CORRELATED TOPIC MODEL DETAILS

JONATHAN HUANG AND TOMASZ MALISIEWICZ

1. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) model by Blei et al [BNJ02] is a generative model for a collection of exchangeable discrete data. LDA has mainly been used to model text corpora, where the notion of exchangeability corresponds to the "bag-of-words" assumption that is commonly employed in such models. More recently, Sivic et al [SRE$^+$05] and Li and Perona [FFP05] have applied the model to collections of images, where the "visual words" correspond to quantized local feature vectors.

The model is simplest to describe for text corpora. In a nutshell, LDA models each document as a mixture over topics, where each vector of mixture proportions is assumed to have been drawn from a Dirichlet distribution. A topic in this model is defined to be a discrete distribution over words from some finite lexicon. For example, if a topic is "astrophysics", then the word "quasar" would presumably be assigned a higher probability than the word "burrito".

More precisely, let D be a collection of M documents and $\mathbf{w}$ be a document with words $w_1, \ldots, w_n$. We assume that each word corresponds to one of K possible topics, and that for each word, there is a latent topic assignment z_i which takes on values $1, \ldots, K$, indexing into the set of topics. By convention, we will represent z_i by a K dimensional vector with one component set to one indicating its value in $1, \ldots, K$ and the rest set to zero. The generative process is as follows:

(1) Draw $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$
(2) For each word $w_n \in \mathbf{w}$,
    (a) Draw a topic $z_i \sim \text{Multinomial}(\theta)$
    (b) Draw a word $w_i \sim \text{Multinomial}(\beta(z_i))$ where $\beta(k)$ is a probability distribution over words corresponding to topic k

The inference task in LDA is to solve for the vector of topic mixture proportions, θ, and the topic assignment z_i for each word, given the words w_i and the model parameters α, β_k. Exact inference in this model involves an intractable integral, but approximations via Markov Chain Monte Carlo sampling [GS04] or mean-field methods have been shown to work well.

2. The Correlated Topic Model

A problem with drawing the topic mixture proportions (θs) from a Dirichlet distribution is that Dirichlet distributions are too simple, and exhibit a near-independence structure.[1]

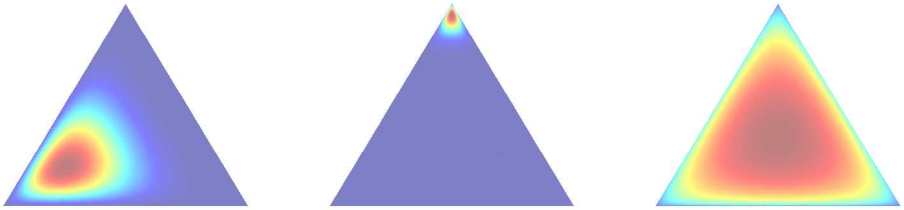

Figure 1. Dirichlet Distributions for various parameter settings on a 2-simplex. Red corresponds to high probability density and blue corresponds to low probability density.

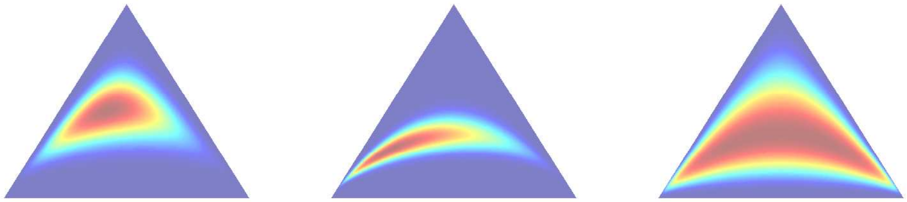

Figure 2. Logistic Normal Distributions for various parameter settings on a 2-simplex. Note that unlike the Dirichlet, its level sets can bound nonconvex regions.

The *Logistic-Normal* distribution [AS80] is an alternative distribution over a simplex which forms a richer class of distributions and better captures inter-component correlations. The process of drawing a k-dimensional Logistic-Normal random variable θ is as follows:

(1) Draw $v \sim N(\mu, \Sigma)$ where $N(\mu, \Sigma)$ is a $k-1$ dimensional Normal distribution.
(2) Define $v_k = 0$.
(3) Let

$$\theta = \frac{\exp v}{\sum_{j=1}^{k} \exp v_j}$$

(This is the projection of $\exp(v)$ to the simplex)

The probability density for θ can be explicitly written as

$$p(\theta; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|} \left(\prod_{j=1}^{k} \theta_j \right)^{-1} \exp \left[-\frac{1}{2} \{\log(\theta/\theta_k) - \mu\} \Sigma^{-1} \{\log(\theta/\theta_k) - \mu\} \right]$$

[1] A way to draw from a Dirichlet distribution is to sample k independent Gamma distributions, concatenate the samples into a vector and divide by their sum. This process shows that the correlations between the components of a Dirichlet random variable arise solely from the fact that they must all sum to one.

[2]

The Correlated Topic Model [BL05] models the same type of data as LDA and only differs in the first step of the generative process. Instead of drawing θ from a Dirichlet distribution it assumes that θ is drawn from a Logistic-Normal Distribution.

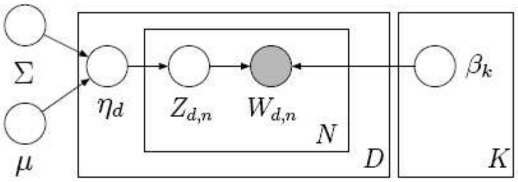

Figure 3. A graphical model representation of the Correlated Topic Model

3. Supervised Training

Expectation-Maximization provides a method for unsupervised training for LDA and CTM. However, in certain cases, if one can observe the latent variables (i.e. each word is labeled by a topic assignment) during the training process, then it makes more sense to take advantage of them. [3]

For LDA, if the topic assignment z_i is observed for each word during training, then we can estimate the α and β independently of each other. Since $\beta_k = P(w|z = k)$, β can be estimated simply by building histograms of words for each topic.

The distribution of z_i given the Dirichlet parameter α is known as a *Polya distribution* and details and code for fitting such a distribution are given by Minka [Min00].

For CTM, the way to estimate β in the supervised case is identical to LDA, but estimating μ and Σ is more involved since the Logistic-Normal distribution is not a member of the exponential family. The distribution of z_i given Logistic Normal parameters is called a Hierarchical Logistic Normal (HLN) distribution and we present a method for fitting one using EM in the appendix.

4. Inference

The details for variational inference are given in Appendix B.

Appendix

Appendix A. Fitting a Hierarchical Logistic-Normal Distribution

We now present the method for fitting the Hierarchical Logistic-Normal (HLN) distribution given by Hoff [Hof03]. The HLN distribution is the same as a CTM distribution except that no words are drawn. Recall the generative process for a given document $\mathbf{w_j}$:

[2] Since θ is actually a k-dimensional vector, we concatenate a zero to the end of μ and pad Σ and Σ^{-1} on the right and bottom by a column and row of zeros respectively.

[3] Note that this only changes the training stage - but the inference step on novel images remains difficult.

(1) Draw $v_j \sim N(\mu, \Sigma)$ where $N(\mu, \Sigma)$ is a $k-1$ dimensional Normal distribution.
(2) Define $v_{jk} = 0$.
(3) Let
$$\theta_j = \frac{\exp v}{\sum_{j=1}^{k} \exp v_j}$$
(4) For each word $w_{ji} \in \mathbf{w_j}$, draw a topic $z_{ji} \sim \text{Multinomial}(\theta)$

Notice that if the v_j are known, then finding the maximum likelihood estimates of μ and Σ is easy. Since they are unknown, the strategy will be instead to alternate between estimating $v_1, \ldots v_m$ for each document, and estimating μ and Σ using EM. Let $\hat{\mathbf{p}}(z)$ be the empirical distribution function (normalized histogram) of the topic assignments in a document. The conditional likelihood of $\mathbf{v}$ given the topic assignments $\mathbf{z} = (z_1, \ldots, z_n)$ for a given document can be written down using Bayes rule:

$$\begin{aligned} P(\mathbf{v}|\mathbf{z}, \mu, \Sigma) &\propto P(\mathbf{z}|\mathbf{v})P(\mathbf{v}|\mu, \Sigma) \\ &= \frac{\exp\left(\sum_{i=1}^{k-1} v_i n \hat{\mathbf{p}}_i\right)}{\left(1 + \sum_{j=1}^{k-1} \exp v_j\right)^n} \exp\left(-\frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)\right) \end{aligned}$$

The conditional log-likelihood and its derivatives are straight-forward (but not fun) to derive:

$$\log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma) = \sum_{i=1}^{k-1} v_i n \hat{\mathbf{p}}_i - n \log\left(1 + \sum_{j=1}^{k-1} \exp v_j\right) - \frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu) + C$$

$$\frac{\partial \log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma)}{\partial \mathbf{v}} = n\left(\hat{\mathbf{p}} - \frac{\exp \mathbf{v}}{1 + \sum_{j=1}^{k-1} \exp v_j}\right) - \Sigma^{-1}(v-\mu)$$

$$\begin{aligned} \frac{\partial^2 \log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma)}{\partial v_i \partial v_j} &= -\Sigma_{ij}^{-1} - n\left[\delta\{i=j\} \frac{\exp v_j}{1 + \sum_{l=1}^{k-1} \exp v_l}\right. \\ &\quad \left. - \left(\frac{\exp v_i}{1 + \sum_{l=1}^{k-1} \exp v_l}\right)\left(\frac{\exp v_j}{1 + \sum_{l=1}^{k-1} \exp x_l}\right)\right] \end{aligned}$$

By maximizing the conditional log-likelihood, the conditional mode of $\mathbf{v}$ can be found. [4]

Let $\hat{\mu}$ be the conditional mode of $\mathbf{v}$ and $\hat{I}$ be the Fisher Information matrix (negative Hessian) evaluated at $\hat{\mu}$. Then asymptotically,

$$f(\mathbf{v}|\mathbf{z}, \mu, \Sigma) \approx \mathcal{N}(\mathbf{v}|\hat{\mu}, \hat{I}^{-1})$$

To estimate the Logistic Normal parameters μ and Σ, we iterate between computing conditional modes, and updating μ, Σ. The algorithm is as follows

(1) Initialize μ_0, Σ_0.
(2) Until convergence,

[4] In practice, we find that (Polak-Riviere) Conjugate Gradient tends to be more dependable than the Newton-Raphson method in high dimensions. We used Carl Rasmussen's Conjugate Gradient Matlab code for this.

(a) For each document $j \in \{1, \ldots, m\}$, estimate $\hat{\mu}_j$ and $\hat{I}_j$ with respect to current model parameters μ_l and Σ_l.

(b) Update μ, Σ:

$$\mu_{l+1} = \frac{1}{m} \sum_{j=1}^{m} \hat{\mu}_j$$

$$\Sigma_{l+1} = \frac{1}{m} \sum_{j=1}^{m} \left[(\hat{\mu}_j - \mu_{l+1})(\hat{\mu}_j - \mu_{l+1})^T + \hat{I}_j^{-1} \right]$$

Appendix B. Variational Inference for the Correlated Topic Model

As in LDA, exact inference in the Correlated Topic Model is intractable. We describe the variational *Mean Field* approximation given by Blei and Lafferty. Suppose $P(x_h, x_v)$ is a distribution where x_h are latent variables and x_v are observable. Inference on this distribution involves the computation of $P(x_v)$ (the denominator in Bayes Rule) which is often an intractable integral. One possible way to get around this is to approximate P by a distribution for which inference is easier.

One view of Mean Field methods [JGJS99] is that they approximate the posterior $P(x_h|x_v)$ with a fully factorized distribution $Q = \prod_{i \in h} Q_i(x_i)$ by minimizing $KL(Q(x_h)||P(x_h|x_v))$. Another equivalent view is that one can maximize a lower bound $J(Q)$ on the observed data log-likelihood:

$$\begin{aligned} J(Q) &= \log P(x_v) - KL(Q(x_h)||P(x_v|x_h)) \\ &= \mathbb{E}_Q[\log P(x_h, x_v)] + H(Q) \end{aligned}$$

with respect to Q subject to the constraint that Q must be a fully factorized probability distribution. This alternative view of maximizing $J(Q)$ is appealing because it suggests the intuition that one should at once maximize both the expected complete data log-likelihood *and* the entropy of Q.

For the CTM, we use the approximating variational distribution

$$Q(\mathbf{v}_{1:K}, z_{1:N} | \lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} Q(\mathbf{v}_i | \lambda_i, \nu_i^2) \prod_{n=1}^{N} Q(z_n | \phi_n)$$

where each $Q(\mathbf{v}_i|\lambda_i, \nu_i^2)$ is a univariate gaussian with parameters λ_i, ν_i^2 and each $Q(z_n|\phi_n)$ is discrete with multinomial parameters ϕ_n. As we described in the previous paragraph, the objective is to find λ, ν^2, ϕ such that the following variational lower bound is maximized:

$$J(Q) = \mathbb{E}_Q[\log P(\mathbf{v}|\mu, \Sigma)] + \sum_{n=1}^{N} \left(\mathbb{E}[\log P(z_n|\mathbf{v})] + \mathbb{E}[\log P(w_n|z_n, \beta)] \right) + H(Q)$$

The second term of $J(Q)$ is not computable, and so yet another variational bound is introduced via a Taylor expansion:

$$\begin{aligned} \mathbb{E}_Q[\log P(z_n|\mathbf{v})] &= \mathbb{E}_Q[\mathbf{v}^T z_n] - \mathbb{E}_Q \left[1 + \log \left(\sum_{i=1}^{k-1} \exp\{v_i\} \right) \right] \\ &\geq \mathbb{E}_Q[\mathbf{v}^T z_n] - \left(\zeta^{-1} \left(\sum_{i=1}^{k-1} \mathbb{E}_Q[\exp\{v_i\}] \right) - 1 + \log(\zeta) \right) \end{aligned}$$

where ζ is a new variational parameter that must be fit.

Expanding everything out for $J(Q)$ yields the following (very ugly) expression:

$$\begin{aligned} J(Q) \quad = \quad & \frac{1}{2}\log|\Sigma^{-1}| - \frac{k-1}{2}\log 2\pi - \frac{1}{2}\left[Tr(\mathrm{diag}(\nu^2)\Sigma^{-1}) + (\lambda-\mu)^T\Sigma^{-1}(\lambda-\mu)\right] \\ & + \sum_{n=1}^{N}\left[\sum_{i=1}^{k-1}\lambda_i\phi_{n,i} - \zeta^{-1}\left(\sum_{i=1}^{k-1}\exp\left\{\lambda + \frac{\nu^2}{2}\right\}\right) + 1 - \log\zeta + \sum_{i=1}^{k-1}\phi_{n,i}\log\beta_{i,w_n}\right] \\ & + \sum_{i=1}^{k-1}\frac{1}{2}\left(\log\nu_i^2 + \log 2\pi + 1 - \sum_{n=1}^{N}\sum_{i=1}^{k-1}\phi_{n,i}\log\phi_{n,i}\right) \end{aligned}$$

To optimize $J(Q)$, Blei and Lafferty give coordinate ascent updates with respect to the variational parameters, which we reproduce here with a few minor changes. [5]

Holding all other parameters fixed, the optimal ζ is:

$$\hat{\zeta} = 1 + \sum_{i=1}^{k-1}\exp\{\lambda_i + \nu_i^2/2\}$$

Maximizing with respect to ϕ gives:

$$\hat{\phi}_{n,i} \propto \begin{cases} \exp\{\lambda_i\}\beta_{i,w_n} & \text{for } i \in \{1,\ldots,k-1\} \\ \beta_{i,w_n} & \text{for } i = k \end{cases}$$

There are no analytic expressions for maximimizing with respect to λ or ν^2, but they can be maximized numerically using a number of methods. The gradient with respect to λ is:

$$\frac{\partial J}{\partial \lambda} = -\Sigma^{-1}(\lambda-\mu) + \sum_{n=1}^{N}\phi_{n,1:K-1} - (N/\zeta)\exp\left(\lambda + \frac{\nu^2}{2}\right)$$

The derivative with respect to ν_i^2 is:

$$\frac{\partial J}{\partial \nu_i^2} = -\frac{1}{2\Sigma_{ii}} - \left(\frac{N}{2\zeta}\right)\exp\left(\lambda_i + \frac{\nu_i^2}{2}\right) + \frac{1}{2\nu_i^2}$$

And the second derivative is given by:

$$\frac{\partial^2 J}{\partial(\nu_i^2)^2} = -\frac{N}{4\zeta}\exp\left(\lambda_i - \frac{\nu_i^2}{2}\right) - \frac{1}{2(\nu_i^2)^2}$$

References

[AS80] J Aitchison and S.M. Shen, *Logistic-normal distributions: Some properties and uses*, Biometrika **67** (1980).

[BL05] David Blei and John Lafferty, *Correlated topic models*, Advances in Neural Information Processing Systems **18** (2005).

[BNJ02] David Blei, Andrew Ng, and Michael Jordan, *Latent dirichlet allocation*, Advances in Neural Information Processing Systems **14** (2002).

[FFP05] Li Fei-Fei and Pietro Perona, *A hierarchical bayesian model for learning natural scene, categories*, IEEE Computer Science Society International Conference of Computer Vision and Pattern Recognition, October 2005.

[GS04] T. Griffiths and M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences, 2004.

[5] In their NIPS '05 paper, the updates were given for an overparameterized version of the Logistic Normal distribution which, while being conceptually simpler, may lead to ill conditioned matrices in the algorithm.

[Hof03] Peter Hoff, *Nonparametric modelling of hierarchically exchangeable data*, Tech. report, Department of Statistics, University of Washington, 2003.

[JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul, *An introduction to variational methods for graphical models*, Machine Learning **37** (1999), no. 2, 183–233.

[Min00] Thomas Minka, *Estimating a dirichlet distribution*, Tech. report, Massachusetts Institute of Technology, 2000.

[SRE^{+}05] Josef Sivic, Bryan Russell, Alexei Efros, Andrew Zisserman, and William Freeman, *Discovering object categories in image collections*, Tech. Report AIM-2005-005, Massachusetts Institute of Technology, 2005.

E-mail address: `jch1@cs.cmu.edu`

E-mail address: `tomasz@cmu.edu`