

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 latter 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 latter 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 interest 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 $i = 1, \dots, N$ 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 $j = 1, \dots, N_s$ pairs of disjoint subsets of observations $\{\tilde{x}_A\}_{ij}$ and $\{\tilde{x}_B\}_{ij}$, where our indices denote the i th 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

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

where we derive these marginal conditional likelihoods from a single underlying joint probability model with parameters θ . 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

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

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

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

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 $\mathcal{D} = \{(\tilde{y}_n, \tilde{x}_{j,n}); n = 1, \dots, N, j = 1 \dots M_n\}$ 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 θ . 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

$$P(x_1, \dots, x_{M_n}, y|\theta) = \frac{1}{Z} \phi(y|\theta_y) \prod_{j=1}^{M_n} \phi(x_j, y|\theta_{x,y}) \quad (4)$$

where

$$Z = \sum_y \sum_{x_1} \dots \sum_{x_{M_n}} \phi(y|\theta_y) \prod_{j=1}^{M_n} \phi(x_j, y|\theta_{x,y}). \quad (5)$$

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

$$P(\mathbf{y}, \{\mathbf{w}\}) = \frac{1}{Z} \exp\left(\mathbf{y}^T \theta_y + \mathbf{y}^T \theta_{x,y}^T \sum_{j=1}^{M_n} \mathbf{w}_j\right). \quad (6)$$

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 θ_y and $\theta_{x,y}$ into θ such that

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

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

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

where

$$Z(\mathbf{y}) = \sum_{\mathbf{w}_1} \dots \sum_{\mathbf{w}_{M_n}} \prod_{j=1}^{M_n} \exp(\mathbf{y}^T \theta_{x,y}^T \mathbf{w}_j) \exp(\mathbf{y}^T \theta_y). \quad (9)$$

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

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

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

