Conditional Topic Random Fields

Jun Zhu JUNZHU@CS.CMU.EDU
Eric P. Xing EPXING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Generative topic models such as LDA are limited by their inability to utilize nontrivial input features to enhance their performance, and many topic models assume that topic assignments of different words are conditionally independent. Some work exists to address the second limitation but no work exists to address both. This paper presents a conditional topic random field (CTRF) model, which can use arbitrary nonlocal features about words and documents and incorporate the Markov dependency between topic assignments of neighboring words. We develop an efficient variational inference algorithm that scales linearly in terms of topic numbers, and a maximum likelihood estimation (MLE) procedure for parameter estimation. For the supervised version of CTRF, we also develop an arguably more discriminative max-margin learning method. We evaluate CTRF on real review rating data and demonstrate the advantages of CTRF over generative competitors, and we show the advantages of max-margin learning over MLE.

1. Introduction

Probabilistic topic models, such as the latent Dirichlet allocation (LDA) model (Blei et al., 2003), have been widely used for inferring a low dimensional representation that captures the latent semantics of textual or image documents. Such low dimensional representations can be used for classifying, clustering, or structurally browsing large corpora.

However, most existing topic models share two key characteristics which could limit their utility. First, they are essentially not "feature-based" models due to the generative nature of the models. More precisely, a topic model specifies the joint likelihood of all the introduced variables, which prevents flexible incorporation of nontrivial features in the data, such as non-local contextual or summary features in an article or image, because directly modeling such features as random variables would result in a prohibitively large state space that makes inference and learning very difficult, if at all possible. One may find convincing arguments of instead preferring a feature-based model for various applications from the celebrated paper on conditional random fields (CRF) (Lafferty et al., 2001). Second, with some exceptions (Gruber et al., 2007; Wallach, 2006; Chen et al., 2009), many topic models assume that the topic assignments of different text or image words are conditionally independent, and do not depend on the ordering of words. Such oversimplifying assumptions can be harmful in many critical applications such as scene classification.

Recently, a number of attempts have been made to address the limitation due to conditional-independence. At the data representation-level, "bag of region pairs" (Gökalp & Aksoy, 2007) or "doublets" (Sivic et al., 2005) are used to incorporate important spatial information in computer vision tasks; a bi-gram language model (Wallach, 2006) is used to consider word ordering information for text mining. At the latent topic-level, structured topic models with Markov properties (Gruber et al., 2007; Verbeek & Triggs, 2007; Wang et al., 2009b) or with latent permutations (Chen et al., 2009) have been proposed to capture the correlations between topic assignments of neighboring words. However, such models are all generative in nature, and cannot flexibly exploit non-trivial features of the entity in question.

To our knowledge, very little advance has been made in the direction of enriching feature usage under a topic model; and no successful attempt exists that considers enhancing feature usage and alleviating conditional-independency jointly. Maybe the most relevant work

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

along this direction is the Dirichlet-multinomial regression (DMR) model (Mimno & McCallum, 2008), which can use arbitrary document-level features to refine the Dirichlet parameter priors in LDA models. However, due to the fully generative nature of the underlying topic models, DMR cannot incorporate word-level features. Moreover, DMR assumes that the topic assignments of different words are conditionally independent. Thus this model still largely suffers from the two limitations discussed above.

In this paper, we propose a new model called *conditional topic random fields* (CTRF), which address both the feature and independence limitations in a simple but rich statistical framework; and we present efficient algorithms for training CTRFs. Underlying our proposed model is a key reliance on a *conditional* scheme, rather than a *generative* scheme, for defining the likelihood function of observations, which can be paralleled to the well-known contrast between a CRF and an HMM. A conditional model specifies a conditional likelihood of the observed documents, and can incorporate arbitrary over-lapping input features. (See (Lafferty et al., 2001) for a discussion on why this is feasible in a conditional model but not in a conventional generative model.) Our proposed model employs a general structure of the GLM to define a conditional distribution of latent topic assignments over words, rather than using a conventional multinomial model as in LDA. It retains all the remaining structure of the original topic models while incorporating arbitrary input features about words and documents; when new features are included, there is no need to alter this general design principle, therefore all the proposed inference and learning algorithms still apply. Moreover, CTRF directly incorporates the Markov dependency between the topic assignments of neighboring words, based on the same GLM principle.

Depending on the nature of training data and applications, like the LDA, CTRFs can be specialized into an unsupervised or a supervised version. We present a maximum likelihood estimation (MLE) algorithm for unsupervised CTRFs. For supervised CTRFs, which aim to discover predictive latent topic representations by incorporating commonly available side information, while the same MLE method can be still applied, we also present a max-margin learning algorithm which is arguably preferred for discriminative tasks. Finally, we demonstrate the advantages of CTRFs and max-margin learning on real review rating data.

The paper is structured as: Sec 2 present CTRFs. Sec 3 presents inference and MLE estimation. Sec 4 presents max-margin learning for supervised CTRFs. Sec 5 presents empirical results and Sec 6 concludes.

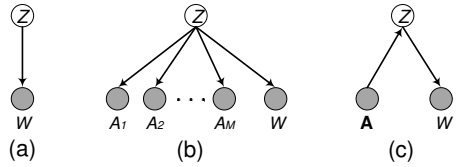

Figure 1. (a) the word generating model in LDA; (b) a generative model for a word and features; and (c) a conditional word generating model with features.

2. Conditional Topic Random Fields

For clarity and self-containedness, we begin with a brief Recap of the standard LDA and supervised LDA.

2.1. LDA and sLDA

The latent Dirichlet allocation (LDA) model (Blei et al., 2003) defines the document likelihood using a hierarchical Bayesian scheme. Specifically, a *topic proportion vector* is defined for each document, which is drawn from a Dirichlet distribution, and document words are each sampled from a topic-specific word distribution specified by a random drawn of the word-topic-assignment from the topic proportion vector. A supervised LDA (sLDA) (Blei & McAuliffe, 2007) introduces a response variable to incorporate commonly available side information (e.g., review rating scores) for discovering predictive topics. Let K be the number of topics; N be the number of the terms in a vocabulary; β be a $K \times N$ matrix; and each β_k be a distribution over the N terms. For a document d, the generating procedure of sLDA is

1. Draw a topic proportion vector $\theta_d|\alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) draw a topic assignment $z_{dn}|\theta \sim \text{Multi}(\theta_d)$.
   (b) draw a word $w_{dn}|z_{dn}, \phi \sim \text{Multi}(\phi_{z_{dn}})$.
3. Draw a response variable: $y|\bar{z}_d, \eta, \delta^2 \sim P(y|\bar{z}_d, \eta, \delta^2)$, where $\bar{z}_d = 1/N \sum_{n=1}^{N} z_{dn}$.

Here, Z_{dn} is a K-dimensional indicator variable that represents the topic assignment of w_{dn} and θ_d is a mixture proportion over topics of the document d. The sLDA model can be used for regression by defining a normal distribution P, or for classification by defining P as a logistic regression model (Wang et al., 2009a). To learn an sLDA, both maximum likelihood estimation (Blei & McAuliffe, 2007) and max-margin learning (Zhu et al., 2009) have been developed.

2.2. Conditional Topic Random Fields

The basic idea of CTRF can be understood from a simple word generating model. In LDA or sLDA, each word is represented as a mixture of latent topics Z as illustrated in Figure 1 (a), where the generating probability is estimated from a document-word count matrix. However, using the word and its count frequency as in LDA models is insufficient in resolving

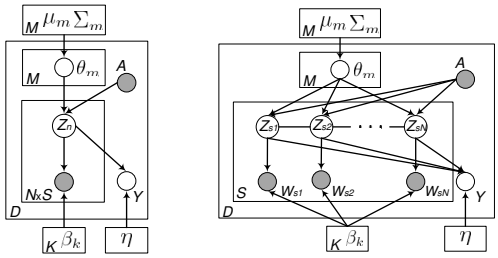

Figure 2. (Left) supervised conditional topic models; (Right) supervised linear chain CTRF.

the word's meaning ambiguity. For example, in a hotel review, the word "good" can be used to describe a positive aspect, while it can also be used to describe a negative aspect when used with a denying word, such as in the sentence "the service is not good". Therefore, exploring a rich set of input features (e.g., nonlocal contextual features) is expected to yield better models in terms of their discovered latent topics, and their performance on prediction tasks, such as regression.

To incorporate a rich set of useful features in a word generating model, a fully generative approach is shown in Figure 1 (b), where we use $\mathbf{a} = \{a_1, \cdots, a_M\}$ to denote a set of local and global features, such as POS tags. In this generative model, we need to assume that the features are conditionally independent and define a conditional distribution $p(a_i|z)$ for each feature i, such as a multinomial distribution for quantized code-words (Sivic et al., 2005) or a Gaussian distribution for continuous data (Welling et al., 2004; Xing et al., 2005). There have been extensive discussions on such models being difficult to learn, because of the possibly prohibitive parameterization cost of defining $p(w, a_{1:M}|z)$, the need of large training corpora to reliably estimate such parameters, and the complexity of unsupervised training due to induced coupling of all features when Z is unobserved. More critically, the inherited conditional independent assumptions in such a model is unrealistic and limits its utility.

An alternative approach to incorporating rich features in a word generating model is conditional modeling, which treats the features as conditions and directly defines the word generating distribution as illustrated in Figure 1(c). The conditional model defines a distribution $p(w|a_{1:M}) = \sum_z p(z|a_{1:M})p(w|z)$, which is just a mixture of latent topic components, where the mixing weights are defined via a richer log-linear model rather than a mere multinomial. As we have discussed, conditional models can incorporate arbitrary input features without changing the model structure and inference algorithm when new features are included.

Based on the conditional word generating model as in Figure 1(c), a *conditional topic model* (CdTM) can be defined as illustrated in Figure 2 (Left). Here, we assume that each document consists of S sentences and each sentence has N words. The generating procedure of CdTM is similar to that of LDA. The difference is that instead of using a multinomial distribution, the mixing weights over topics in CdTM are defined with a generalized linear model (GLM)

$$p(z_{sn}|\theta, \mathbf{a}) = \frac{\exp\{\theta^\top \mathbf{f}(z_{sn}, \mathbf{a})\}}{\sum_{z'_{sn}} \exp\{\theta^\top \mathbf{f}(z'_{sn}, \mathbf{a})\}},$$

where $\mathbf{f}$ is a vector of feature functions that are defined on arbitrary features related to words and documents. As a Bayesian model, the latent parameter variable θ can have an arbitrary prior. Here, we choose the independent multivariate normal distribution, i.e., $p(\theta|\mu, \Sigma) = \prod_m p(\theta_m|\mu_m, \Sigma_m) = \prod_m \mathcal{N}(\mu_m, \Sigma_m)$. One nice result of this choice is that the CdTM with a normal prior is a generalization of the correlated topic model (CTM) (Blei & Lafferty, 2006), which uses only one trivial feature that equals to one for any word.

Although the CdTM can address the feature limitation, it still suffers from the conditional independence limitation. In order to address both limitations, we further propose the *conditional topic random fields* (CTRF). The linear chain CTRF is shown in Figure 2 (Right), where the topic assignments of the words in the same sentence are mutually influenced through a conditional random field (Lafferty et al., 2001). The generating procedure of a CTRF is

1. For $m \in \{1, \cdots, M\}$, sample $\theta_m \sim p(\theta_m|\mu_m, \Sigma_m)$
2. For $s \in \{1, \cdots, S\}$
   (a) sample $z_s \sim \mathrm{P}_{ctrf}(z_s|\theta, \mathbf{a})$
   (b) for $n \in \{1, \cdots, N\}$, sample $w_{sn} \sim \mathrm{Multi}(\beta_{z_{sn}})$

where $\mathrm{P}_{ctrf}(z_s|\theta, \mathbf{a})$ is a conditional topic random field over the topic assignments of all the words in one sentence. By the random field theory (Lafferty et al., 2001), $\mathrm{P}_{ctrf}(z_s|\theta, \mathbf{a})$ has a log-linear form

$$p(z_s|\theta, \mathbf{a}) = \exp\Big(\theta^\top \mathbf{f}(\mathbf{a}, z_s) - \mathcal{A}_s(\theta, \mathbf{a})\Big),$$

where $\mathcal{A}_s(\theta, \mathbf{a}) = \log(\sum_{z_s} \exp(\theta^\top \mathbf{f}(\mathbf{a}, z_s)))$ is the log-partition function. For a linear chain CTRF, we have both *singleton* feature functions $f(\mathbf{a}, z_{sn})$ and *pairwise* feature functions $f(\mathbf{a}, z_{sn}, z_{sn+1})$, and $\mathbf{f}(\mathbf{a}, z_s) = \sum_n [\mathbf{f}(\mathbf{a}, z_{sn}) + \mathbf{f}(\mathbf{a}, z_{sn}, z_{sn+1})]$ is the cumulative feature function value on a sentence. For simplicity, we assume that a pairwise feature function $f(\mathbf{a}, z_{sn}, z_{sn+1})$ equals to zero if $z_{sn} \neq z_{sn+1}$. This assumption is true in many applications[1], for example, in text documents we would expect that neighboring words in the same sentence should be assigned to the same topic if they have some feature patterns. With this assumption, each θ_m is a K-dimensional vector and each element is associated with a (singleton or

[1] Gruber et al. (2007) made a more strict assumption that words in the same sentence are also in the same topic.

pairwise) topic assignment. We use $\mathbf{f}_m$ to denote a K-dimensional column vector of which only one element is non-zero according to the topic assignment; $\mathbf{f}$ is a MK-dimensional vector by stacking $\mathbf{f}_m$; θ is a stacking vector of θ_m; and $\mathcal{I}$ and $\mathcal{P}$ are the index sets of singleton and pairwise features, respectively.

Now, we explain the response variable Y in Figure 2, which has been ignored on purpose in the above discussions. As in sLDA (Blei & McAuliffe, 2007), the variable Y is for incorporating the commonly available supervised side information, such as rating scores in hotel reviews or scene categories associated with images (Wang et al., 2009a). By exploiting supervised side information, supervised CTRF (sCTRF) are expected to discover predictive topic representations, as have been demonstrated in (Blei & McAuliffe, 2007; Zhu et al., 2009; Wang et al., 2009a). Depending on the properties of Y, its generating model can be defined accordingly. For continuous Y, as in sLDA, $p(y|\bar{z}, \eta, \delta^2) = \mathcal{N}(\eta^\top \bar{z}, \delta^2)$, where $\bar{z} = \frac{1}{NS}\sum_{sn} z_{sn}$; and for categorical Y, the distribution can be a logistic regression model. Here, we consider the continuous case. Extension to the discrete case can be similarly done as in (Zhu et al., 2009; Wang et al., 2009a).

Since the supervised CTRF subsumes the unsupervised CTRF and CdTM, we will stick to the sCTRF when presenting the posterior inference and parameter estimation algorithms, and highlight the necessary modifications when applied to other models.

3. Posterior Inference and Estimation

The difficulty of inferring the posterior distribution of latent variables and estimating the unknown parameters in CTRF arises from two aspects. First, the topic distributions of different words in the same sentence are strongly correlated in the topic random fields. Second, the conditional distribution $p(z_s|\theta, \mathbf{a})$ contains a log-partition function that involves a summation over an exponential number of latent topic assignments. Since exact inference is intractable, we develop an efficient variational inference algorithm to obtain an approximation of the posterior distribution. Our method relies on a forward-backward message passing procedure on the linear-chain, which scales linearly with respect to the number of topics. For parameter estimation, we present the standard maximum likelihood estimation (MLE) that is applicable for estimating both supervised and unsupervised CTRF.

3.1. Posterior Inference

An Approximate Bound: By applying the Jensen's inequality, we obtain a lower bound of the log-likelihood of a document under an sCTRF, i.e., $\log p(y, \mathbf{w}|\mu, \Sigma, \eta, \delta^2, \beta) \geq L$, where

$$L \triangleq \sum_m \mathbb{E}[\log p(\theta_m|\mu_m, \Sigma_m)] + \sum_s \mathbb{E}[\log p(z_s|\theta, \mathbf{a})] + \sum_{sn} \mathbb{E}[\log p(w_{sn}|z_{sn}, \beta)] + \mathbb{E}[\log p(y|\bar{z}, \eta, \delta^2)] + H(q),$$

where $H(q) = -\mathbb{E}[\log q]$ is the entropy of the variational distribution q. Here, we make the mean-field assumption about q and it has the factorized form

$$q(\theta, \mathbf{z}|\lambda, \sigma^2, \psi) = \prod_{mk} q(\theta_{mk}|\lambda_{mk}, \sigma^2_{mk}) \prod_{sn} q(z_{sn}|\psi_{sn}),$$

where λ_{mk} and σ^2_{mk} are the mean and variance of the univariate normal distribution $q(\theta_{mk}|\lambda_{mk}, \sigma^2_{mk})$, and ψ_{sn} is the K-dimensional parameter of the topic distribution of z_{sn}. We will use σ^2 to denote the stacking vector of σ_{mk} similar to θ and define $\Lambda \triangleq \text{diag}(\sigma^2)$.

With this assumption, all the terms except the second one can be efficiently calculated, similar as in CTM (Blei & Lafferty, 2006). We omit the details for saving space. For the second term, we have

$$\mathbb{E}[\log p(z_s|\theta, \mathbf{a})] = \lambda^\top \mathbb{E}[\mathbf{f}(\mathbf{a}, z_s)] - \mathbb{E}[\mathcal{A}_s(\theta, \mathbf{a})],$$

which cannot be efficiently calculated because of the expectation of the log-partition function $\mathcal{A}_s(\theta, \mathbf{a})$. Therefore, we need to further relax the lower bound L. Let $h_s(q) = \sum_{z_s} \mathbb{E}[\exp(\theta^\top \mathbf{f}(\mathbf{a}, z_s))]$, we have[2]

$$\mathbb{E}[\mathcal{A}_s(\theta, \mathbf{a})] \leq \frac{h_s(q)}{\zeta_s} - 1 + \log \zeta_s, \tag{1}$$

where ζ_s is a positive variational parameter.

With the assumption of q and a little algebra, we have

$$\mathbb{E}[\exp(\theta^\top \mathbf{f}(\mathbf{a}, z_s))] = \exp\{\lambda^\top \mathbf{f}(\mathbf{a}, z_s) + \frac{\mathbf{f}(\mathbf{a}, z_s)^\top \Lambda \mathbf{f}(\mathbf{a}, z_s)}{2}\}.$$

For the univariate CdTM, $h_s(q)$ can be exactly calculated. However, for a non-trivial chain-structured random field, since the quadratic term $\mathbf{f}(\mathbf{a}, z_s)^\top \Lambda \mathbf{f}(\mathbf{a}, z_s)$ couples the topic assignments of all the words in one sentence, it is intractable to exactly compute $h_s(q)$. Here, we approximate[3] this term as

$$\mathbb{E}[\exp(\theta^\top \mathbf{f}(\mathbf{a}, z_s))] \approx \prod_n \mathbb{E}[\exp\{\theta^\top \mathbf{f}_n(\mathbf{a}, z_s)\}] = \prod_n \exp\{\lambda^\top \mathbf{f}_n(\mathbf{a}, z_s) + \frac{\mathbf{f}_n(\mathbf{a}, z_s)^\top \Lambda \mathbf{f}_n(\mathbf{a}, z_s)}{2}\}, \tag{2}$$

where $\mathbf{f}_n(\mathbf{a}, z_s) = \mathbf{f}(\mathbf{a}, z_{sn}, z_{sn+1})$. Now, we can develop a forward-backward message passing procedure to compute $h_s(q)$. Specifically, we use $\mathbf{f}^2_n(\mathbf{a}, z_s)$ to denote the element-wise product $\mathbf{f}_n(\mathbf{a}, z_s) \cdot \mathbf{f}_n(\mathbf{a}, z_s)$. Then, we have

$$h_s(q) = \sum_{z_s} \exp\Big(\sum_n \lambda^\top \mathbf{f}_n(\mathbf{a}, z_s) + \sum_n \frac{1}{2}(\sigma^2)^\top \mathbf{f}^2_n(\mathbf{a}, z_s)\Big).$$

[2] Due to the inequality $\log x \leq a^{-1}x - 1 + \log a, \ \forall a > 0$.

[3] By assuming $\mathbf{f}_n(\mathbf{a}, Z_s)$ are independent.

Since both $\mathbf{f}_n$ and $\mathbf{f}_n^2$ depend only on the cliques (i.e., nodes and edges in a linear chain CTRF), $h_s(q)$ and its gradients $\nabla_\lambda h_s(q)$ and $\nabla_{\sigma^2} h_s(q)$ can be efficiently computed with a forward-backward massage passing procedure (Lafferty et al., 2001), which has a complexity $\mathcal{O}(NK)$ because the transition matrices are diagonal due to the assumption of pairwise feature functions. Our empirical studies show that this simple approximation method works well. Developing a tighter approximation is our future work.

Now, substituting the above inequality in Eq. (1) and the approximation in Eq. (2) into the variational lower bound L, we can get an approximate bound $\mathcal{L}$, which is a function of $(q, \zeta, \mu, \Sigma, \eta, \delta^2, \beta)$.

Inference: Given the parameters $(\mu, \Sigma, \eta, \delta^2, \beta)$, the inference can be done with a coordinate ascent method, iteratively optimizing $\mathcal{L}$ with respect to each variational parameter while holding all the others fixed. Let q^{old} be the current solution of q, the update rule for ζ_s is simple, i.e., $\zeta_s = h_s(q^{old})$. For the parameters ψ_{sn}, we have the update rule

$$\psi_{snk} \propto \exp\Big(\log \beta_{kw_{sn}} + \sum_{m\in\mathcal{I}} \lambda_m^\top \mathbf{f}_m(\mathbf{a}, k) + \sum_{i\in\mathcal{N}_n} m_{i\to n}(k) + \frac{y}{SN\delta^2}\eta_k - \frac{2\eta^\top \psi_{-sn} + \eta_k}{2S^2N^2\delta^2}\eta_k\Big), \quad (3)$$

where $\mathcal{N}_n$ is the set of neighbors of the node n; $m_{i\to n}(k) = \sum_j \psi_{sij} \sum_{m\in\mathcal{P}} \lambda_m^\top \mathbf{f}_m(\mathbf{a}, j, k)$ is the mean field message from node i; and $\psi_{-sn} = \sum_{ij} \psi_{ij} - \psi_{sn}$. From this update rule, we can explicitly see the additive contributions from different parts in the topic distribution. The first term is from the multinomial word generating model; the second term considers the local features; the third term incorporates the mutual dependency between neighbors; and the last two terms are from the supervised response model. For unsupervised CTRF, the update rule contains only the first three terms.

For the parameters λ and σ^2, since we cannot get closed form update rules, we apply gradient descent methods. The gradients for λ and σ^2 are

$$\nabla_{\lambda_m} \mathcal{L} = \gamma_m + \sum_s \mathbb{E}[\mathbf{f}_m(\mathbf{a}, z_s)] - \sum_s \zeta_s^{-1} \nabla_{\lambda_m} h_s(q)$$

$$\nabla_{\sigma^2_{mk}} \mathcal{L} = -\frac{1}{2}\Sigma^{-1}_{kk} + \sum_s \zeta_s^{-1} \nabla_{\sigma^2_{mk}} h_s(q) + \frac{1}{2\sigma^2_{mk}},$$

where $\gamma_m = \Sigma_m^{-1}(\lambda_m - \mu_m)$. Note that the Newton method as used in (Blei & Lafferty, 2006) optimizes with respect to each dimension of σ^2 in the log-domain. This coordinate ascent procedure is expensive, especially when the number of features is large. Here, we use the much faster L-BFGS method (Liu & Nocedal, 1989) for λ and $\log \sigma^2$ jointly.

3.2. Max-Likelihood Estimation

Given training set $\mathcal{D}$, where each document d is associated with a true response y_d, parameter learning is to estimate the topics β, Gaussian parameters (μ, Σ), and the response model parameters (η, δ^2). The most common method is maximum likelihood estimation (MLE), which has been used to learn LDA and sLDA.

For MLE, we optimize the approximate bound $\sum_d \mathcal{L}_d$ with a variational EM procedure, which iteratively performs an E-step and an M-step until the bound converges. In E-step, we perform variational inference for each document as discussed above. In M-step, we maximize $\sum_d \mathcal{L}_d$ with respect to the unknown parameters. This can be performed similarly as in CTMs (Blei & Lafferty, 2006) for (μ, Σ, β) and sLDAs (Blei & McAuliffe, 2007) for (η, δ^2). We omit the details due to space limitation.

4. Joint Max-margin & Max-Likelihood Learning for sCTRF

For supervised CTRFs, our goal is to discover predictive latent topic representations that are suitable for prediction tasks. Besides MLE, another arguably more discriminative approach is the max-margin learning, which has been applied to learn discriminative latent topic models in MedLDA (Zhu et al., 2009). In this section, we present an alternative margin-based discriminative approach for learning supervised CTRFs.

With the approximate bound $\mathcal{L}$, we formally define a joint max-margin and max-likelihood estimator of sCTRF as the solution to the following problem

$$\min_{\Theta, q, \xi, \xi^\star} \; -\sum_d \mathcal{L}_d(\Theta, q) + \frac{1}{2}\gamma\eta^\top\eta + C\sum_{d=1}^D (\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d: \begin{cases} y_d - \eta^\top \mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d \\ -y_d + \eta^\top \mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d^\star \end{cases}$$

where γ and C are regularization constants; ϵ is a precision parameter; ξ are non-negative slack variables; and Θ denotes all the parameters $(\mu, \Sigma, \eta, \delta^2, \beta)$.

This constrained optimization problem can be efficiently solved with Lagrangian methods. We introduce a pair of lagrange multipliers α and $\alpha^\star$ for the two constraints associated with each document, and we perform alternative minimization over the Lagrangian functional. With the same mean field assumption of q, we get the same gradients for λ and σ^2 in the posterior inference as above. For ψ, the update rule is similar to Eq. (3), but with an additional term

$$\psi_{snk} \propto \exp\Big(\log \beta_{kw_{sn}} + \sum_{m\in\mathcal{I}} \lambda_m^\top \mathbf{f}_m(\mathbf{a}, k) + \sum_{i\in\mathcal{N}_n} m_{i\to n}(k) + \frac{y}{SN\delta^2}\eta_k - \frac{2\eta^\top \psi_{-sn} + \eta_k}{2S^2N^2\delta^2}\eta_k + \frac{\alpha - \alpha^\star}{SN}\eta_k\Big),$$

where the last term arises from the max-margin constraints and it plays a role of regularizing the topic assignment. Once the model makes a bad prediction for document d, one of the lagrange multipliers α_d and $\alpha_d^\star$ will be non-zero. Then, the last term in the exponential will bias the topic assignment to favor a better prediction. This regularization effect will yield a more predictive topic representation. For μ, Σ, β and δ^2, the update rules are the same as those in MLE. Finally, the parameter η can be efficiently estimated by solving a standard SVM regression problem, which is the same as in MedLDA (Zhu et al., 2009).

5. Experiments

In this section, we report empirical results of the supervised CTRF on real review rating data. Our goals are to demonstrate that sCTRF can discover good topic representations and can make accurate predictions.

5.1. Data Set

We build a real data set by randomly crawling hotel reviews from TripAdvisor[4], where each review is associated with a global rating score and five aspect rating scores for the aspects– *Value*, *Rooms*, *Location*, *Cleanliness*, and *Service*. This data set is very interesting and can be used for many data mining tasks, for example, extracting the textual mentions of each aspect (Titov & McDonald, 2008). In these experiments, we focus on predicting the global rating scores for reviews.

To avoid too short and too long reviews, we only keep those reviews whose character length is between 1500 and 6000. On TripAdvisor, the global ratings rank from 1 to 5. We randomly select 1000 reviews for each rating and the data set consists of 5000 reviews in total. We uniformly partition it into training and testing sets. For each review, we use the NLProcessor[5] to do part-of-speech (POS) tagging and noun phrase (NP) chunking, and extract the following features:

. **POS-Tag**: We distinguish four types of POS tags, that is, Adjective, Noun, Adverb, and Verb. Each type includes all its subcategories, e.g., Adjective includes "JJ" (Adjective), "JJR" (comparative Adjective), and "JJS" (superlative Adjective).

. **WordNet**: WordNet[6] is a large lexical database of English. We navigate it with some seeds of positive (e.g., good, excellent, etc) and negative (e.g., bad, painful, etc) words, and identify whether a word is positive or negative based on the synonym and antonym relationship. Words without strong relationship with the seeds are treated as neutral. For a positive or negative word, we also identify whether a denying word (e.g., not, no, etc.) appears before it within a word distance of 4.

[4]http://www.tripadvisor.com/

[5]http://www.infogistics.com/textanalysis.html

[6]http://wordnet.princeton.edu/

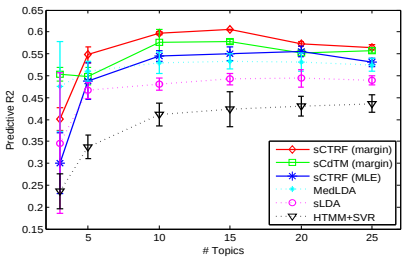

Figure 3. The predictive R^2 of different models.

. **NP-Chunking**: We define pairwise feature functions for those words that are in the same noun or verb phrase, or the conjunction "and" or "or" appears between them.

By removing a standard list of stopping words and those terms whose count frequency is less than 5, we build a dictionary with 12000 terms.

5.2. Prediction Accuracy

Similar as in (Blei & McAuliffe, 2007), we treat the problem of predicting rating scores as a regression problem. We take logarithm to make the response variables approximately normal. We compare sCTRF with sCdTM, sLDA, MedLDA, and hidden Markov topic models (HTMM) (Gruber et al., 2007). For the unsupervised HTMM, we feed the discovered topic representations to a linear support vector regression (SVR) (Smola & Schölkopf, 2003) to make it consistent with the regression model as integrated in MedLDA and sCTRF. For sCTRF, we report two sets of results achieved with maximum likelihood estimation and max-margin training.

Fig. 3 shows the predictive R^2 (Blei & Lafferty, 2006) scores of different models. First, since supervised topic models can leverage the side information (e.g., rating scores) to discover latent topic representations, they generally outperform the decoupled two-step procedure as adopted in unsupervised topic models. Second, the feature-based sCdTM (with max-margin training) outperforms MedLDA, which only uses the word count feature. Third, sCTRF (max-margin) slightly outperforms sCdTM (max-margin) because of the incorporation of Markov dependency[7]. Therefore, the reason for the superior performance of sCTRF (MLE or max-margin) compared to sLDA or MedLDA is because sCTRF can incorporate both rich input features and Markov dependency. Finally, the max-margin based methods (e.g., sCTRF (max-margin)) generally outperform likelihood-based methods (e.g., sCTRF (MLE)) due to a regularization effect introduced by constraints. For a linear SVR with word count features, the predictive R^2 is about 0.56, which is comparable to the best performance of MedLDA but worse than the best performance of max-margin sCTRF.

[7]Similar conclusions observed for MLE method.

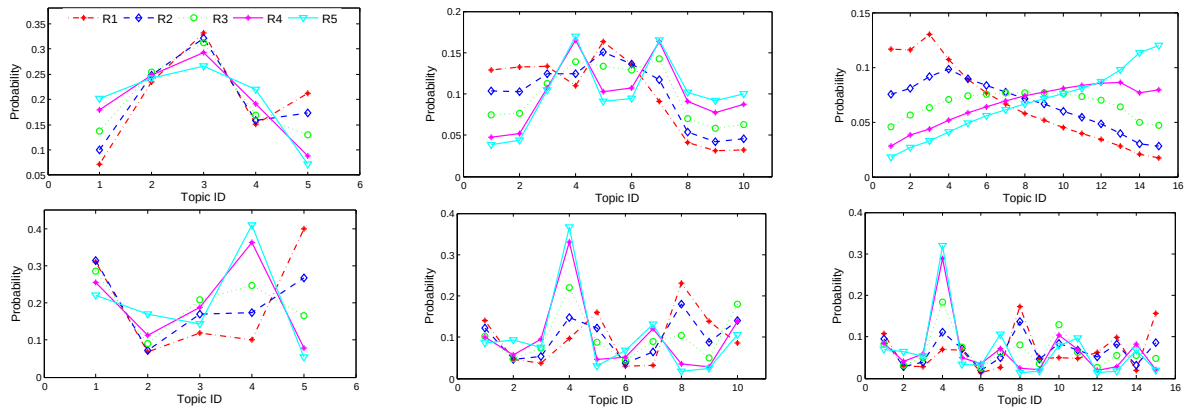

Figure 4. The average distribution over topics for documents of the same rating score. The top row is for max-margin sCTRF and the bottom row is for MedLDA. For each model, we use 5 (Left), 10 (Middle), and 15 (Right) topics.

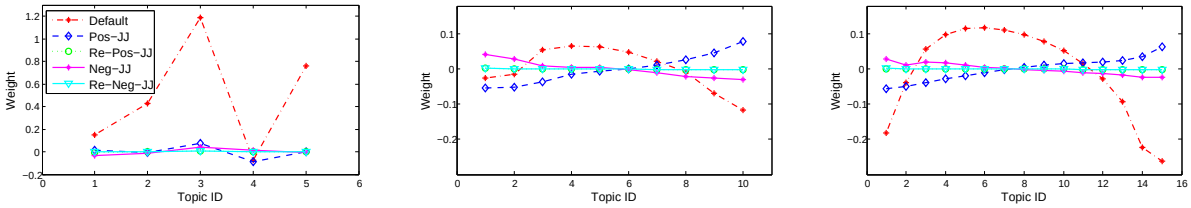

Figure 5. The weights of five features in max-margin sCTRF with topic number 5 (Left), 10 (Middle), and 15 (Right).

5.3. Characterization of Topic Modeling

In this section, we show some interesting properties of sCTRF on topic modeling, which provide insights for the outstanding performance in regression.

5.3.1. Average Topic Distribution

Fig. 4 shows the average mixture distribution over topics for the two best representatives–max-margin sCTRF and MedLDA. The average is taken over testing documents that have the same rating score. We denote the rating scores from small to large by $R_1, \cdots$, and R_5. We use curves that connect the topic-probability points to show the trends of probability change. We can see that the curves of sCTRF show a consistent smooth change from R_1 to R_5. For example, when the topic number is 10 or 15, the lowest rating R_1 has high probabilities on the first several topics while low probabilities on the last several topics; but the highest rating R_5 is on the opposite side, i.e., having low probabilities on the first several topics and high probabilities on the last several topics. In fact, as shown in Table 1 for a 10 topic sCTRF, topics T1 and T2 are about the negative side of a hotel, while topics T9 and T10 are on the positive side. Therefore, reviews with a low rating score have high probabilities to describe negative aspects (e.g., T1) but low probabilities to describe positive aspects (e.g., T10). Overall, for the ratings from R_1 to R_5, the probability mass is smoothly changed with more probability distributed on positive topics and less probability on negative topics. When topic number is small (e.g., 5), the topics do not show a regular negative/positive pattern due to the dominating effect of the *Default* feature, as we shall see.

The smooth change of the probability curves in the max-margin sCTRF result in a better prediction in a linear regression model, as shown in Fig. 3. However, for MedLDA, the probability curves tend to mix together and do not show a consistently smooth change. Moreover, the topics discovered by MedLDA do not show a similar positive/negative pattern[8], as detailed below. Similarly, the likelihood-based sCTRF obtains a better topic representation than sLDA and a better predictive R^2. Details are deferred to a full version.

5.3.2. Feature Weights and Sample Topics

Fig. 5 shows the weights of the feature functions that are defined with five features: *Default*–equal to one for any word; *Pos-JJ*–positive adjective; *Neg-JJ*–negative adjective; *Re-Pos-JJ*–positive adjective that has a denying word before it; and *Re-Neg-JJ*–negative adjective that has a denying word before it. Note that the *Default* feature is equivalent to the word counts used in LDA models. Again, we use curves that connect the topic-weight points to show the change trends. We can see that when the topic number is small, the default feature dominates; but when the topic number is large (e.g., 10 or 15), the default feature tends to discover prominent latent topics that are common for all the documents, e.g., T3 to T8 in Table 1. In contrast, both the positive and negative adjective features tend to discover topics that are more discriminative for rating prediction, e.g., T1 and T10.

[8] All results are automatic, without sorting of the topics.

Table 1. Top words in different topics by 10 topic max-margin sCTRF and 10 topic MedLDA.

Max-Margin CTRF										MedLDA									
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
told	booked	place	hotel	room	room	hotel	beach	beach	great	hotel	place	room	room	beach	hotel	hotel	hotel	resort	hotel
dirty	room	hotel	food	hotel	hotel	area	pool	resort	good	location	day	day	table	pool	room	room	pool	beach	people
room	told	room	bar	check	parking	staff	resort	pool	nice	good	house	hotel	bed	resort	breakfast	told	food	night	room
front	front	days	day	stay	desk	pool	food	ocean	lovely	guests	stay	night	breakfast	room	rooms	asked	good	food	desk
asked	asked	time	pool	night	stay	breakfast	island	island	beautiful	service	great	stay	shower	area	stay	staff	beach	time	front
hotel	hotel	day	time	rooms	night	day	kids	kids	excellent	place	time	desk	bathroom	view	area	day	bar	pool	stay
bad	back	night	service	bed	breakfast	view	trip	good	wonderful	area	trip	front	night	nice	good	manager	night	room	day
small	called	people	holiday	floor	bed	location	service	great	comfortable	staff	back	told	place	ocean	staff	back	day	great	time
worst	manager	stay	room	desk	floor	service	day	restaurants	beach	rooms	beautiful	door	coffee	great	nice	reception	room	service	parking
poor	left	water	people	door	rooms	walk	staff	enjoyed	friendly	time	island	rooms	small	bedroom	bathroom	rooms	holiday	good	pool
manager	stay	rooms	night	time	time	restaurant	restaurants	loved	fresh	restaurant	lovely	place	hotel	chairs	floor	time	staff	day	nice
called	work	experience	restaurant	morning	day	time	time	trip	large	stay	dinner	back	tv	unit	night	stay	people	staff	service
rude	decided	food	water	back	clean	food	view	area	amazing	pool	experience	floor	water	kids	small	night	area	water	back
reception	finally	resort	drinks	stayed	staff	room	area	food	fantastic	experience	food	time	looked	island	n't	service	restaurant	people	check
broken	checked	staff	rooms	day	bathroom	bit	ocean	walk	perfect	site	night	bed	door	time	bed	booked	great	nice	lobby

Table 1 shows the top words for the topics discovered by 10 topic max-margin sCTRF and MedLDA. We can see that for sCTRF the topics show a regular positiveness/negativeness pattern. For example, the topics T1 and T2 are on the negative aspects of a hotel while T9 and T10 describe positive aspects, and in between, the topics from T3 to T8 tend to be neutral and describe the most common aspects, such as room, food, staff, etc. Overall, from T1 to T10, the positiveness of the topics increases while the negativeness decreases. This is consistent with the weight change of the *Pos-JJ* and *Neg-JJ* features over topics as shown in Fig. 5. However, the topics discovered by MedLDA do not show a regular pattern on positiveness or negativeness. For example, the positive words such as "good", "nice", or "great" appear in most of the topics, and within each topic the words are not so coherent, in terms of POS tags or meanings. The reason for discovering such unpurified topics is because MedLDA only uses the *Default* (i.e., word count) feature, which tends to discover common aspects of a hotel, as shown in Fig. 5.

6. Conclusions

We have presented the conditional topic random field (CTRF), a general framework that incorporates arbitrary input features in latent topic models and the Markov dependency between topic assignments of neighboring words. We develop efficient inference and MLE parameter estimation algorithms. For the supervised CTRF, we also develop an arguably more discriminative max-margin learning method. On real review rating data, we demonstrate the interesting characterization of CTRF on topic modeling and the advantages of max-margin training on prediction tasks.

Acknowledgements

This work is supported by ONR N000140910758, NSF IIS-0713379, NSF Career DBI-0546594, and an Alfred P. Sloan Research Fellowship to EPX.

References

Blei, D. & Lafferty, J. Correlated topic models. *NIPS*, 2006.

Blei, D. & McAuliffe, J. Supervised topic models. *NIPS*, 2007.

Blei, D., Ng, A., & Jordan, M. Latent Dirichlet allocation. *JMLR*, (3):993–1022, 2003.

Chen, H., Branavan, S.R.K., Barzilay, R., & Karger, D. Content modeling using latent permutations. *JAIR*, (36):129–163, 2009.

Gökalp, D. & Aksoy, S. Scene classification using bag-of-regions representations. *CVPR*, 2007.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. Hidden topic markov models. *AISTATS*, 2007.

Lafferty, J., McCallum, A., & Pereira, F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

Liu, D. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, (45):503–528, 1989.

Mimno, D. & McCallum, A. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *UAI*, 2008.

Sivic, J., Russell, B., Efros, A., Zisserman, A., & Freeman, W. Discovering objects and their locatioins in images. *ICCV*, 2005.

Smola, A. & Schölkopf, B. A tutorial on support vector regression. *Statistics and Computing*, 2003.

Titov, I. & McDonald, R. A joint model of text and aspect ratings for sentiment summarization. *ACL*, 2008.

Verbeek, J. & Triggs, B. Region classification with markov field aspect models. *CVPR*, 2007.

Wallach, H. Topic modeling: Beyond bag-of-words. *ICML*, 2006.

Wang, C., Blei, D., & Fei-Fei, L. Simultaneous image classification and annotation. *CVPR*, 2009a.

Wang, C., Thiesson, B., Meek, C., & Blei, D. Markov topic models. *AISTATS*, 2009b.

Welling, M., Rosen-Zvi, M., & Hinton, G. Exponential family harmoniums with an application to information retrieval. *NIPS*, 2004.

Xing, E., Yan, R., & Hauptmann, A. Mining associated text and images with dual-wing harmoniums. *UAI*, 2005.

Zhu, J., Ahmed, A., & Xing, E. MedLDA: Maximum margin supervised topic models for regression and classification. *ICML*, 2009.