Multi-Conditional Learning: Generative/Discriminative Training for Clustering and Classification

Andrew McCallum, Chris Pal, Greg Druck and Xuerui Wang
Department of Computer Science
140 Governors Drive
University of Massachusetts
Amherst, MA 01003–9264
{mccallum,pal,gdruck,xuerui}@cs.umass.edu

Abstract

This paper presents multi-conditional learning (MCL), a training criterion based on a product of multiple conditional likelihoods. When combining the traditional conditional probability of "label given input" with a generative probability of "input given label" the later acts as a surprisingly effective regularizer. When applied to models with latent variables, MCL combines the structure-discovery capabilities of generative topic models, such as latent Dirichlet allocation and the exponential family harmonium, with the accuracy and robustness of discriminative classifiers, such as logistic regression and conditional random fields. We present results on several standard text data sets showing significant reductions in classification error due to MCL regularization, and substantial gains in precision and recall due to the latent structure discovered under MCL.

Introduction

Conditional-probability training, in the form of maximum entropy classifiers (Berger et al., 1996) and conditional random fields (CRFs) (Lafferty et al., 2001; Sutton & McCallum, 2006), has had dramatic and growing impact on natural language processing, information retrieval, computer vision, bioinformatics, and other related fields. However, discriminative models tend to overfit the training data, and a prior on parameters typically provides limited relief. In fact, it has been shown that in some cases generative naïve Bayes classifiers provide higher accuracy than conditional maximum entropy classifiers (Ng & Jordan, 2002). We thus consider alternative training criteria with reduced reliance on parameter priors, which also combine generative and discriminative learning.

This paper presents *multi-conditional learning*, a family of parameter estimation objective functions based on a product of multiple conditional likelihoods. In one configuration of this approach, the objective function is the (weighted) product of the "discriminative" probability of label given input, and the "generative" probability of the input given label. The former aims to find a good decision boundary, the later aims to model the density of the input, and the single set of parameters in our naïve-Bayes-structured model thus strives for both. All regularizers provide some additional constraints on parameter estimation. Our experimental results on a variety of standard text data sets show that this density-estimation constraint is a more effective regularizer than "shrinkage toward zero," which is the basis of traditional regularizers, such as the Gaussian prior—reducing error by nearly 50% in some cases. As well as improving accuracy, the inclusion of a density estimation criterion helps improve confidence prediction.

In addition to simple conditional models, there has been growing interest in conditionally-trained models with latent variables (Jebara & Pentland, 1998; McCallum et al., 2005; Quattoni et al., 2004). Simultaneously there is immense interested in generative "topic models," such as latent Dirichlet allocation, and its progeny, as well as their undirected analogues, including the harmonium models (Welling et al., 2005; Xing et al., 2005; Smolensky, 1986).

In this paper we also demonstrate multi-conditional learning applied to latent-variable models. MCL discovers a latent space projection that captures not only the co-occurrence of features in input (as in generative models), but also provides the ability to accurately predict designated outputs (as in discriminative models). We find that MCL is more robust than the conditional criterion alone, while also being more purposeful than generative latent variable models. On the document retrieval task introduced in Welling et al. (2005), we find that MCL more than doubles precision and recall in comparison with the generative harmonium.

In latent variable models, MCL can be seen as a form of semi-supervised clustering—with the flexibility to operate on relational, structured, CRF-like models in a principled way. MCL here aims to combine the strengths of CRFs (handling auto-correlation and non-independent input features in making predictions), with the strengths of topic models (discovering co-occurrence patterns and useful latent projections). This paper sets the stage for various interesting future work in multi-conditional learning. Many configurations of multi-conditional learning are possible, including ones with more than two conditional probabilities. For example, transfer learning could naturally be configured as the product of conditional probabilities for the labels of each task, with some latent variables and parameters shared. Semi-supervised learning could be configured as the product of conditional probabilities for predicting the label, as well as predicting each input given the others. These configurations are the subject of ongoing work.

Multi-Conditional Learning and MRFs

In the following exposition we first present the general framework of multi-conditional learning. We then derive the equations used for multi-conditional learning in several structured Markov Random Field (MRF) models. We introduce discrete hidden (sub-class) variables into naïve MRF models, creating multi-conditional mixtures, and discuss how multi-conditional methods are derived. We then construct binary word occurrence models coupled with hidden *continuous* variables, as in the exponential family harmonium, demonstrating the advantages of multi-conditional learning for these models also.

The MCL Framework

Consider a data set consisting of $i = 1, \ldots, N$ instances. We will construct probabilistic models consisting of discrete observed random variables $\{x\}$, discrete hidden variables $\{z\}$ and continuous hidden variables $\mathbf{z}$. Denote an outcome of a random variable as $\tilde{x}$. Define $j = 1, \ldots, N_s$ pairs of disjoint subsets of observations $\{\tilde{x}_A\}_{ij}$ and $\{\tilde{x}_B\}_{ij}$, where our indices denote the ith instance of the variables in subset j. We will construct a multi-conditional objective by taking the product of different conditional probabilities involving these subsets and we will use α_j to weight the contributions of the different conditionals. Using these definitions the optimal parameter settings under our multi-conditional criterion are given by

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i,j} \sum_{\{z\}_{ij}} \int P\big(\{\{\tilde{x}_A\}, \{z\}, \mathbf{z}\}_{ij} | \{\tilde{x}_B\}_{ij}; \boldsymbol{\theta}\big)^{\alpha_j} \mathrm{d}\mathbf{z}_{ij}, \tag{1}$$

where we derive these marginal conditional likelihoods from a single underlying joint probability model with parameters $\boldsymbol{\theta}$. Our underlying joint probability model may itself be normalized locally, globally or using some combination of the two.

For the experiments in this paper we will partition observed variables into a set of "labels" $\mathbf{y}$ and a set of "features" $\mathbf{x}$. We define two pairs of subsets: $\{x_A, x_B\}_1 = \{\mathbf{y}, \mathbf{x}\}$ and $\{x_A, x_B\}_2 = \{\mathbf{x}, \mathbf{y}\}$. We then construct multi-conditional objective functions $\mathcal{L}_{MC}$ with the following form

$$\begin{aligned} \mathcal{L}_{MC} &= \log\left(P(\mathbf{y}|\mathbf{x})^{\alpha} P(\mathbf{x}|\mathbf{y})^{\beta}\right) \\ &= \alpha \mathcal{L}_{y|x}(\boldsymbol{\theta}) + \beta \mathcal{L}_{x|y}(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

In this configuration one can think of our objective as having a generative component $P(\mathbf{x}|\mathbf{y})$ and a discriminative component $P(\mathbf{y}|\mathbf{x})$. Another attractive definition using two pairs is: $\{x_A, x_B\}_1 = \{\mathbf{y}, \mathbf{x}\}$ and $\{x_A, x_B\}_2 = \{\mathbf{x}, \emptyset\}$, giving rise to objectives of the following form

$$\mathcal{L} = \log(P(\mathbf{y}|\mathbf{x})^{\alpha} P(\mathbf{x})^{\beta}), \tag{3}$$

which represents a way of restructuring a joint likelihood to concentrate modeling power on a conditional distribution of interest. This objective is similar to the approach advocated in Minka (2005).

Naïve MRFs for Documents

The graphical descriptions of the naïve Bayes model for text documents (Nigam et al., 2000) and the multinomial logistic regression or maximum entropy (Berger et al., 1996) model can be written with similar naïve graphical structures. Here we consider naïve MRFs which can also be represented by a similar graphical structure but define a joint distribution in terms of unnormalized potential functions.

Consider data $\mathcal{D} = \{(\tilde{y}_n, \tilde{x}_{j,n}); n = 1, \ldots, N, j = 1 \ldots M_n\}$ where there are N instances and within each instance there are M_n realizations of discrete random variables $\{x\}$. We will use y_n to denote a single discrete random variable for a class label. Model parameters are denoted by $\boldsymbol{\theta}$. For a collection of N documents we thus have M_n word events for each document. The joint distribution of the data can be modeled using a set of naïve MRFs, one for each observation such that

$$P(x_1, \ldots, x_{M_n}, y | \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}} \phi(y|\boldsymbol{\theta}_y) \prod_{j=1}^{M_n} \phi(x_j, y | \boldsymbol{\theta}_{x,y}) \tag{4}$$

where

$$\mathcal{Z} = \sum_{y} \sum_{x_1} \cdots \sum_{x_{M_n}} \phi(y|\boldsymbol{\theta}_y) \prod_{j=1}^{M_n} \phi(x_j, y|\boldsymbol{\theta}_{x,y}). \tag{5}$$

If we define potential functions $\phi(\cdot)$ to consist of exponentiated linear functions of *multinomial* variables (sparse vectors with a single 1 in one of the dimensions), $\mathbf{y}$ for labels and $\mathbf{w}_j$ for each word, a naïve MRF can be written as

$$P(\mathbf{y}, \{\mathbf{w}\}) = \frac{1}{\mathcal{Z}} \exp\left(\mathbf{y}^T \boldsymbol{\theta}_y + \mathbf{y}^T \boldsymbol{\theta}_{x,y}^T \sum_{j=1}^{M_n} \mathbf{w}_j\right). \tag{6}$$

To simplify our presentation, consider now combining our multinomial word variables $\{\mathbf{w}\}$ such that $\mathbf{x} = [\sum_{j=1}^{M_n} \mathbf{w}_j; 1]$. One can also combine $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_{x,y}$ into $\boldsymbol{\theta}$ such that

$$P(\mathbf{y}, \mathbf{x}) = \frac{1}{\mathcal{Z}} \exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x}) \tag{7}$$

Under this model, to optimize $\mathcal{L}_{MC}$ from (2) we have

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x})}{\sum_{\mathbf{y}} \exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x})} \text{ and } P(\mathbf{x}|\mathbf{y}) = \frac{\exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x})}{Z(\mathbf{y})} \tag{8}$$

where

$$Z(\mathbf{y}) = \sum_{\mathbf{w}_1} \cdots \sum_{\mathbf{w}_{M_n}} \prod_{j=1}^{M_n} \exp(\mathbf{y}^T \boldsymbol{\theta}_{x,y}^T \mathbf{w}_j) \exp(\mathbf{y}^T \boldsymbol{\theta}_y). \tag{9}$$

The gradients of the log conditional likelihoods contained in our objective can then be computed using:

$$\begin{aligned} \nabla \mathcal{L}_{y|x}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \left(\mathbf{x}_n \mathbf{y}_n^T - \frac{\sum_{\mathbf{y}} \exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x}_n) \mathbf{x}_n \mathbf{y}^T}{\sum_{\mathbf{y}} \exp(\mathbf{y}^T \boldsymbol{\theta}^T \mathbf{x}_n)} \right) \\ &= N\Big(\langle \mathbf{x}\mathbf{y}^T \rangle_{\tilde{P}(\mathbf{x},\mathbf{y})} - \langle \langle \mathbf{x}\mathbf{y}^T \rangle_{P(\mathbf{y}|\mathbf{x})} \rangle_{\tilde{P}(\mathbf{x})} \Big) \end{aligned} \tag{10}$$

where $\langle \cdot \rangle_{P(\mathbf{x})}$ denotes the expectation with respect to distribution $P(\mathbf{x})$ and we use $\tilde{P}(\mathbf{x})$ to denote the empirical distribution of the data, the distribution obtained placing a delta

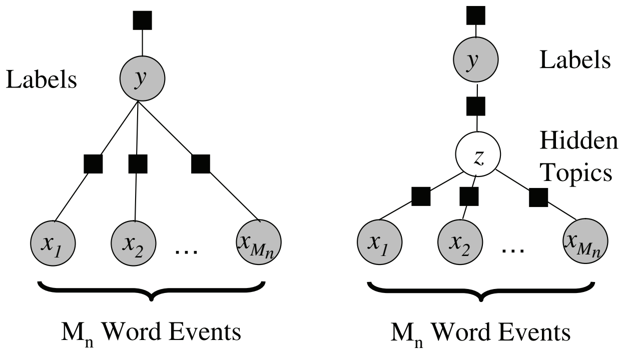

Figure 1: (Left) A factor graph (Kschischang et al., 2001) for a naïve MRF. (Right) A factor graph for a mixture of naïve MRFs. In these models each word occurrence is a draw from a discrete random variable; there are M_n random variables in document n.

function on each data point and normalized by N. To compute $\nabla\mathcal{L}_{x|y}(\boldsymbol{\theta}_{x,y})$, we observe that

$$P(\mathbf{x}|\mathbf{y}) = \prod_{j=1}^{M_n} P(\mathbf{w}_j|\mathbf{y}) = \prod_{j=1}^{M_n} \left(\frac{\exp(\mathbf{y}^T \boldsymbol{\theta}_{x,y}^T \mathbf{w}_j)}{\sum_{\mathbf{w}_j} \exp(\mathbf{y}^T \boldsymbol{\theta}_{x,y}^T \mathbf{w}_j)} \right), \tag{11}$$

and therefore

$$\nabla\mathcal{L}_{x|y}(\boldsymbol{\theta}_{x,y}) = \sum_{n=1}^{N} \sum_{j=1}^{M_n} \Big(\tilde{\mathbf{w}}_{j,n} \tilde{\mathbf{y}}_n^T - \mathbf{w}_{j,n} \tilde{\mathbf{y}}_n^T P(\mathbf{w}_{j,n}|\tilde{\mathbf{y}}_n) \Big). \tag{12}$$

Mixtures of Naïve MRFs

We can extend the basic naïve MRF model shown in Figure 1 (Left) by adding a hidden subclass variable as illustrated (Right). In a mixture of naïve MRFs the joint distribution of the data for each observation can be modeled using

$$P(\{x\}, y, z|\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}} \phi(y|\boldsymbol{\theta}_y) \phi(y, z|\boldsymbol{\theta}_{y,z}) \prod_{j=1}^{M_n} \phi(x_j, z|\boldsymbol{\theta}_{x,z}), \tag{13}$$

where the $\phi(y, z|\boldsymbol{\theta}_{y,z})$ potential encodes a sparse compatibility function relating labels or classes to a subset of states of the hidden discrete variable z.

To optimize a mixture of naïve MRFs, we use the expected gradient algorithm (Salakhutdinov et al., 2003). In this model we can compute the gradient of the complete log likelihood and this gradient decomposes with respect to our expectation such that the following computation can be efficiently performed,

$$\begin{aligned}\nabla\mathcal{L}_{x|y}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}} \ln P(\{x\}|y; \boldsymbol{\theta}) \\ &= \sum_{z} P(z|\{x\}, y; \boldsymbol{\theta}) \frac{\partial}{\partial\boldsymbol{\theta}} \ln P(\{x\}, z|y; \boldsymbol{\theta}).\end{aligned} \tag{14}$$

For example, the gradient for the "weights" λ_{x_e,z_s} comprising the elements of the potential function parameters $\boldsymbol{\theta}_{x,z}$ are computed from

$$\begin{aligned}\frac{\partial\mathcal{L}_{x|y}(\boldsymbol{\theta})}{\partial\lambda_{x_e,z_s}} = \sum_{n=1}^{N} \sum_{j=1}^{M_n} \Bigg[& \sum_{z_n} P(z_n|\{\tilde{x}\}_n, \tilde{y}_n; \boldsymbol{\theta}) f_{x_e,z_s}(\tilde{x}_{j,n}, z_n) \\ & - \sum_{z_n} \sum_{\{x\}_n} P(\{x\}_n, z_n|\tilde{y}_n; \boldsymbol{\theta}) f_{x_e,z_s}(x_{j,n}, z_n) \Bigg],\end{aligned} \tag{15}$$

where $f_{x_e,z_s}(x, z)$ are binary feature functions evaluating to one when the state of $x = x_e$ *and* the state of $z = z_s$. The updates for the potentials function parameters using $\mathcal{L}_{y|x}$ take a form similar to the standard "maximum entropy" gradient computations, augmented with a hidden variable. We term mixture models trained my multi-conditional learning *multi-conditional mixtures* (MCM).

Harmonium Structured Models

A harmonium model (Smolensky, 1986) is a two layer Markov Random Field (MRF) consisting of observed variables and hidden variables. Like all MRFs, the model we present here will be defined in terms of a globally normalized product of (unnormalized) potential functions defined upon subsets of variables. A harmonium can also be described as a type of restricted Boltzmann machine (Hinton, 2002). In the following we present a new type of exponential family multi-attribute harmonium, extending the models used in Welling et al. (2005) and the dual-wing harmonium work of Xing et al. (2005).

Our exponential family harmonium structured model can be written as

$$\begin{aligned}P(\mathbf{x}, \mathbf{z}|\boldsymbol{\Theta}) = \exp\Bigg\{ & \sum_i \boldsymbol{\theta}_i^T \boldsymbol{f}_i(\mathbf{x}_i) + \sum_j \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\mathbf{z}_j) \\ & + \sum_i \sum_j \boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{z}_j) - A(\boldsymbol{\Theta}) \Bigg\},\end{aligned} \tag{16}$$

where $\mathbf{z}$ is a vector of continuous valued hidden variables, $\mathbf{x}$ is a vector of observations, $\boldsymbol{\theta}_i$ represents parameter vectors (or weights), $\boldsymbol{\theta}_{ij}$ represents a parameter vector on a cross product of states, $\boldsymbol{f}_i$ denotes feature functions, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j\}$ is the set of all parameters and A is the log-partition function or normalization constant. A harmonium model factorizes the third term of (16) into $\boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{z}_j) = \boldsymbol{f}_i(\mathbf{x}_i)^T \mathbf{W}_{ij}^T \boldsymbol{f}_j(\mathbf{z}_j)$, where $\mathbf{W}_{ij}^T$ is a parameter matrix with dimensions $a \times b$, i.e., with rows equal to the number of states of $\boldsymbol{f}_i(\mathbf{x}_i)$ and columns equal to the number of states of $\boldsymbol{f}_j(\mathbf{z}_j)$. In the models we construct here we will use *binary* word occurrence vectors that have dimension M_v, the size of our vocabulary. This is in contrast to our models in the previous section where we had a different number of discrete word events M_n for each document n. We will denote one of the observed input variables x_d as a discrete label denoted as y in Figure 2.

Figure 2 illustrates a multi-attribute harmonium model as a factor graph. A harmonium represents the factorization of a joint distribution for observed and hidden variables using

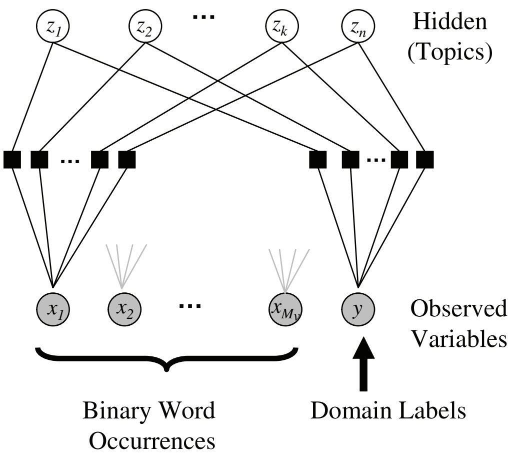

Figure 2: A factor graph for a multi-attribute harmonium model or two layer MRF.

a globally normalized product of local functions. In our experiments here we shall use the harmonium's factorization structure to define an MRF and we will then define sets of marginal conditionals distributions of some *observed* variables given others that are of particular interest so as to form our multi-conditional objective.

Importantly, using a globally normalized joint distribution with this construction it is also possible to derive two consistent conditional models, one for hidden variables given observed variables and one for observed variables given hidden variables (Welling et al., 2005). The conditional distributions defined by these models can also be used to implement sampling schemes for various probabilities in the underlying joint model. However, it is important to remember that the original model parameterization is not defined in terms of these conditional distributions. In our experiments below we use a joint model with a form defined by (16) with $\mathbf{W}^T = [\mathbf{W}_b^T \mathbf{W}_d^T]$ such that the (exponential family) conditional distributions consistent with the joint model are

$$P(\mathbf{z}_n|\tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{z}_n; \hat{\boldsymbol{\mu}}, \mathbf{I}), \quad \hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \mathbf{W}^T\tilde{\mathbf{x}} \tag{17}$$

$$P(\mathbf{x}_b|\tilde{\mathbf{z}}) = \mathcal{B}(\mathbf{x}_b; \hat{\boldsymbol{\theta}}_b), \quad \hat{\boldsymbol{\theta}}_b = \boldsymbol{\theta}_b + \mathbf{W}_b\tilde{\mathbf{z}} \tag{18}$$

$$P(\mathbf{x}_d|\tilde{\mathbf{z}}) = \mathcal{D}(\mathbf{x}_d; \hat{\boldsymbol{\theta}}_d), \quad \hat{\boldsymbol{\theta}}_d = \boldsymbol{\theta}_d + \mathbf{W}_d\tilde{\mathbf{z}}, \tag{19}$$

where $\mathcal{N}()$, $\mathcal{B}()$ and $\mathcal{D}()$ represent Normal, Bernoulli and Discrete distributions respectively. The following equation can be used to represent the marginal distribution of $\mathbf{x}$,

$$P(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} - A(\boldsymbol{\theta}, \boldsymbol{\Lambda})\}, \tag{20}$$

where $\boldsymbol{\Lambda} = \frac{1}{2}\mathbf{W}\mathbf{W}^T$ and $\boldsymbol{\theta}$ combines $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_b$. The labels for this model are the discrete random variable (i.e. $\mathbf{y} = \mathbf{x}_d$) and the features are the binary variables.

In an exponential family model with exponential function $\mathbf{F}(\mathbf{x};\theta)$, it is easy to verify that the gradient of the log marginal likelihood $\mathcal{L}$ of the observed data $\mathbf{x}$, can be expressed

$$\frac{\partial \mathcal{L}(\theta;\mathbf{x})}{\partial \theta} = N\left[\left\langle \frac{\partial \mathbf{F}(\mathbf{x};\theta)}{\partial \theta} \right\rangle_{\tilde{P}(\mathbf{x})} - \left\langle \frac{\partial \mathbf{F}(\mathbf{x};\theta)}{\partial \theta} \right\rangle_{P(\mathbf{x};\theta)}\right], \tag{21}$$

where $\langle\cdot\rangle_{\tilde{P}(\mathbf{x})}$ denotes the expectation under the empirical distribution, $\langle\cdot\rangle_{P(\mathbf{x})}$ is an expectation under the models marginal distribution and N is the number of data elements. We can thus compute the gradient of the log-likelihood with respect to the weight matrix $\mathbf{W}$ using

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^T} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\mathbf{W}^T\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T - \frac{1}{N_s}\sum_{j=1}^{N_s}\mathbf{W}^T\tilde{\mathbf{x}}_{i,(j)}\tilde{\mathbf{x}}_{i,(j)}^T\right), \tag{22}$$

where N_d are the number of vectors of observed data, $\tilde{\mathbf{x}}_{i,(j)}$ are samples indexed by j and N_s are the number of samples used per data vector, computed using Gibbs sampling with conditionals (17), (18) and (19). In our experiments here we have found it possible to use either one or a small number of Markov Chain Monte Carlo (MCMC) (Andrieu et al., 2003) steps initialized from the data vector (the contrastive divergence approach (Hinton, 2002)). Standard MCMC approximations for expectations are also possible. We use straightforward gradient-based optimization for model parameters with a learning rate and a momentum term. Finally, for conditional likelihood and multi-conditional likelihood based learning, gradient values can be obtained from

$$\begin{aligned}\frac{\partial \mathcal{L}_{MC}}{\partial \theta} = N\Bigg[(\alpha+\beta)\Bigg[\left\langle \frac{\partial \mathbf{F}(\mathbf{x}_b,\mathbf{x}_d;\theta)}{\partial \theta}\right\rangle_{\tilde{P}(\mathbf{x}_b,\mathbf{x}_d)} \\ - \alpha\left\langle\left\langle \frac{\partial \mathbf{F}(\mathbf{x}_b,\mathbf{x}_d;\theta)}{\partial \theta}\right\rangle_{P(\mathbf{x}_d|\mathbf{x}_b;\theta)}\right\rangle_{\tilde{P}(\mathbf{x}_b)}\Bigg] \\ - \beta\left\langle\left\langle \frac{\partial \mathbf{F}(\mathbf{x}_b,\mathbf{x}_d;\theta)}{\partial \theta}\right\rangle_{P(\mathbf{x}_b|\mathbf{x}_d;\theta)}\right\rangle_{\tilde{P}(\mathbf{x}_d)}\Bigg]\Bigg]\end{aligned} \tag{23}$$

Relationships to Other Work

Theoretical and empirical results in Ng and Jordan (2002) have supported the notion that, while a discriminative model may have a lower asymptotic error (with more data), the error rate of classifications based on an analogous generative model can often approach an asymptotically higher error rate faster. Hybrids methods combining generative and discriminative methods are appealing in that they have the potential to draw upon the strengths of both approaches. For example, in Raina et al. (2003), a high dimensional subset of parameters are trained under a joint likelihood objective while another smaller subset of parameters are trained under a conditional likelihood objective. In contrast, in our approach all parameters are optimized under a number of conditional objectives.

In Corduneanu and Jaakkola (2003), a method characterized as information regularization is formulated for using information about the marginal density of unlabeled data to constrain an otherwise free conditional distribution. Their approach can be thought of as a method for penalizing decision boundaries that occur in areas of high marginal density. In terms of the regularization perspective, our multi-conditional approach uses additional or auxiliary conditional distributions derived from an underlying joint probability model as regularizers. Furthermore, our approach is defined within the context of an underlying joint model. It is our belief that these additional conditional distributions in our objective function can serve as a regularizer for the conditional distributions we primarily care about, the probability of labels. As such, we weight the conditional distributions differently in our objective.

With equal weighting of conditionals and an appropriate definition of subsets of variables, the method can be seen as a type of pseudo-likelihood (Besag, 1975). However, our goals are quite different, in that we are not trying to *approximate* a joint likelihood, but rather, we wish to explicitly optimize for the conditional distributions in our objective.

The mixtures of naïve MRFs we present resemble the multiple mixture components per class approach used in Nigam et al. (2000). The conditional distributions arising for our labels given our data are also related to mixtures of experts (Jordan & Jacobs, 1994), conditional mixture models (Jebara & Pentland, 1998), simple mixtures of maximum entropy models (Pavlov et al., 2002), and mixtures of conditional random fields (McCallum et al., 2005; Quattoni et al., 2004). The continuous latent variable model we present here is similar to the dual wing harmonium or two layer random field presented in Xing et al. (2005) for mining text and images. In that approach a lower dimensional representation of image and text data is obtained by optimizing the joint likelihood of a harmonium model.

Experimental Results

In this section, we present experimental results using multi-conditional objective functions in the context of the models described. First, we apply naïve Markov random fields to document classification and show that the multi-conditional training provides better regularization than the traditional Gaussian prior. Next, we demonstrate mixture forms of the model on both real and synthetic data, including an example of topic discovery. Finally, we show that in harmonium-structured models, the multi-conditional objective provides a quantitatively better latent space.

Naïve MRFs and MCL as Regularization

We use the objective function $\alpha\mathcal{L}_{y|x}(\boldsymbol{\theta})+\beta\mathcal{L}_{x|y}(\boldsymbol{\theta})$ in naïve MRFs and compare to the generative naïve Bayes model and the discriminative maximum entropy model for document classification. We present extensive experiments with common text data sets, which are briefly described below.

- 20 Newsgroups is a corpus of approximately 20,000 newsgroup messages. We use the entire corpus (abbreviated as *news*), as well as two subsets (*talk* and *comp*).
- The industry sector corpus is a collection of corporate webpages split into about 70 categories. We use the entire corpus (*sector*), as well as three subsets: *healthcare*, financial (*finan*), and *technology*.
- The movie review corpus (*movie*) is a collection of user movie reviews from the Internet Movie Database, compiled by Bo Pang at Cornell University. We used the polarity data set (v2.0), where the task is to classify the sentiment of each review as positive or negative.
- The *sraa* data set consists of 73,218 UseNet articles from four discussion groups: simulated auto racing, simulated aviation, real autos, and real aviation.
- The Web Knowledge Base (*webkb*) data set consists of webpages from four universities that are classified into faculty, student, course, and project (we discard the categories of staff, department, and other).

We determine α and β, the weights of each component of our objective function, and the Gaussian prior variance σ^2 using cross validation. Specifically, we use 10-fold cross-validation, with 5 folds used for choosing these parameters and 5 folds used for testing. The models tend to be quite sensitive to the values of α and β. Additionally, because there is no longer a guarantee of convexity, thoughtful initialization of parameters is sometimes required. In future work, we hope to more thoroughly understand and control for these engineering issues.

During preprocessing, we remove words that only occur once in the each corpus, as well as stopwords, HTML, and email message headers. We also test with small-vocabulary versions of each data set in which the vocabulary size is reduced to 2000 using information gain.

The results are presented in Table 1. The parenthesized values are the standard deviations of the test accuracy across the cross validation folds. On 15 of 20 data sets, we show improvements over both maximum entropy and naïve Bayes. Although the differences in accuracy are small in some cases, the overall trend across data sets illustrates the potential of MCL for regularization. In fact, the difference between the mean accuracy for maximum entropy and MCL is larger than the difference between the mean accuracies of naïve Bayes and maximum entropy. Across all data sets, the mean MCL accuracy is significantly greater than the mean accuracies of naive Bayes ($p = 0.001$) and maximum entropy ($p = 0.0002$) under a one-tailed paired *t*-test.

We also found that in 10 of 15 data sets on which we also calculated the area under the accuracy/coverage curve, MCL provided better confidence estimates.

Mixtures of Naïve MRFs

In order to demonstrate the ability of multi-conditional mixtures to successfully classify data that is not linearly separable, we perform the following synthetic data experiments. Four class labels are each associated with four 4-dimensional Gaussians, having means and variances uniformly sampled between 0-100. Positions of data points generated from the Gaussians are rounded to integer values. For some samples of the Gaussian means and variances—*e.g.* an XOR configuration—a significant portion of the data would be misclassified by the best linear separator. MCMs, however, can learn and combine multiple linear decision boundaries. A MCM with two hidden subclasses per class attains an accuracy of 75%, whereas naïve Bayes, maximum entropy, and non-mixture multi-conditional naïve MRFs have accuracies of 54%, 52%, and 56%, respectively. With explicitly-constructed XOR positioning, MCM attains 99%, while the others yield less than 50%.

Running these MCMs on the *talk* data set yields "topics" similar to latent Dirichlet allocation (LDA) (Blei et al., 2003), except that parameter estimation is driven to discover topics that not only re-generate the words, but also help predict the class label; (thus MCM can also be understood as a "semi-supervised" topic model). Furthermore, MCM topics

Data	Naive Bayes	MaxEnt	MCL
news	85.3 (0.61)	82.9 (0.82)	**85.9 (0.89)**
news (2000)	76.4 (0.88)	77.4 (0.81)	**77.7 (0.48)**
comp	**85.1 (1.78)**	83.7 (0.68)	83.4 (0.94)
comp (2000)	81.8 (1.36)	82.2 (0.75)	**84.0 (1.05)**
talk	**84.6 (1.02)**	82.3 (1.43)	83.7 (1.27)
talk (2000)	83.7 (2.17)	81.6 (2.27)	**84.3 (1.21)**
sector	75.6 (2.05)	**88.0** (1.13)	87.4 (0.84)
sector (2000)	73.9 (0.78)	82.0 (1.03)	**83.2 (1.56)**
tech	91.0 (1.33)	91.8 (2.24)	**93.1 (1.69)**
tech (2000)	92.9 (2.46)	91.4 (2.03)	**94.5 (1.81)**
finan	**92.3 (2.36)**	89.2 (1.52)	91.5 (2.57)
finan (2000)	87.3 (3.31)	89.6 (1.82)	**94.6 (1.79)**
health	93.5 (4.36)	94.0 (3.74)	**95.5 (4.00)**
health (2000)	95.0 (5.00)	91.0 (3.39)	**95.5 (4.30)**
movie	78.6 (1.20)	82.6 (2.96)	**82.7 (2.50)**
movie (2000)	90.9 (1.98)	88.8 (1.96)	**94.0 (1.05)**
sraa	95.9 (0.15)	96.1 (0.23)	**96.7 (0.09)**
sraa (2000)	93.7 (0.20)	94.7 (0.13)	**95.0 (0.21)**
webkb	87.9 (2.14)	**92.4 (0.84)**	**92.4 (1.04)**
webkb (2000)	84.7 (1.20)	92.4 (1.07)	**92.7 (1.40)**
mean	86.5 (6.73)	87.7 (5.39)	**89.4 (5.76)**

Table 1: Document classification accuracies for naive Bayes, maximum entropy, and MCL.

are defined not only by positive word associations, but also by prominent negative word associations. The words with most positive and negative $\boldsymbol{\theta}_{x,z}$ are shown in Table 2.

Lower-variance Conditional Mixture Estimation

Consider data generated from two classes, each with four sub-classes drawn from 2-D isotropic Gaussians (similar to the example in Jebara and Pentland (2000)). The data are illustrated by red ◦'s and blue ×'s in Figure 3. Using joint, conditional, and multi-conditional likelihood, we fit mixture models with two (diagonal covariance, *i.e.* naïve) subclasses using conditional expected gradient optimization (Salakhutdinov et al., 2003). The figure depicts the parameters of the best models found under our objectives using ellipses for constant probability under the model.

From this illustrative example, we see that the parameters estimated by joint likelihood would completely fail to classify ◦ versus × given location. In contrast, the conditional objective focuses completely on the decision boundary, however, in 30 random initializations, this produced parameters with very high variance, and little interpretability. Our multi-conditional objective, however, optimizes for both class label prediction and class-conditioned density, yielding good classification accuracy, and sensible, low-variance parameter estimates.

Multi-Conditional Harmoniums

We are interested in the quality of the latent representations obtained when optimizing multi-attribute harmonium structured models under standard (joint) maximum likelihood (ML), conditional likelihood (CL) and multi-conditional likelihood (MCL) objectives. We use a similar testing strategy to Welling et al. (2005) but focus on comparing the different latent spaces obtained with the various optimization objectives. As in Welling et al. (2005), we used the reduced 20 newsgroups data set prepared in MATLAB by Sam Roweis. In this data set, 16242 documents are represented by 100 word vocabulary binary occurrences and are labeled as one of four domains.

To evaluate the quality of our latent space, we retrieve documents that have the same domain label as a test document based on their cosine coefficient in the latent space when observing only binary occurrences. We randomly split data into a training set of 12,000 documents and a test set of 4242 documents. We use a joint model with a corresponding full rank multi-variate Bernoulli conditional for binary word occurrences and a discrete conditional for domains. Figure 4 shows the precision-recall results. ML-1 is our model with no domain label information. ML-2 is optimized with domain label information. CL is optimized to predict domains from words and MCL is optimized to predict both words from domains and domains from words. From Figure 4 we see that the latent space captured by the model is more relevant for domain classification when the model is optimized under the CL and MCL objectives. MCL more than doubles the precision and recall at reasonable values of the counterparts.

Topic 1 (gun control)		Topic 2 (Waco incident)	
guns	1.27	nra	1.63
texas	1.19	assault	1.52
gun	1.18	waco	1.21
enforcement	1.14	compound	1.19
...	...	...	
president	-0.83	employer	-0.90
peace	-0.85	cult	-0.94
years	-0.88	terrorists	-1.02
feds	-1.17	matthew	-1.15

Table 2: Two MCM-discovered "topics" associated with the `politics.guns` label in a run on *talk* data set. On the left, discussion about gun control in Texas. The negatively-weighted words are prominent in other classes, including `politics.misc`. On the right, discussion about the gun rights of David Koresh when the NRA stormed their compound in Waco, TX. Aspects of the Davidian cult, however, were discussed in `religion.misc`.

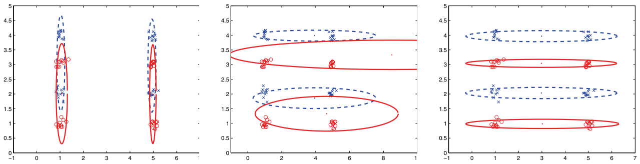

Figure 3: (Left) Joint likelihood optimization. (Middle) One of the many near optimal solutions found by conditional likelihood optimization. (Right) An optimal solution found by our multi-conditional objective.

Discussion and Conclusions

We have presented multi-conditional learning in the context of naïve MRFs, mixtures of naïve MRFs and harmonium-structured models. For Naive MRFs, we show that multi-conditional learning provides improved regu-

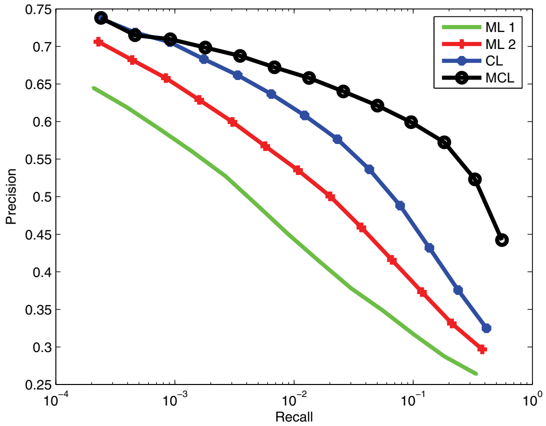

Figure 4: Precision-recall curves for the "20newsgroups" data using ML, CL and MCL with 20 latent variables. Random guessing is a horizontal line at .25.

larization, and flexible, robust mixtures. In the context of harmonium-structured models our experiments show that multi-conditional contrastive-divergence-based optimization procedures can lead to latent document spaces with superior quality.

Multi-conditional learning is well suited for multi-task and semi-supervised learning, since multiple prediction tasks are easily and naturally defined in the MCL framework. In recent work by Ando and Zhang (2005), semi-supervised and multi-task learning methods are combined. Their approach involves auxiliary prediction problems defined for unlabeled data such that model structures arising from these tasks are also useful for another classification problem of particular interest. Their approach involves finding the principal components of the parameters space for auxiliary tasks. One can similarly use the MCL approach to define auxiliary conditional distributions among features. In this way MCL is a natural framework for semi-supervised learning. We are presently exploring MCL in these multi-task and semi-supervised settings.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Central Intelligence Agency, the National Security Agency and the National Science Foundation under NSF grant #IIS-0326249, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010.

References

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*, 39–72.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, *24*, 179–195.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Corduneanu, A., & Jaakkola, T. (2003). On information regularization. *Proceedings of Uncertainty in Artificial Intelligence*.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*, 1771–1800.

Jebara, T., & Pentland, A. (1998). Maximum conditional likelihood via bound maximization and the CEM algorithm. *In Neural Information Processing Systems (NIPS), 11*.

Jebara, T., & Pentland, A. (2000). On reversing Jensen's inequality. *NIPS 13*.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.

Kschischang, F. R., Frey, B., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*, 498–519.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. ICML*, 282–289.

McCallum, A., Bellare, K., & Pereira, F. (2005). A conditional random field for discriminatively-trained finite-state string edit distance. *Conference on Uncertainty in AI (UAI)*.

Minka, T. (2005). Discriminative models, not discriminative training. *MSR-TR-2005-144*.

Ng, A. Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *NIPS 14*.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

Pavlov, D., Popescul, A., Pennock, D., & Ungar, L. (2002). Mixtures of conditional maximum entropy models. *NEC Research Institute Technical Report NECI*.

Quattoni, A., Collins, M., & Darrell, T. (2004). Conditional random fields for object recognition. *NIPS 17*, 1097–1104.

Raina, R., Shen, Y., Ng, A. Y., & McCallum, A. (2003). Classification with hybrid generative/conditional models. *NIPS*.

Salakhutdinov, R., Roweis, S., & Ghahramani, Z. (2003). Optimization with EM and expectation-conjugate-gradient. *Proc. ICML*.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In D. Rumehart and J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. volume 1: Foundations*, 194–281. MIT Press.

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar (Eds.), *Introduction to statistical relational learning*. MIT Press. To appear.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2005). Exponential family harmoniums with an application to information retrieval. *NIPS*, 1481–1488.

Xing, E., Yan, R., & Hauptmann, A. G. (2005). Mining associated text and images with dual-wing harmoniums. *Proc. Uncertainty in Artificial Intelligence*.