(\mathbf{x})^T \beta_j - \log Z(\beta_j) - \log Y(\mathbf{x}) + \log \pi(j) - \log Z(\pi) \; . \quad (11)$$

The $\log Y(\mathbf{x})$-term can be left out, since it is the same for all components j. By inserting Eq. (11) into Eq. (10), we get

$$p(c|\mathbf{x}, \beta) = \exp\{T(c)^T \left(\mathbf{B}^T T(\mathbf{x}) - \log Z(\beta) + \log \pi - \log Z(\pi)\right) - \log Z(p(\mathbf{x}|\beta))\}. \quad (12)$$

The π and $\log Z(\pi)$ can be incorporated into the matrix $\mathbf{B}$, similarly to the log-linear case. The vector $\log Z(\beta)$ consists of components $\log Z(\beta_j)$.

Now, when the generative model and the GLM have been expressed in the exponential family notation in Equations (12) and (3), respectively, we will point out their difference. In case of multinomials considered in this paper, the $Y(c)$ in (3) is zero because of the form of the sufficient statistics. Of the remaining terms within the exponent, the last one in both models is the normalization term. The essential difference then is the term $\log Z(\beta_j)$ in (12). In case of multinomial distribution it removes one degree of freedom in the model. This can be shown by adding a displacement λ to each component of β_j, which does not change the predictions of the model (12). GLM, in contrast, does not have such a restriction.

The generative model in effect introduces prior information into GLMs: assuming that the generative model for $\mathbf{x}$ is (nearly) correct, we can restrict the (effective) parameter space of β. The restriction provides an additional benefit, since by mapping the parameters to their dual parameters (through $\log Z(\beta)$), the values of β can be interpreted in terms of sufficient statistics of $\mathbf{x}$.

The model is very similar to GLMs (with random effects), since both models define a probabilistic mapping to θ. However, in discriminative joint density models the uncertainty is defined for values of $\mathbf{x}$, whereas in GLMs with random effects the uncertainty is defined for θ. The discriminative joint density models, however, have an additional benefit: By expressing the noise terms for individual $\mathbf{x}$, we form a generative distribution for $\mathbf{x}$.

4.2 Connection to Maximum Entropy Discrimination

In maximum entropy discrimination (MED) [13], *discriminative functions* of the form $\mathcal{L}(X|\theta) = \log \frac{p_+}{p_-}$ are optimized. The p_+, p_- denote probabilistic models for the class $+$ and $-$, respectively. In contrast, the discriminative joint density modeling cost function can be expressed by

$$\frac{p_+}{p_+ + p_-} = \frac{1}{1 + \exp\{-\log \frac{p_+}{p_-}\}} = \frac{1}{1 + \exp\{-\mathcal{L}(X|\theta)\}} \quad . \tag{13}$$

The cost function thus is a monotonic (sigmoid) transformation of the MED objective function.

The main advantage of the discriminative joint density modeling cost function over MED is that the output is the probability of the corresponding class, thus expressing directly the level of uncertainty in class prediction. Generalization to the case of several classes is also simpler and more straightforward to implement.

4.3 Missing Data

It is of interest to know whether the estimate $p(c|\mathbf{x}, \theta)$ can benefit from data where the $\mathbf{x}$ is incomplete for some data items. Let us denote vectors with missing

values by $\mathbf{x} = [\mathbf{y}\ \mathbf{z}]$, where $\mathbf{y}$ is the missing data and $\mathbf{z}$ the known components. The conditional log-likelihood with missing data can then be written as

$$\mathcal{L} = \sum_{i \in D_{full}} \log p(c_i|\mathbf{x}_i, \theta) + \sum_{i \in D_{miss}} \int p(\mathbf{y}|c_i, \mathbf{z}_i, \theta) \log p(c_i|\mathbf{y}, \mathbf{z}_i, \theta) d\mathbf{y} \quad , \quad (14)$$

where we denote by D_{full} the data set with all entries known, and by D_{miss} the data with missing entries. In order to infer the value for missing data, we need to make a distributional assumption for $\mathbf{y}$. A feasible one is $p(\mathbf{y}|c_i, \theta)$ used in the generative model. If the data really has been generated from the model family this is the correct assumption, but in the real world the performance depends on how close the model family is to the "true" generative distribution.

Practical Implementation. There are several possibilities to optimize Eq. (14). We now present a simple approach that makes computations tractable by constructing a lower bound for $\mathcal{L}_{miss}$, the cost function for the missing part of the data. For discriminative joint density models it can be written as

$$\mathcal{L}_{miss} = \sum_{i \in D_{miss}} \langle \log p(c_i, \mathbf{y}, \mathbf{z}_i|\theta) \rangle_{p(\mathbf{y}|c_i, \mathbf{z}_i, \theta)} - \langle \log \sum_j p(j, \mathbf{y}, \mathbf{z}_i|\theta) \rangle_{p(\mathbf{y}|c_i, \mathbf{z}_i, \theta)} \ . \quad (15)$$

The latter term can be upper bounded (and thus we obtain a lower bound for $\mathcal{L}_{miss}$) by applying Jensen inequality

$$\langle \log \sum_j p(j, \mathbf{y}, \mathbf{z}_i|\theta) \rangle_{p(\mathbf{y}|c_i, \mathbf{z}_i, \theta)} \leq \log \sum_j \langle p(j, \mathbf{y}, \mathbf{z}_i|\theta) \rangle_{p(\mathbf{y}|c_i, \mathbf{z}_i, \theta)} \leq \log \sum_j p(j, \mathbf{z}_i|\theta),$$

where the last expression follows from $\langle p(\mathbf{y}|j, \mathbf{z}_i, \theta) \rangle_{p(\mathbf{y}|c_i, \mathbf{z}_i, \theta)} \leq 1$. A simple lower bound of the cost function for missing data then follows

$$\mathcal{L}_{miss} \geq \sum_i \langle \log p(c_i, \mathbf{y}, \mathbf{z}_i|\theta) \rangle_{p(\mathbf{y}|c_i, z_i, \theta)} - \sum_j \log p(j, \mathbf{z}_i|\theta) \quad , \quad (16)$$

where the missing values $\mathbf{y}$ are replaced by their expectation under $p(\mathbf{y}|c_i, z_i, \theta)$ in the first term, and omitted in the second term.

4.4 Discriminative Document Modeling

The mixture of unigrams model [14] is a hidden variable model that generates word counts for documents. The model assumes that each document is generated from a mixture of M hidden "topics", $\sum_{j=1}^{M} \pi_j p(\mathbf{x}_i|\beta_j)$, where j is the index of the topic, and β_j the multinomial parameters that generate words from the topic. The vector $\mathbf{x}_i$ is the observed word counts for document i, and π_j the probability of generating the words from the topic j. In its simplest form with one topic per class the model is a naive Bayes classifier.

In a discriminative mixture of unigrams the document vector is generated from a mixture of topics (multinomials), where each class is assigned a subset of topics. In this paper we will illustrate the functionality of the discriminative mixture of unigrams in two cases: either with one or five topic vectors per class.

5 Experiments

We used the Reuters data set [15]. A subset of 4000 documents from four categories was selected, 1000 from each category. The categories were: Corporate-Industrial (CCAT), Economics and Economic Indicators (ECAT), Government and Social (GCAT), and Securities and Commodities Trading and Markets (MCAT). Each of the selected documents was classified to only one of the four classes. The words that occurred less than 200 times in the whole subset were left out, thus leaving 1952 words. The data set was then split into equal-sized training and test sets.

The second data set was the MNIST data[6]. The data consists of gray level images of handwritten digits. The data was thresholded to ones and zeros with a threshold gray level value of 128 (with a maximum of 255) before evaluating the models. The training and test data sets each consisted of 10000 samples, each sample being a binary image of 784 pixels.

A discriminative mixture of unigrams model (d-MUM) with one and five components was applied. Reference methods included the naive Bayes classifier, loglinear regression, k-means algorithm (where each class was modelled by its centroid), and k-nearest neighbor search (k-NN), where the size of the neighborhood was chosen by dividing the full training set to training and validation sets.[7] The k-means and k-NN algorithms were computed using dot product and Hellinger distances. The classification accuracies for the test data set are reported in Table 1. With the Reuters data the performances of the loglinear model and

Table 1. Classification accuracies for the test sets. Comparisons $^{(1),(2),(3)}$: significant ($p < 0.01$) difference (McNemar's test).

Method	Accuracy (%)	
	Reuters	MNIST
k-means	79.9	64.1
k-means (Hellinger)	81.9	76.0
k-NN	(5-nn) 74.6	(9-nn) 84.8
k-NN (Hellinger)	(5-nn) 86.9 $^{(1)}$	(5-nn) 94.9
naive Bayes	59.0	68.9
loglinear	92.2	90.9 $^{(2)(3)}$
d-MUM 1 component	92.5 $^{(1)}$	90.5 $^{(2)}$
d-MUM 5 components	92.3	93.2 $^{(3)}$

d-MUM are roughly equal. With MNIST data, the loglinear model is better than

[6] Available at `http://yann.lecun.com/exdb/mnist/`

[7] The computational complexity of k-NN is not comparable to the other methods, since the method computes pairwise distances between every data point pair, whereas in the other methods only C "prototypes" are used.

1-component d-MUM, but loses to 5-component d-MUM. Both models clearly outperform the joint likelihood (naive Bayes) model.

In a second experiment the MNIST teaching data was corrupted by randomly replacing pixels with missing values. The experiment was run for 10, 30, 50, and 75 % missing data. A baseline comparison method was logistic regression where missing values were imputed by the mean of the known pixel values for the given pixel and class. We also compared to the current state-of-the-art, k-NN imputation which has been reported to outperform several other methods [16].

The discriminative MUM compares favourably to the k-NN imputation with missing values computed based on the 10 nearest neighbors. Besides being more accurate, our method is considerably faster, since k-NN imputation is $\mathcal{O}(N^2)$, where N is the amount of samples[8]. This is an additional cost, since the optimization durations for the loglinear model and discriminative MUM (with missing value imputation) are roughly equal.

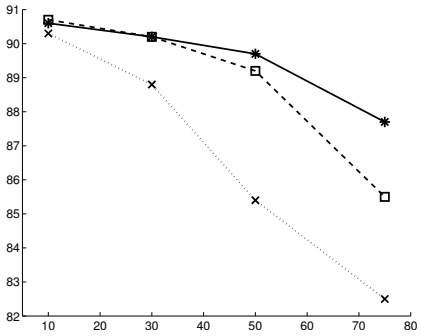

Fig. 3. Performance with missing data. The performance of discriminative MUM (solid line) compared to logistic regression with k-NN imputation (dashed line) and imputation by the mean of the class (dotted line). Horizontal axis: Percentage of missing data. Vertical axis: Classification accuracy (%). The difference between k-NN imputation and d-MUM is significant with 75 %, and mildly significant (p=0.033) with 50% missing data.

6 Discussion

The aim of this paper has been to set the stage for further contributions on discriminative joint density models. Several theoretical connections were explored. We have also shown that the paradigm can be easily applied to discriminative document modeling with a simple case of mixture of unigrams model introduced in this paper, and that the generative mechanism for $\mathbf{x}$ in discriminative joint density models still contains useful information for example in predicting missing values.

[8] For computational reasons, we divided the data set to blocks of 1000 samples and then imputed the missing values. This took more than 12 hours for each data set.

Acknowledgements. This work was supported in part by Academy of Finland, decision 79017, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. All rights are reserved because of other commitments.

References

1. McCullagh, P., Nelder, J.A.: Generalized Linear Models. 2nd edn. CRC Press (1990)
2. Rubinstein, Y.D., Hastie, T.: Discriminative vs informative learning. In Heckerman, D., Mannila, H., Pregibon, D., Uthurusamy, R., eds.: Proc. ACM KDD. AAAI Press (1997) 49–53
3. Kontkanen, P., Myllymäki, P., Tirri, H.: Classifier learning with supervised marginal likelihood. In Breese, J., Koller, D., eds.: Proc. UAI'01, Morgan Kaufmann Publishers (2001) 277–284
4. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in NIPS 14. MIT Press, Cambridge, MA (2002) 841–848
5. Nádas, A., Nahamoo, D., Picheny, M.A.: On a model-robust training method for speech recognition. IEEE Tr. on Acoustics, Speech, and Signal Processing **39** (1988) 1432–1436
6. Povey, D., Woodland, P., Gales, M.: Discriminative MAP for acoustic model adaptation. In: Proc. IEEE ICASSP'03. Volume 1. (2003) 312–315
7. Buntine, W.: Variational extensions to EM and multinomial PCA. In Elomaa, T., Mannila, H., Toivonen, H., eds.: Proc. ECML-2002. Springer-Verlag (2002) 23–34
8. Efron, B.: The geometry of exponential families. The Annals of Statistics **6** (1978) 362–376
9. Schall, R.: Estimation in generalized linear models with random effects. Biometrika **78** (1991) 719–727
10. Salojärvi, J., Puolamäki, K., Kaski, S.: Expectation maximization algorithms for conditional likelihoods. In: Proc. ICML-2005. (2005) in press.
11. Sharma, S.: Applied Multivariate Techniques. John Wiley & Sons, Inc. (1996)
12. Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., Kaski, S.: Combining eye movements and collaborative filtering for proactive information retrieval. In: Proc. SIGIR 2005. (2005) in press.
13. Jaakkola, T.S., Meila, M., Jebara, T.: Maximum entropy discrimination. In Solla, S.A., Leen, T.K., Müller, K.R., eds.: Advances in NIPS 12. MIT Press, Cambridge, MA (2000) 470–476
14. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning **39** (2000) 103–134
15. Lewis, D.D., Yang, Y., Rose, T., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5** (2004) 361–397
16. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics **17** (2001) 520–525