NOTES OF CONDITIONAL RANDOM FIELD

XIANGLI CHEN

1. Introduction

HMMs and stochastic grammars [2] are generative models, assigning a joint probability to paired observation and label sequences. A generative model needs to enumerate all possbile observation sequences that the inference problem for such model is intractable. This difficulty is one of the main motivation for looking at conditinal models as an alternative. A conditional model does not expand modeling effort on the ovservations and the conditional probability of the label sequence can depend on arbitrary, non-independent features of the observation sequence without forcing the model to account for the distribution of those dependencies. Maximum entropy Markov models (MEMMs) [3] are conditional probabilistic sequence models. But like other non-generative finite-state models based on next-state classifiers, such as discriminative markov models, MEMMs has the label bias problem: the transitions leaving a given state compete only against each other, rather than against all other transitions in the model.

2. Conditional Random Field

Conditonal random field (CRFs)[2] is a sequence modeling framework that has all the advantages of MEMMs but also solves the label bias problem in a principled way. CRFs perform better than HMMs and MEMMs when the true data distribution has higher-order dependencies than the model, as is ofter the case in practice. If the graph $G = (V, E)$ of Y is a tree and the joint distribution is positive, by Hammersley Clifford Theorem[1], the conditional model takes the form:

$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{1}{P(X)Z}\phi(X)\prod_{e\in E}\phi_e(Y_e, X)\prod_{v\in V}\phi_v(Y_v, X).$$

where Z is the normalized constant with respect to the joint distribution. We can also denote it to be:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{e\in E}\psi_e(\mathbf{y}_e, \mathbf{x}) + \sum_{v\in V}\psi_v(\mathbf{y}_v, \mathbf{x})\right)$$

In addition, we assume the conditional model is a probability distribution of exponential families,

$$p(\mathbf{y}|\mathbf{x}) = \exp(\phi(\mathbf{x}, \mathbf{y})\cdot\theta - g(\theta|\mathbf{x})).$$

Apply Hammersley Clifford Theorem for exponential families,

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{e\in E}\phi_e(\mathbf{y}_e, \mathbf{x})\cdot\theta_e + \sum_{v\in V}\phi_v(\mathbf{y}_v, \mathbf{x})\cdot\theta_v\right).$$

Date: July 6, 2017.

We consider a chain-structured CRFs with label sequence $Y = (Y_1, Y_2, ..., Y_n)$. The cliques are (x, y_i) and y_{i-1}, y_i's. In general, CRFs assume the unknown parameter does not depend on the position of sequence. As a consequence, the conditional model takes the following form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{n}\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, \mathbf{x}, i) + \sum_k \mu_k s_k(y_i, \mathbf{x}, i)\right)\right)$$

where t_j's and s_k's are feature functions. Note that the cliques are (x, y_i) and y_{i-1}, y_i's or more precisely formulation would be:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{n}\left(\sum_j \lambda_j t_j(y_{i-1}, y_i) + \sum_k \mu_k s_k(y_i, x_i)\right)\right).$$

To obtain a more uniform notation, we specifies a feature vector $\mathbf{f}(\mathbf{y}, \mathbf{x}, i)$. An element of the feature vector $f_j(\mathbf{y}, \mathbf{x}, i)$ is either a state feature $s_j(y_i, \mathbf{x}, i)$ or transitive feature $t_j(y_{i-1}, y_i, \mathbf{x}, i)$. (In general, we let $t_j(y_{i-1}, y_i, \mathbf{x}, i) = 0$, when $i = 1$.) Let,

$$F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(\mathbf{y}, \mathbf{x}, i).$$

Then a general conditional model for CRFs is,

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j F_j(\mathbf{y}, \mathbf{x})\right) = \frac{1}{Z(\mathbf{x})} \exp\left(\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})\right).$$

We can train a CRF[4] by maximizing the log-likelihood of a given training set $T = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{N}$.

$$\mathcal{L}_{\boldsymbol{\lambda}} = \sum_k \log p_{\boldsymbol{\lambda}}(\mathbf{y}_k|\mathbf{x}_k)$$

The gradient with respect to $\boldsymbol{\lambda}$,

$$\nabla \mathcal{L}_{\boldsymbol{\lambda}} = \sum_k \left[\mathbf{F}(\mathbf{y}_k, \mathbf{x}_k) - \mathbb{E}_{p_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}_k)} \mathbf{F}(\mathbf{Y}, \mathbf{x}_k)\right].$$

3. Dynamic Programming

The chanllenging part of computing the gradient of CRF is $\mathbb{E}_{p_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}_k)} \mathbf{F}(\mathbf{Y}, \mathbf{x}_k)$. Here we use dynamic programming method. Let $y \in \mathcal{Y}$, given $\mathbf{x}$, define n transition matrix $\{\mathbf{M}_i(\mathcal{Y}, \mathbf{x}) | i = 1, ..., n\}$, where each $\mathbf{M}_i(\mathcal{Y}, \mathbf{x})$ is a $|\mathcal{Y} \times \mathcal{Y}|$ matrix with elements of the form:

$$\mathbf{M}_i[y_{i-1} = y, y_i = y'] = \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(y_{i-1} = y, y_i = y', \mathbf{x}, i)\right)$$

For each λ_j, define n feature matrix $\{\mathbf{Q}_{ji}(\mathcal{Y}, \mathbf{x}) | i = 1, .., n\}$, where each $\mathbf{Q}_{ji}(\mathcal{Y}, \mathbf{x})$ is a $|\mathcal{Y} \times \mathcal{Y}|$ matrix with elements of the form:

$$\mathbf{Q}_{ji}[y_{i-1} = y, y_i = y'] = f_j(y_{i-1} = y, y_i = y', \mathbf{x}, i)$$

Note that $F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f(y_{i-1}, y_i, \mathbf{x}, i)$. Then,

$$\begin{aligned} \mathbb{E}_{p_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}_k)} F_j(\mathbf{Y}, \mathbf{x}_k) &= \sum_{\mathbf{y} \in \mathcal{Y}} p_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}_k) F_j(\mathbf{y}, \mathbf{x}_k) \\ &= \sum_{i=1}^{n} \frac{\alpha_{i-1}^T (\mathbf{Q}_{ji} \circ \mathbf{M}_i) \beta_i}{Z_{\boldsymbol{\lambda}}(\mathbf{x})} \end{aligned}$$

where $\circ$ is the element-wise matrix product, $Z_{\boldsymbol{\lambda}}(\mathbf{x})$ is the norm constant, α_i and β_i is defined by:

$$\alpha_i^T = \begin{cases} \alpha_{i-1}^T \mathbf{M}_i & 0 < i \leq n \\ \mathbf{1}^T & i = 0 \end{cases}.$$

$$\beta_i = \begin{cases} \mathbf{M}_{i+1} \beta_{i+1} & 0 \leq i < n \\ \mathbf{1} & i = n \end{cases}.$$

Note that $Z_{\boldsymbol{\lambda}}(\mathbf{x}) = \alpha_n^T \mathbf{1}$.

References

[1] John M Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. 1971.

[2] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[3] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, pages 591–598, 2000.

[4] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

Computer Science Department, University of Illinois at Chicago, Chicago, IL 60607
E-mail address: xchen40@uic.edu
URL: https://www.cs.uic.edu/~xchen/