MMH: Maximum Margin Supervised Harmoniums

Ning Chen[†‡] NINGCHEN@CS.CMU.EDU
Jun Zhu[‡] JUNZHU@CS.CMU.EDU
[†]Dept. of Computer Science & Technology, TNList Lab, Tsinghua University, Beijing 100084 China
[‡]School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

Abstract

Exponential family Harmoniums (EFH) are undirected topic models that enjoy nice properties such as fast inference compared to directed topic models. Supervised EFHs can utilize documents' side information for discovering predictive latent topic representations. However, existing likelihood based estimation does not yield conclusive results. This paper presents a max-margin approach to learning supervised EFHs for joint latent topic discovery and classification. The learning problem is efficiently solved with coordinate descent. We demonstrate the advantages of the max-margin approach on video data classification and retrieval.

1. Introduction

Probabilistic topic models have shown great success in extracting latent semantic structures of large collections of documents. Although directed topic models (e.g., latent Dirichlet allocation (LDA) (Blei et al., 2003)) have gained much more attention than undirected topic models (e.g., Harmoniums (Welling et al., 2004)), the undirected topic models do enjoy some nice and rather orthogonal properties (e.g., fast inference due to the encoded conditional independence), which make them more preferable in some applications, such as information retrieval (Welling et al., 2004) and multi-modal video data analysis (Xing et al., 2005). Recent work (Salakhutdinov & Hinton, 2009) also demonstrates that an undirected topic model can yield better generalization ability than LDA.

Most existing undirected topic models are unsupervised and they are incapable in utilizing the widely available side information, such as categories associated with images and rating scores associated with movie reviews. Recent work (Blei & McAuliffe, 2007; Wang et al., 2009; Lacoste-Julien et al., 2008; Zhu et al., 2009) has shown that by utilizing such supervised side information a directed topic model can discover informative latent topic representations, which may be more relevant to our applications (e.g., prediction). For undirected topic models, however, very few work has been done to explore such useful supervised information. The only exception is the hierarchical harmonium model or tri-wing Harmonium (TWH) (Yang et al., 2007). However, the likelihood-based learning method as used in (Yang et al., 2007) does not yield conclusive results, e.g., TWH does not show improvements in classification or retrieval.

In this paper, we explore the arguably more discriminative max-margin principle to train undirected topic models (e.g., Harmoniums) when supervised information is available. Our results on video data analysis show that by doing max-margin training, supervised Harmoniums can achieve better performance in classification and retrieval tasks. This better performance is due to the discovered more discriminative latent topic representations. Our work is motivated by and can be viewed as an undirected counterpart of the recently proposed MedLDA model (Zhu et al., 2009). The resultant optimization problem is efficiently solved by iteratively learning a multi-class SVM.

2. Harmonium Models

We begin with a brief recap of existing harmonium models and setting up the ground for our approach.

The dual-wing harmonium (DWH) (Xing et al., 2005) is an extension of EFH (Welling et al., 2004) for inferring latent topics from heterogenous input data, such as word counts and color histogram for video analysis. The model of DWH is shown in Fig. 1, where $\mathbf{H} := \{H_k\}$ are hidden variables and the variables on each plane (e.g., $\mathbf{X} := \{X_n\}$) represent one type of input data. DWH defines a joint distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h})$, which is fully determined by the three between-layer conditionals $p(\mathbf{x}|\mathbf{h})$, $p(\mathbf{z}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{x}, \mathbf{z})$, according to

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

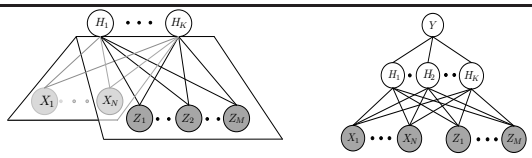

Figure 1. (L) dual-wing and (R) hierarchical Harmoniums.

the constructive definition (Xing et al., 2005). Moreover, due to the conditional independency encoded in DWH, each conditional has a factorized form (e.g., $p(\mathbf{h}|\mathbf{x},\mathbf{z}) = \prod_k p(h_k|\mathbf{x},\mathbf{z})$), which makes the inference easier, compared to directed topic models. Formally, the joint distribution of DWH has a log-linear form

$$p(\mathbf{x},\mathbf{z},\mathbf{h}) \propto \exp\Big\{\sum_i \theta_i^\top \phi(x_i) + \sum_j \eta_j^\top \psi(z_j) + \sum_k \lambda_k^\top \varphi(h_k) + \sum_{ik} \phi(x_i)^\top \mathbf{W}_i^k \varphi(h_k) + \sum_{ik} \psi(z_j)^\top \mathbf{U}_j^k \varphi(h_k)\Big\}, \quad (1)$$

where $\phi(x_i)$, $\psi(z_j)$ and $\varphi(h_k)$ are the features of x_i, z_j and h_k, respectively; and $\Theta := \{\theta_i, \eta_j, \lambda_k, \mathbf{W}_i^k, \mathbf{U}_j^k\}$ denotes the parameters. Then, the conditional distributions can be easily derived with shifted parameters

$$\begin{aligned} p(x_i|\mathbf{h}) &= \exp\{\hat{\theta}_i^\top \phi(x_i) - A_i(\hat{\theta}_i)\} \\ p(z_j|\mathbf{h}) &= \exp\{\hat{\eta}_j^\top \psi(z_j) - B_j(\hat{\eta}_j)\} \\ p(h_k|\mathbf{x},\mathbf{z}) &= \exp\{\hat{\lambda}_k^\top \varphi(h_k) - C_k(\hat{\lambda}_k)\}, \end{aligned} \quad (2)$$

where $\hat{\theta}_i = \theta_i + \sum_k \mathbf{W}_i^k \varphi(h_k)$, $\hat{\eta}_j = \eta_j + \sum_k \mathbf{U}_j^k \varphi(h_k)$ and $\hat{\lambda}_k = \lambda_k + \sum_i \mathbf{W}_i^k \phi(x_i) + \sum_j \mathbf{U}_j^k \psi(z_j)$; and A_i, B_j and C_k are log-partition functions.

As we have stated, the unsupervised EFH and DWH ignore the widely available supervised information. To incorporate such supervised information for discovering predictive latent topic representations, the tri-wing harmonium (TWH) or hierarchical harmonium (Yang et al., 2007) was proposed. The model of TWH is shown in Fig. 1 (R). To define a classification model, TWH uses the latent representation $\mathbf{H}$ as input data and defines the probability distribution

$$p(y|\mathbf{h}) = \frac{\exp\{\mathbf{V}^\top \mathbf{f}(\mathbf{h}, y)\}}{\sum_{y'} \exp\{\mathbf{V}^\top \mathbf{f}(\mathbf{h}, y')\}}, \quad (3)$$

where $\mathbf{f}(\mathbf{h}, y)$ is the feature vector whose elements from $(y-1)K+1$ to yK are those of $\mathbf{h}$ and all others are 0. Accordingly, $\mathbf{V}$ is a stacking parameter vector of T sub-vectors $\mathbf{V}_y$, of which each one corresponds to a class label y. Then, the joint distribution $p(\mathbf{x},\mathbf{z},\mathbf{h},y)$ has the same form as in Eq. (1), but with an additional term of $\mathbf{V}^\top \mathbf{f}(\mathbf{h}, y) = \mathbf{V}_y^\top \mathbf{h}$ in the exponential.

Note that for brevity, we have used one multi-valued discrete variable Y ($y \in \{1, \cdots, T\}$) with a softmax transformation to replace the original representation that uses T conditionally independent binary variables. The subtle difference is analogous to the difference between a multi-class SVM (Crammer & Singer, 2001) and the approach that builds multiple binary SVMs for multi-class classification.

In (Yang et al., 2007), parameter estimation is done by maximizing the joint data likelihood. However, as reported in (Yang et al., 2007), TWH does not yield improved performance as compared to the naive method that combines an unsupervised DWH for discovering latent representations and an SVM for classification. One of our motivations to develop an integrated large-margin approach is to investigate the reason why the supervised TWH does not yield improvements compared to the unsupervised DWH with a downstream combination of large-margin classifiers. As we shall see, integrating the large-margin principle into an objective function for joint latent representation discovery and classification can yield much better results, which demonstrate the usefulness of supervision.

3. MMH: Max-margin Harmoniums

Now, we present the max-margin approach for learning supervised undirected topic models for classification. For brevity, we consider the general multi-class classification, where $y \in \{1, \cdots, T\}$, as defined above.

3.1. Problem Definition

Likelihood-based approaches pay additional efforts in defining a normalized probabilistic model as in Eq. (3). An arguably more discriminative way to define a classification model is to directly estimate the decision boundary, which is the essential idea underlying the very successful large-margin classifiers (e.g., SVMs). Here, we integrate the large-margin idea into the learning of supervised harmonium models, analogous to the recent development of max-margin supervised LDA (MedLDA) (Zhu et al., 2009).

More specifically, as in the log-linear model in Eq. (3), we assume that the discriminant function $F(y, \mathbf{h}; \mathbf{V})$ is linear, that is, $F(y, \mathbf{h}; \mathbf{V}) = \mathbf{V}^\top \mathbf{f}(\mathbf{h}, y)$, where $\mathbf{f}$ and $\mathbf{V}$ are defined the same as above. For prediction, we take the expectation over the latent variable $\mathbf{H}$ and define the prediction rule as

$$y^* := \arg\max_y \mathbb{E}_{\mathbf{H}}[F(\mathbf{H}, y; \mathbf{V})], \quad (4)$$

where the expectation is taken over the posterior distribution $p(\mathbf{H}|\mathbf{x},\mathbf{z})$ or its variational approximation.

Now, learning is to find an optimal $\mathbf{V}^*$ that minimizes a loss function. Here, we minimize the hinge loss, as used in SVMs. Given training data $\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^D$, the hinge loss of the predictive rule (4) is

$$\mathcal{R}_{hinge}(\mathbf{V}) := \frac{1}{D} \sum_d \max_y [\Delta\ell_d(y) - \mathbf{V}^\top \mathbb{E}_{\mathbf{H}}[\Delta\mathbf{f}_d(y)]],$$

where $\Delta\ell_d(y)$ is a loss function that measures how different the prediction y is compared to the true label y_d, and $\mathbb{E}_{\mathbf{H}}[\Delta\mathbf{f}_d(y)] = \mathbb{E}_{\mathbf{H}}[\mathbf{f}(\mathbf{H}_d, y_d)] - \mathbb{E}_{\mathbf{H}}[\mathbf{f}(\mathbf{H}_d, y)]$. It can be proved that the hinge loss is an upper bound of the empirical loss $\mathcal{R}_{emp} := \frac{1}{D}\sum_d \Delta\ell(y_d^*)$.

Applying the principle of *regularized risk minimization*, we define the learning problem of MMH as

$$\text{MMH}: \min_{\Theta,\mathbf{V}} \; L(\Theta) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\mathcal{R}_{hinge}(\mathbf{V}), \quad (5)$$

where $L(\Theta) := -\sum_d \log p(\mathbf{x}_d, \mathbf{z}_d)$ is the negative data likelihood of DWH and C_1 and C_2 are non-negative constants, which can be selected via cross-validation.

The rationale underlying MMH is that: we want to find a latent topic representation $p(\mathbf{h}|\mathbf{x}, \mathbf{z})$ and a prediction model $\mathbf{V}$ which on one hand tend to predict as accurate as possible on training data, while on the other hand tend to explain the data well. The regularizor can prevent over-fitting.

3.2. Optimization

To solve the problem (5), we first make concrete instantiations of the model distributions. However, our procedure is generic for any instantiations.

Specifically, we consider the video analysis problem (Xing et al., 2005), where each shot is represented as a vector of word features $\mathbf{x}$ and color features $\mathbf{z}$. Each dimension x_i is a Bernoulli variable that denotes whether the ith term of a dictionary appears or not in the shot, and each dimension z_j is a real number that denotes the normalized color histogram of the keyframe in the shot. We assume each real-valued H_k follows a univariate Gaussian distribution. Therefore, we define the conditional distributions as

$$p(x_i = 1|\mathbf{h}) = 1/(1 + \exp(-\alpha_i - \mathbf{W}_{i\cdot}\mathbf{h}))$$
$$p(z_j|\mathbf{h}) = \mathcal{N}(z_j|\sigma_j^2(\beta_j + \mathbf{U}_{j\cdot}\mathbf{h}), \sigma_j^2)$$
$$p(h_k|\mathbf{x}, \mathbf{z}) = \mathcal{N}(h_k|\mathbf{x}^\top\mathbf{W}_{\cdot k} + \mathbf{z}^\top\mathbf{U}_{\cdot k}, 1),$$

where $\mathbf{W}_{i\cdot}$ and $\mathbf{W}_{\cdot k}$ denote the ith row and kth column of $\mathbf{W}$, respectively. Alike for other notations.

With the above definitions, we have $\mathbb{E}_{\mathbf{H}}[\mathbf{f}(\mathbf{H}, y)] = \mathbf{f}(\mathbf{v}, y)$, where $\mathbf{v}$ is a K-dimensional vector and $\mathbf{v}_k = \mathbf{x}^\top\mathbf{W}_{\cdot k} + \mathbf{z}^\top\mathbf{U}_{\cdot k}$. The data likelihood is

$$p(\mathbf{x}, \mathbf{z}) \propto \exp\left\{\alpha^\top\mathbf{x} + \beta^\top\mathbf{z} - \frac{1}{2}\sum_j \frac{z_j^2}{\sigma_j^2} + \frac{1}{2}\sum_k (\mathbf{x}^\top\mathbf{W}_{\cdot k} + \mathbf{z}^\top\mathbf{U}_{\cdot k})^2\right\}.$$

Variational approximation: Since the normalization factor of $p(\mathbf{x}, \mathbf{z})$ is generally intractable to compute, we apply the contrastive divergence method (Xing et al., 2005) to derive a variational approximation of the negative log-likelihood $-\log p(\mathbf{x}, \mathbf{z})$

$$\mathcal{L}^v(q_0, q_1) := R(q_0(\mathbf{x}, \mathbf{z}, \mathbf{h}), p(\mathbf{x}, \mathbf{z}, \mathbf{h})) - R(q_1(\mathbf{x}, \mathbf{z}, \mathbf{h}), p(\mathbf{x}, \mathbf{z}, \mathbf{h})),$$

where $R(q, p)$ is the relative entropy, and q_0 is a variational distribution with $\mathbf{x}$ and $\mathbf{z}$ clamped to their observed values while q_1 is a distribution with all variables free. For q, we make the naive mean field assumption that q is a product of singleton marginal over the variables[1] $q(\mathbf{x}, \mathbf{z}, \mathbf{h}) = \prod_i q(x_i)\prod_j q(z_j)\prod_k q(h_k)$.

[1]Unlike previous work (Xing et al., 2005; Yang et al., 2007), we don't need to assume the parametric forms of q.

Solving the approximate problem: Substituting the variational approximation $\mathcal{L}^v$ into problem (5), we get an approximate objective function $\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$. Then, we can apply co-ordinate descent to minimize $\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$. Specifically, for q_0 and q_1, we keep $(\Theta, \mathbf{V})$ fixed and update each marginal as

$$\begin{array}{l} q(x_i) = p(x_i|\mathbb{E}_{q(\mathbf{H})}[\mathbf{H}]), \quad q(z_j) = p(z_j|\mathbb{E}_{q(\mathbf{H})}[\mathbf{H}]) \\ q(h_k) = p(h_k|\mathbb{E}_{q(\mathbf{X})}[\mathbf{X}], \mathbb{E}_{q(\mathbf{Z})}[\mathbf{Z}]) \end{array} \quad (6)$$

For q_0, (x_i, z_j) are clamped at their observed values, and only $q(h_k)$ is updated. The distribution q_1 is achieved by performing the above updates starting from q_0. Several iterations can yield a good q_1.

After we have inferred q_0 and q_1, parameter estimation can be done by alternating between (1) keep Θ fixed and estimate $\mathbf{V}$: this problem is learning a multi-class SVM, which can be efficiently done with existing solvers; and (2) keep $\mathbf{V}$ fixed and estimate Θ: this can be solved with sub-gradient descent, where the sub-gradient is computed as

$$\delta\alpha_i = \mathbb{E}_{q_0}[x_i] - \mathbb{E}_{q_1}[x_i], \quad \delta\beta_j = \mathbb{E}_{q_0}[z_j] - \mathbb{E}_{q_1}[z_j]$$
$$\delta\mathbf{W}_{ik} = \mathbb{E}_{q_0}[x_i h'_k] - \mathbb{E}_{q_1}[x_i h'_k] - C_2\sum_d \mathbb{I}(\bar{y}_d \neq y_d)\sum_i \mathbb{E}_{q_1}[x_i]$$
$$\delta\mathbf{U}_{jk} = \mathbb{E}_{q_0}[z_j h'_k] - \mathbb{E}_{q_1}[z_j h'_k] - C_2\sum_d \mathbb{I}(\bar{y}_d \neq y_d)\sum_j \mathbb{E}_{q_1}[z_j]$$

where $h'_k = \mathbf{x}^\top\mathbf{W}_{\cdot k} + \mathbf{z}^\top\mathbf{U}_{\cdot k}$; $\mathbb{I}(\cdot)$ is an indicator function; and $\bar{y}_d = \arg\max_y[\Delta\ell_d(y) + \mathbf{V}^\top\mathbb{E}_{q_1}[\mathbf{f}(\mathbf{H}_d, y)]$ is the *loss-augmented prediction*. Based on the definition of q_0, the expectations $\mathbb{E}_{q_0}[x_i]$ and $\mathbb{E}_{q_0}[z_j]$ are actually the count frequency of x_i and z_j, respectively.

Note that in our integrated max-margin formulation, the sub-gradients of $\mathbf{W}$ and $\mathbf{U}$ contain an additional term (i.e., the third term) compared to the standard DWH with contrastive divergence approximation. This additional term introduces a regularization effect to the latent topic model. If the prediction $\bar{y}$ differs from the true label, this term will be non-zero and it biases MMH towards discovering a better representation for prediction.

4. Experiments

We report some empirical results of MMH on video data analysis. Our goals are to illustrate the benefits of max-margin learning in MMH for both latent topic discovery and final classification and retrieval tasks. Due to space limitation, we report a part of the experiments. More results are deferred to a full extension.

We use the TRECVID 2003 video data set (Xing et al., 2005), which contains 1078 manually annotated video shots that belong to 5 categories. Each shot is represented as a 1894-dim vector of text features and a 165-dim vector of HSV color histogram, which is extracted from the associated keyframe. We evenly split this data set into training and testing sets.

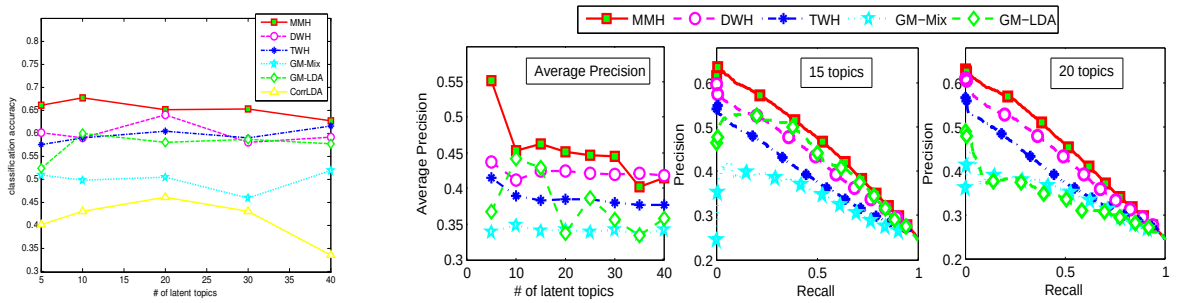

Figure 2. (Left) classification accuracy; and (Right) the average precision curve and the two precision-recall curves.

4.1. Classification

We compare the classification accuracy of MMH with DWH, TWH, Gaussian Mixture model (GM-Mix), Gaussian Mixture LDA (GM-LDA), and Correspondence LDA (CorrLDA). See (Blei & Jordan, 2003) for the details of the last three models. We use the SVM^{struct} [2] to solve the sub-step of learning the parameter $\mathbf{V}$ in MMH and build the SVM classifier for unsupervised models (i.e., DWH, GM-Mix, GM-LDA, and CorrLDA). Fig. 2 (left) shows the prediction accuracy of different models on testing data when topic numbers are changed from 5 to 40. We can see that the max-margin based supervised Harmonium (i.e., MMH) performs consistently better than any other methods. In contrast, the likelihood-based supervised Harmonium (i.e., TWH) does not show any improvements compared to the two-stage approach of combining the unsupervised DWH for topic discovery and an SVM for classification. These results demonstrate that supervised information can help in discovering informative latent topic representations that are more suitable for prediction if the model is appropriately learned, e.g., by using the discriminative max-margin method. The reasons for the inferior performance of the other models (e.g., CorrLDA and GM-Mix) are provided in (Xing et al., 2005; Yang et al., 2007).

4.2. Retrieval

In this task, each test sample is treated as a query and training samples are ranked based on the cosine similarity between a training sample and the given query. The similarity is computed based on the discovered latent topic representations. A sample is considered relevant to the query if it belongs to the same category as the query. We evaluate the retrieval results by computing the average precision (AP) and precision-recall curve. Fig. 2 (right) compares MMH with four other models when the topic number changes. Here, we show the precision-recall curves when the topic number is set at 15 and 20. We can see that for the AP measure, MMH outperforms all other methods in most cases, and MMH consistently outperforms all the other methods in the measure of precision-recall curve.

[2] http://svmlight.joachims.org/svm_multiclass.html

5. Conclusions and Discussions

We have presented the max-margin supervised Harmoniums (MMH), with a co-ordinate descent method provided for solving the learning problem. By optimizing one single objective function, MMH integrates the max-margin principle into the latent topic discovery process. Empirical results on video data demonstrate the promise of MMH on classification and retrieval.

Currently, we only present the results on classification and retrieval tasks, and MMH only uses the unsupervised DWH as the underlying topic model. In the full extension, more experimental results on latent topic discovery and the extension to using supervised TWH as the underlying topic model will be provided. It turns out that the latter extension is much easier than that in directed topic models (e.g., MedLDA).

References

Blei, David and McAuliffe, Jon D. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Blei, David M. and Jordan, Michael I. Modeling annotated data. In *ACM SIGIR*, pp. 127–134, 2003.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Crammer, Koby and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, (2):265–292, 2001.

Lacoste-Julien, Simon, Sha, Fei, and Jordan, Michael I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Salakhutdinov, Ruslan and Hinton, Geoffrey. Replicated softmax: an undirected topic model. In *NIPS*, 2009.

Wang, Chong, Blei, David, and Li, Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

Welling, Max, Rosen-Zvi, Michal, and Hinton, Geoffrey. Exponential family harmoniums with an application to information retrieval. In *NIPS*, pp. 1481–1488, 2004.

Xing, Eric P., Yan, Rong, and Hauptmann, Alexander G. Mining associated text and images with dual-wing harmoniums. In *UAI*, 2005.

Yang, Jun, Liu, Yan, Xing, Eric P., and Hauptmann, Alexander G. Harmonium models for semantic video representation and classification. In *SDM*, 2007.

Zhu, Jun, Ahmed, Amr, and Xing, Eric P. Medlda: Maximum margin supervised topic models for regression and classification. In *ICML*, 2009.