

Convex Relaxations of Latent Variable Training

Yuhong Guo and Dale Schuurmans

読む人: 栗原 賢一 (東工大)

注意: NIPSでのタイトルはConvex Relaxations of EM となっている。現在、web で落とせる pdf のタイトルは Convex Relaxations of Latent Variable Training

概要

- 隠れ変数のあるモデルのパラメータ推定を relaxation + semi-definite programming (SDP) で解きます。
- けっこう広いクラスの問題に適用できます。
- ただし、(#学習データ by #学習データ) の行列の SDP を解くので、学習データは100個程度で、いっぱいいっぱい。

論文の味わいどころ

- convex relaxation はそんなに簡単なことではないと、まず釘をさす (by Lemma 1)
- 不幸な Lemma 1 を解決する方法を提案
- 後はひたすら式変形 (退屈)

背景 - 確率モデルとパラメータ推定

- 例: 正規分布 $N(x; m, \sigma)$
- データ: x_1, \dots, x_n
- 最尤推定では、 m と σ が一意に **closed form** で求まる。
- optimization 不要。

背景 - 確率モデルとパラメータ推定

- 例: 二つの混合正規分布
 - $N(x; \mu_1, \sigma_1)$ と $N(x; \mu_2, \sigma_2)$
 - $p(x) = \pi N(x; \mu_1, \sigma_1) + (1-\pi) N(x; \mu_2, \sigma_2)$
- データ: x_1, \dots, x_n , **隠れ変数: y_1, \dots, y_n**
 - 各データ x_i は、どちらかの正規分布から生成された
- 最尤推定では、 $\mu_1, \mu_2, \sigma_1, \sigma_2, \pi$ が closed form で求まらない。
- optimization 必要。
 - e.g. expectation maximization (EM) で対数尤度を最大化
 - **local optima**

目的

- 隠れ変数のあるモデルでパラメータ推定
- 目的関数
 - $\min_y \min_w - \sum_i \log P(x_i, y_i | w)$
 - 観測 $X=(x_1, \dots, x_n)$, 隠れ変数 $Y=(y_1, \dots, y_n)$
 - c.f. EM は $\min_w - \sum_i \log P(x_i | w)$
- convex relaxation したい

残念な補題 (Lemma 1)

- 準備
 - 例: 二つの混合正規分布
 - $p(x) = \pi N(x; m_1, \sigma_1) + (1-\pi) N(x; m_2, \sigma_2)$
 - 1 と 2 というラベルに意味はなく、交換しても分布は同じ
- Lemma 1
 - もし対数尤度が convex でラベルの交換に対して不変であれば、最適な $P(Y|X)$ は uniform

$M=YY^T$ で最適化

- $\min_y \min_w - \sum_i \log P(x_i, y_i | w)$
- の代わりに
- $\min_M \min_w - \sum_i \log P(x_i, y_i | w)$
- where $M=YY^T$

Convex Relaxation 準備

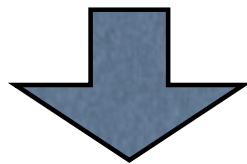
- 例題
- $p(Y_i | \Phi_i, W) = \exp(\Phi_i W Y_i^\top - A(W, \Phi_i))$
 - $\Phi_i = [0, \dots, 1, \dots, 0]$
 - $Y_i = [0, \dots, 1, \dots, 0]$
 - 正規化; $A(W, \Phi_{i:}) = \log \sum_a \exp(\Phi_{i:} W \mathbf{1}_a)$
- 目的関数

$$\min_W \left(\sum_i A(W, \Phi_{i:}) \right) - \text{tr}(\Phi W Y^\top) + \frac{\alpha}{2} \text{tr}(W^\top W)$$

Convex Relaxation 準備

$$\min_W \left(\sum_i A(W, \Phi_{i:}) \right) - \text{tr}(\Phi W Y^\top) + \frac{\alpha}{2} \text{tr}(W^\top W)$$

$$A(W, \Phi_{i:}) = \log \sum_a \exp(\Phi_{i:} W \mathbf{1}_a).$$



exact

$$\min_B \left(\sum_i A(B, \Phi_{i:}) \right) - \text{tr}(K B M) + \frac{1}{2\alpha} \text{tr}(B^\top K B M)$$

subject to $B \leq I, B\mathbf{1} = 0$

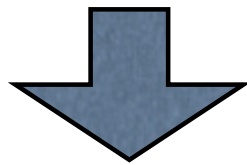
$$M = Y Y^\top \quad W = \frac{1}{\alpha} \Phi^\top B Y \quad K = \Phi \Phi^\top$$

Convex Relaxation 準備

$$\min_B \left(\sum_i A(B, \Phi_{i:}) - \text{tr}(KBM) + \frac{1}{2\alpha} \text{tr}(B^\top KBM) \right)$$

subject to $B \leq I, B\mathbf{1} = 0$

$$M = YY^\top \quad W = \frac{1}{\alpha} \Phi^\top BY \quad K = \Phi\Phi^\top$$



exact Aの消去

$$\max_{\Lambda \geq 0, \Lambda\mathbf{1}=\mathbf{1}} -\text{tr}(\Lambda \log \Lambda^\top) - \mathbf{1}^\top \Lambda \log(M\mathbf{1}) - \frac{1}{2\alpha} \text{tr}((I - \Lambda)^\top K(I - \Lambda)M)$$

$$B = I - \Lambda$$

Convex Relaxation

- ベイジアンネット

$$\min_{\{Y^h\}} \sum_j \min_{\mathbf{w}_j} -\log P(z_j^i | \mathbf{z}_{\pi(j)}^i, \mathbf{w}_j) + \frac{\alpha}{2} \mathbf{w}_j^\top \mathbf{w}_j$$

concave w.r.t. Λ
convex w.r.t. M

$$= \min_{\{M^h\}} \sum_j \max_{\Lambda_j \geq 0, \Lambda_j \mathbf{1} = \mathbf{1}} -\text{tr}(\Lambda_j \log \Lambda_j^\top) - \mathbf{1}^\top \Lambda_j \log(M^j \mathbf{1}) - \frac{1}{2\alpha} \text{tr}((I - \Lambda_j)^\top K^j (I - \Lambda_j) M^j)$$

subject to $M^h = Y^h Y^{h\top}, Y^h \in \{0, 1\}^{t \times v_h}, Y^h \mathbf{1} = \mathbf{1}$

\Updownarrow exact

$$M \in \{0, 1\}^{t \times t}, \text{diag}(M) = \mathbf{1}, M = M^\top, M \succeq 0, \text{rank}(M) = v$$

\Updownarrow relaxation

$$M^h \in [0, 1]^{t \times t}, \text{diag}(M^h) = \mathbf{1}, M^h = M^{h\top}, M^h \succeq 0$$

適用範囲

- 観測データは、連続でも離散でもよい
- 隠れ変数は離散
- ベイジアンネット
 - マルコフランダムフィールドは駄目っぽい

実験

- ベイジアンネットの実験
- 学習データのサイズは100

Bayesian networks	Fully Supervised		Viterbi EM		Convex EM	
	Train	Test	Train	Test	Train	Test
Synth1	7.23 \pm .06	7.90 \pm .04	11.29 \pm .44	11.73 \pm .38	8.96 \pm .24	9.16 \pm .21
Synth2	4.24 \pm .04	4.50 \pm .03	6.02 \pm .20	6.41 \pm .23	5.27 \pm .18	5.55 \pm .19
Synth3	4.93 \pm .02	5.32 \pm .05	7.81 \pm .35	8.18 \pm .33	6.23 \pm .18	6.41 \pm .14
Diabetes	5.23 \pm .04	5.53 \pm .04	6.70 \pm .27	7.07 \pm .23	6.51 \pm .35	6.50 \pm .28
Pima	5.07 \pm .03	5.32 \pm .03	6.74 \pm .34	6.93 \pm .21	5.81 \pm .07	6.03 \pm .09
Cancer	2.18 \pm .05	2.31 \pm .02	3.90 \pm .31	3.94 \pm .29	2.98 \pm .19	3.06 \pm .16
Alarm	10.23 \pm .16	12.30 \pm .06	11.94 \pm .32	13.75 \pm .17	11.74 \pm .25	13.62 \pm .20
Asian	2.17 \pm .05	2.33 \pm .02	2.21 \pm .05	2.36 \pm .03	2.70 \pm .14	2.78 \pm .12

average loss \pm standard deviation

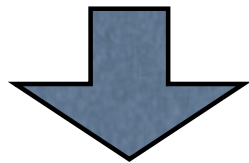
まとめ

- Y で relaxation するのはよくない(ラベルのパーミュテーション)
- $M = YY^T$ で relaxation する
- $\log p(Y|X)$ や $\log p(X,Y)$ を convex relaxation することができることを示した
- 実験により、Viterbi EM よりはよいことがわかった

補足; A の消去

$$\min_W \left(\sum_i A(W, \Phi_{i:}) \right) - \text{tr}(\Phi W Y^\top) + \frac{\alpha}{2} \text{tr}(W^\top W)$$

where $A(W, \Phi_{i:}) = \log \sum_a \exp(\Phi_{i:} W \mathbf{1}_a)$,



Fenchel conjugate of A

$$A(w, \Phi_{i:}) = \max_{\Theta_{i:}} \text{tr}(\Theta_{i:}^T \Phi_{i:} W) - A^*(\Theta_{i:})$$

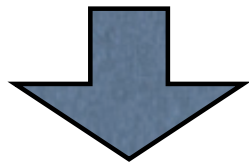
where

$$A^*(\Theta_{i:}) = \sup_{w, \Phi_{i:}} \text{tr}(\Theta_{i:}^T \Phi_{i:} W) - A(W, \Phi_{i:}) = \Theta_{i:} \log \Theta_{i:}^T$$

補足; A の消去

$$\min_W \left(\sum_i A(W, \Phi_{i:}) \right) - \text{tr}(\Phi W Y^\top) + \frac{\alpha}{2} \text{tr}(W^\top W)$$

$$\text{where } A(W, \Phi_{i:}) = \log \sum_a \exp(\Phi_{i:} W \mathbf{1}_a)$$



Fenchel conjugate of A で A を消去

$$\max_{\Theta} \min_W -\text{tr}(\Theta \log \Theta^\top) - \text{tr}((Y - \Theta)^\top \Phi W) + \frac{\alpha}{2} \text{tr}(W^\top W)$$

$$\text{subject to } \Theta \geq 0, \Theta \mathbf{1} = \mathbf{1}$$



Wを解く

$$\max_{\Theta} -\text{tr}(\Theta \log \Theta^\top) - \frac{1}{2\alpha} \text{tr}((Y - \Theta)^\top \Phi \Phi^\top (Y - \Theta))$$

$$\text{subject to } \Theta \geq 0, \Theta \mathbf{1} = \mathbf{1}$$