Weighted Lower Linear Envelope Potentials

Chang Li

15. Januar 2016

1 Introduction

2 Proposal

2.1 Background

2.2 Extensions

2.2.1 Weighted Lower Linear Envelope and Large Margin Optimization

Suppose we have a binary MRFs $\mathbf{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{0, 1\}$. A higher-order potential $\psi_c^H(\mathbf{y}_c)$ is an arbitrary function defined on cliques $\mathbf{y}_c = \{y_i : i \in c\}$ where $c \subseteq \{1, \ldots, n\}$. Gould[1] proposed it with a weighted lower linear envelope potential expression

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_{k=1,\ldots,K} \left\{ a_k W_c(\mathbf{y}_c) + b_k \right\} \tag{1}$$

where $(a_k, b_k) \in \mathbb{R}^2$ are linear function parameters and

$$W_c(\mathbf{y}_c) = \sum_{i \in c} w_i^c y_i$$

where c is a clique. w_i^c is a per-variable non-negative weights for every nodes in each clique and satisfies $\sum_{i \in c} w_i^c = 1$.

Gould[1] introduced a non-negative linear parameter vector $\boldsymbol{\theta} \in \mathbb{R}^K$ and a feature map $\boldsymbol{\phi}(\mathbf{y}) \in \mathbb{R}^K$

$$\theta_k = \begin{cases} b_1 & \text{for } k = 0 \\ a_1 & \text{for } k = 1 \\ a_{k-1} & \text{for } k = 2, \ldots, K \end{cases}$$

$$\phi_k = \begin{cases} 1 & \text{for } k = 0 \\ W(\mathbf{y}) & \text{for } k = 1 \\ \left(\dfrac{k-1}{K} - W(\mathbf{y}) \right) \left[\left[W(\mathbf{y}) > \dfrac{k-1}{K} \right] \right] & \text{for } k = 2, \ldots, K \end{cases}$$

to rewrite the Equation 1 into a linear formulation

$$\psi_c^H(\mathbf{y}_c) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{y}) \tag{2}$$

where $D\boldsymbol{\theta} \geq \mathbf{0}$. Hence the Equation 2 can be optimized using max-margin framework with an additional linear constraint.

2.2.2 Structured SVM with latent variables

Suppose we have a combined feature function $\Psi(\mathbf{x}, \mathbf{y})$ defined on a sample set $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ describes the relationship between input $\mathbf{x}$ and structured output $\mathbf{y}$. $\mathbf{x}_i$ denots the ith sample and $\mathbf{y}_i$ denotes its structured label vector. The Structured SVM approach [2] is to maximize a linear discriminant function $\mathcal{F} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for a given input $\mathbf{x}_i$ to derive the structured label $\mathbf{y}_i$

$$f_w(x) = \arg\max_{\mathbf{y} \in \mathcal{Y}} F_w(\mathbf{x}, \mathbf{y})$$

Since $\mathcal{F}$ is linear, we can rewrite this as

$$f_w(x) = \arg\max_{\mathbf{y} \in \mathcal{Y}} w \cdot \Psi(\mathbf{x}, \mathbf{y})$$

The task of finding the $y^* \in \arg\max_{\mathbf{y} \in \mathcal{Y}} w \cdot \Psi(\mathbf{x}, \mathbf{y})$ is called 'Inference'.

An arbitrary loss function $\Delta(\mathbf{y}, \hat{\mathbf{y}}) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is usde to quantify the loss associated with the prediction $\hat{\mathbf{y}}$ and the true output vector $\mathbf{y}$. Δ is non-negative. $\Delta = 0$ when $\mathbf{y} = \hat{\mathbf{y}}$ and $\Delta > 0$ for other cases. Suppose a finite training set $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ is generated according to a fixed but unobserved distribution $P(\mathbf{x}, \mathbf{y})$, our function can be learnt through minimizing the loss on the training sample S

$$R_S^{\Delta}(f_w) = \frac{1}{n} \sum_{i=1}^{n} \Delta(\mathbf{y}_i, f_w(\mathbf{x}_i))$$

To include unobserved information in the model, Yu[3] extended the joint feature function $\Psi(\mathbf{x}, \mathbf{y})$ with a latent variable $\mathbf{h} \in \mathcal{H}$ to $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$. So the inference problem becomes

$$f_w(x) = \arg\max_{(\mathbf{y} \times \mathbf{h}) \in \mathcal{Y} \times \mathcal{H}} w \cdot \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$$

Accordingly, the loss function can be extended as

$$\Delta((\mathbf{y}_i, \mathbf{h}_i^*(w)), (\hat{\mathbf{y}}_i(w), \hat{\mathbf{h}}_i(w)))$$

where

$$\mathbf{h}_i^*(w) = \arg\max_{\mathbf{h} \in \mathcal{H}} w \cdot \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h})$$

$$(\hat{\mathbf{y}}_i(w), \hat{\mathbf{h}}_i(w)) = \arg\max_{(\mathbf{y} \times \mathbf{h}) \in \mathcal{Y} \times \mathcal{H}} w \cdot \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})$$

Yu[3] proved that the extended loss has an upper bound

$$\Delta((\mathbf{y}_i, \mathbf{h}_i^*(w)), (\hat{\mathbf{y}}_i(w), \hat{\mathbf{h}}_i(w))) \leq \left(\max_{(\hat{\mathbf{y}} \times \hat{\mathbf{h}}) \in \mathcal{Y} \times \mathcal{H}} [w \cdot \Psi(\mathbf{x}_i, \hat{\mathbf{y}}, \hat{\mathbf{h}}) + \Delta(\mathbf{y}_i, \hat{\mathbf{y}}, \hat{\mathbf{h}})] \right) - \max_{\mathbf{h} \in \mathcal{H}} w \cdot \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h})$$

Hence the optimization problem for Structural SVMs with latent variables becomes

$$\min_w \left[\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \left(\max_{(\hat{\mathbf{y}} \times \hat{\mathbf{h}}) \in \mathcal{Y} \times \mathcal{H}} [w \cdot \Psi(\mathbf{x}_i, \hat{\mathbf{y}}, \hat{\mathbf{h}}) + \Delta(\mathbf{y}_i, \hat{\mathbf{y}}, \hat{\mathbf{h}})] \right) - C \sum_{i=1}^{n} \left(\max_{\mathbf{h} \in \mathcal{H}} w \cdot \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \right) \right]$$

3 Milestones

Bibliography

[1] S. Gould, "Learning weighted lower linear envelope potentials in binary markov random fields,"

[2] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, pp. 1453–1484, 2005.

[3] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1169–1176, ACM, 2009.