Max-Margin Latent Dirichlet Allocation for Image Classification and Annotation

Yang Wang
yangwang@uiuc.edu
Dept of Computer Science, University of Illinois at Urbana Champaign
Greg Mori
mori@cs.sfu.ca
School of Computing Science, Simon Fraser University

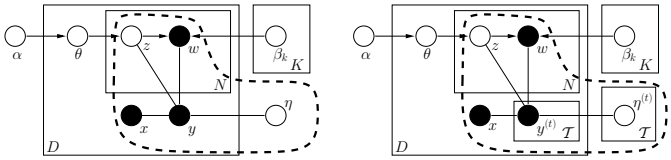

Figure 1: Graphical illustrations of (left) MMLDA^c for image classification; (right) MMLDA^a for image annotation. The variables enclosed by the dashed line are involved in the max-margin component of the model.

Much work in image classification and labeling uses topic models (e.g. LDA [1]), which are a class of powerful tools originally proposed in text modeling and have gained much popularity in computer vision recently. Despite the success of topic models in visual recognition, we believe there are some limitations of the way that topic models are used in computer vision. First of all, most topic models unsupervised. This means the topics discovered by topic models are not necessarily the ones used for discriminative tasks, such as image classification. To address this issue, several supervised variants of topic models have been developed. But the limitation of those models is that most of them assume the "bag-of-words" image representation, i.e. an image is represented by a collection of unordered feature descriptors computed from small local patches. Although the "bag-of-words" representation has been proven successful, other more holistic image representations (e.g. GIST [2]) have been shown to be powerful in many applications too. It is desirable to have the best of both worlds and design a model that can exploit both types of feature representation.

In this paper, we propose the *max-margin latent Dirichlet allocation* (MMLDA), a variant of MedLDA [3]. We introduce two different versions of MMLDA, called MMLDA^c for image classification, and MMLDA^a for image annotation. MMLDA^c is based on MedLDA. The main difference is that MedLDA only uses the latent topics as the feature vector for classification, while MMLDA^c uses latent topics together with any other image features. This extension allows MMLDA^c to make use of image features (e.g. GIST) that cannot be easily represented as bag-of-words. MMLDA^a is an extension of MMLDA^c for image annotation. In image annotation, the goal is to choose a set of annotation terms (also called *tags*) to describe an image. Since an image can be associated with more than one tag, image classification is a multi-label classification. In MMLDA^a, various tags are implicitly coupled by the latent topics defined in the model. Training MMLDA^a results in topic representations that are suitable for predicting those tags.

MMLDA^c: We use $\mathbf{x}$ to denote an image. We use $\mathbf{w}$ to denote the bag-of-words representation of $\mathbf{x}$, e.g. $\mathbf{w}$ can be obtained by vector-quantization of SIFT descriptors. The topic assignment of the words in the document is denoted by $\mathbf{z}$. We assume a linear discriminative function of the form $F(y,\mathbf{z},\mathbf{w},\mathbf{x},\eta) = \eta_y^\top f(\mathbf{z},\mathbf{w},\mathbf{x})$. Note the definition of $F(\cdot)$ is similar to that in MedLDA. In fact, if we assume $f(\mathbf{z},\mathbf{w},\mathbf{x}) = \bar{z} = \frac{1}{N}\sum_{n=1}^N z_n$, we can recover $F(\cdot)$ in MedLDA. So the definition of $F(\cdot)$ in MMLDA is a strict generalization of that in MedLDA. One important thing to remember is that since $\mathbf{z}$ is not observed, $f(\mathbf{z},\mathbf{w},\mathbf{x})$ is actually a random vector implicitly defined by the distribution on Z.

We assume $f(\mathbf{z},\mathbf{w},\mathbf{x})$ is a concatenation of two sub-vectors $f(\mathbf{z},\mathbf{w},\mathbf{x}) = \text{cat}(\bar{z}; g(\mathbf{w},\mathbf{x}))$, where $g(\mathbf{w},\mathbf{x})$ is a vector defined on $\mathbf{w}$ and $\mathbf{x}$, $\bar{z}$ is defined as $\bar{z} = \frac{1}{N}\sum_{n=1}^N z_n$ similar to sLDA and MedLDA, $\text{cat}(\mathbf{a};\mathbf{b})$ denotes the concatenation of two vectors $\mathbf{a}$ and $\mathbf{b}$. Notice that we do not have any assumption on the form of $g(\mathbf{w},\mathbf{x})$, it can be any feature vector extracted from the image, e.g. histogram of words, GIST descriptors, or both. Similarly, we assume η_y is also a concatenation of two sub-vectors $\eta_y = \text{cat}(\zeta_y; \nu_y)$, so that $\eta_y^\top f(\mathbf{z},\mathbf{w},\mathbf{x}) = \zeta_y{}^\top \bar{z} + \nu_y{}^\top g(\mathbf{w},\mathbf{x})$. Fig. 1 (a) shows a graphical illustration of MMLDA^c.

Similar to MedLDA, we learn the model parameter by solving an optimization problem as follows:

$$\min_{q,\alpha,\beta,\eta,\xi} -\mathscr{L}(q) + \frac{1}{2}\lambda||\eta||^2 + \tau\sum_{d=1}^D \xi_d, \ \text{s.t.}\forall d, y: \quad \xi_d \geq 0 \tag{1a}$$

$$(\eta_{y_d}^\top - \eta_y^\top)\mathbb{E}[f(\bar{Z}_d,\mathbf{w}_d,\mathbf{x}_d)] \geq \Delta(y,y_d) - \xi_d \tag{1b}$$

where $\mathscr{L}(q)$ is a lower bound on the log-likelihood of the data (similar to LDA). Without $\mathscr{L}(q)$, the optimization in Eq.(1) simply defines a multi-class SVM. The intuition behind Eq.(1) is that it will set the model parameters to explain the data well, and simultaneously discover topics suitable for image classification.

MMLDA^a: Let us assume an image annotation task with $\mathscr{T}$ possible tags. For a given image $\mathbf{x}$, our goal is to predict a $\mathscr{T}$ dimensional binary vector $\mathbf{y} = (y^{(1)}, y^{(2)}, ..., y^{(\mathscr{T})})$, where $y^{(t)}$ is 1 or 0 indicating the presence/absence of the t-th tag for this image. One simple solution is to formulate image annotation as $\mathscr{T}$ binary classification problems, where a binary classifier is learned separately to predict the presence/absence of each tag based on the image features. During testing, we run $\mathscr{T}$ binary classifiers to independently predict whether each tag should be chosen to describe an unseen image. The disadvantage of this approach is that those $\mathscr{T}$ are learned independently of each other, but the annotations of an image are usually correlated, e.g. annotation terms such as "car", "road", "sky" tend to appear together. In our work, we directly use the latent topics as the low-dimensional space shared by tag classifiers.

We assume the classifier for the t-th tag is a binary linear SVM. We use $\eta^{(t)}$ to denote the parameter of this SVM classifier. Accordingly, $\eta^{(t)}$ has two sub-parts corresponding to vector obtained via latent topics $\mathbb{E}(\bar{Z}_d)$, and the vector obtained via $g(\mathbf{w},\mathbf{x})$. The training data are in the form of $\{\mathbf{x}_d, \mathbf{y}_d\}_{d=1}^D$, where the label $\mathbf{y}_d$ is a $\mathscr{T}$ dimensional binary vector $\mathbf{y}_d = (y_d^{(1)}, y_d^{(2)}, ..., y_d^{(\mathscr{T})})$, and $y_d^{(t)} = 0$ or 1. We use $\Delta^a(\mathbf{y},\mathbf{y}_d)$ to denote the loss incurred by predicting $\mathbf{y}$ while the ground-truth annotation is $\mathbf{y}_d$. We assume $\Delta^a(\mathbf{y},\mathbf{y}_d)$ decomposes into the summation of $\mathscr{T}$ per-tag losses as $\Delta^a(\mathbf{y},\mathbf{y}_d) = \sum_{t=1}^{\mathscr{T}} \Delta(y^{(t)}, y_d^{(t)})$, where $y_d^{(t)}$ denotes the ground-truth label for the t-th tag for document d, $y^{(t)}$ denotes an arbitrary label for the t-th tag. The loss $\Delta(y^{(t)}, y_d^{(t)})$ is the 0-1 loss.

Then we can formulate the multi-label image annotation using the following optimization problem:

$$\min_{q,\alpha,\beta,\eta,\xi} -\mathscr{L}(q) + \frac{1}{2}\lambda\sum_{t=1}^{\mathscr{T}}||\eta^{(t)}||^2 + \tau\sum_{d=1}^D\sum_{t=1}^{\mathscr{T}}\xi_d^{(t)}$$

$$\text{s.t.} \quad \forall d,t,y^{(t)}: \quad \xi_d^{(t)} \geq 0$$

$$\eta_{y_d^{(t)}}^{(t)\top}\mathbb{E}[f(\bar{Z}_d,\mathbf{w}_d,\mathbf{x}_d)] - \eta_{y^{(t)}}^{(t)\top}\mathbb{E}[f(\bar{Z}_d,\mathbf{w}_d,\mathbf{x}_d)] \geq \Delta(y^{(t)},y_d^{(t)}) - \xi_d^{(t)}$$

The intuition of the optimization problem is that we want to find model parameters that explain the data well, and at the same time discover topics suitable for predicting all the tags.

We have develop efficient learning and inference algorithm for the models. We demonstrate the effectiveness of the models on two datasets.

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

[2] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[3] Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning*, 2009.