NIPS2006 **読み会**

Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation

G. Cawley, N. Talbot, M. Girolami
(University of Glasgow)

このスライドの全ての数式・表は紹介論文から引用しています。

論文紹介者：

坪井祐太

日本IBM東京基礎研究所 / NAIST博士課程

論文紹介理由に代えて NIPS 2006 Tutorial の紹介

D. Klein, “Machine Learning for Natural Language Processing: New Developments and Challenges”

テキスト処理（自然言語処理）の課題

(Recurring Issues in NLP Models, tutorial slides P.9 より)

- 学習時間がかかり過ぎることが原因で、識別手法(discriminative methods)が利用できないことがある。
- 百万以上の素性数を扱える必要性
  - データ数より素性数が多いことが殆どだが、それでも統計的手法は役立つ
- カーネル法は殆ど常に計算コストが大きく適用不可。 全ては主問題(Primal) で扱いたい。
- 学習データに現れない未知の現象・言葉を適切に扱える必要性
- 現実のシステムでは、非定常性が強い
- パイプラインシステムのため、（システムの信頼性を考慮して）調整された確率が必要。積み重なるエラーへの対応も必要。
  - （単語推定→品詞推定→構文構造推定→意味推定→…）

論文の見どころ

- 目的：マルチクラスの分類・分布推定
  - スパースなロジスティック回帰
- 手法：ハイパーパラメータ無しモデル
  - 解析的に正則化項の（ハイパー）パラメータを積分消去
  - クロスバリデーションでモデル選択する手法と同等の分類性能で、5-100倍速い！
- 貢献：手法は先行研究(Williams 1995)で完成
  - 実験で有効性を示した
  - Logistic Regressionの文脈で先行研究を紹介…

Multinomial Logistic Regression

$$P(t \mid \mathbf{x}) = \frac{\exp(\langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}, t) \rangle)}{\sum_{\tilde{t} \in \mathbf{Y}} \exp(\langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}, \tilde{t}) \rangle)}$$

- x:**入力素性ベクトル** (d**次元**)
- t:**出力ベクトル**(c**次元**) (**出力クラスの行に**1)
- w: **重みベクトル**(**モデルパラメータ** c×d**次元**)
- l**個の学習データ**D={(x^n,t^n)} (n=1…l)

条件分布

$$p(t_i^n | \boldsymbol{x}^n) = y_i^n = \frac{\exp\{a_i^n\}}{\sum_{j=1}^{c} \exp\{a_j^n\}}$$

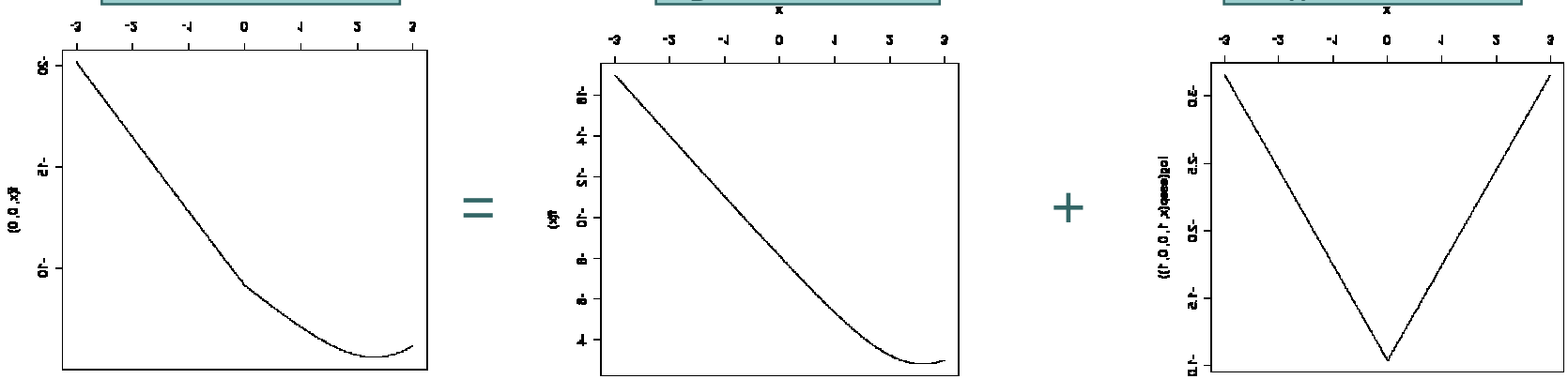
$$\text{where} \qquad a_i^n = \sum_{j=1}^{d} w_{ij} x_j^n$$

パラメータwの
事後確率最大化 (MAP)推定

- l個の学習データD={(x^n,t^n)} (n=1…l)
- αは正則化項のパラメータ (0≦α≦∞)

$$L = E_{\mathcal{D}} + \alpha E_{\mathcal{W}} \qquad E_W = \sum_{i=1}^{c}\sum_{j=1}^{d} |w_{ij}|$$

L=対数損失関数 = E_D=負対数尤度 + αE_W:L1正則化項



対数損失関数の偏微分

$$L = E_{\mathcal{D}} + \alpha E_{\mathcal{W}} \qquad E_W = \sum_{i=1}^{c}\sum_{j=1}^{d} |w_{ij}|$$

- 最小点では、負対数尤度の偏微分の絶対値がαより小さいとw_{ij}は0になる→スパース
- L1は正則化項の微分値が定数(sign(w_{ij})×α)で、$w_{ij} \neq 0$で常に一定のペナルティ。
- 一方、L2は0から遠くでは大きなペナルティだが、0の周辺ではペナルティがすくない。

$$\left|\frac{\partial E_{\mathcal{D}}}{\partial w_{ij}}\right| = \alpha \quad \text{if } |w_{ij}| > 0 \qquad \text{and} \qquad \left|\frac{\partial E_{\mathcal{D}}}{\partial w_{ij}}\right| < \alpha \quad \text{if } |w_{ij}| = 0.$$

Eliminating the Regularisation Parameters (1)

- **ベイズ的正則化項のパラメータαの決定**
  - evidence **が最大になるモデルを選択**
  - **αで積分消去**(integrated out) **←こっち**

事後分布∝尤度×事前分布:
p(w|D) ∝ p(D|w) p(w)

αで周辺化して事前確率p(w)**を求める**

$$p(\boldsymbol{w}) = \int p(\boldsymbol{w}|\alpha)p(\alpha)d\alpha.$$

Eliminating the Regularisation Parameters (2)

- p(w)はLaplace分布(log p(w)∝ α Σ |w|)
- p(α)は無情報事前分布
  - p(α)=1/αと置く (Jeffrey's Prior)

Wは0でないw$_{ij}$の数

$$P(\boldsymbol{w}) = \left(\frac{\alpha}{2}\right)^{W} \exp\{-\alpha E_{\mathcal{W}}\} = \frac{1}{2^{W}} \int_{0}^{\infty} \alpha^{W-1} \exp\{-\alpha E_{\mathcal{W}}\} d\alpha.$$

(Gamma integral, $\int_0^\infty x^{\nu-1} e^{-\mu x} dx = \frac{\Gamma(\nu)}{\mu^{\nu}}$)

$$p(\boldsymbol{w}) = \frac{1}{2^{W}} \frac{\Gamma(W)}{E_{\mathcal{W}}^{W}}$$

($E_{\mathcal{W}} = \sum_{i=1}^{c} \sum_{j=1}^{d} |w_{ij}|$)

Eliminating the Regularisation Parameters (3)

- **αを積分消去した負の**log p(w)

$$p(\boldsymbol{w}) = \frac{1}{2^W} \frac{\Gamma(W)}{E_{\mathcal{W}}^W} \quad \Longrightarrow \quad -\log p(\boldsymbol{w}) \propto W \log E_{\mathcal{W}},$$

$$\left(E_{\mathcal{W}} = \sum_{i=1}^{c} \sum_{j=1}^{d} |w_{ij}| \right)$$

- **αを積分消去した対数損失関数**
  (目的関数)

$$M = E_{\mathcal{D}} + W \log E_{\mathcal{W}}$$

他の手法との比較

- LASSOなど他のL1ロジスティック回帰手法は、Cross-validationでハイパーパラメータの決定が必要 (CVはExpensive)
- Relevance Vector Machine (RVM)はハイパーパラメータの決定不要だが計算量が$O(d^3)$：dは次元数

SBMLR(提案)とSMLR（クロスバリデーションでα探索)の比較実験結果

- ベンチマークデータセットの結果

Benchmark	Error Rate		Cross Entropy		Sparsity		$\log_{10}\frac{T_{\mathrm{SMLR}}}{T_{\mathrm{SBMLR}}}$
	SBMLR	SMLR	SBMLR	SMLR	SBMLR	SMLR	
Covtype	0.4051	**0.4041**	**0.9590**	0.9733	**0.4312**	0.3069	0.6965
Crabs	**0.0350**	0.0500	0.1075	**0.0891**	**0.2708**	0.0635	2.7949
Glass	0.3318	**0.3224**	**0.9398**	0.9912	0.4400	**0.4700**	1.9445
Iris	**0.0267**	**0.0267**	**0.0792**	0.0867	**0.4067**	**0.4067**	1.9802
Isolet	**0.0475**	0.0513	**0.1858**	0.2641	**0.9311**	0.8598	1.3110
Satimage	0.1610	**0.1600**	0.3717	**0.3708**	**0.3694**	0.2747	1.3083
Viruses	**0.0328**	**0.0328**	0.1670	**0.1168**	**0.8118**	0.7632	2.1118
Waveform	**0.1290**	0.1302	**0.3124**	0.3131	0.3712	**0.3939**	1.8133
Wine	**0.0225**	0.0281	0.0827	**0.0825**	**0.6071**	0.5524	2.5541

同じぐらいの性能

(=2で100倍早い)
SMLRより5-100倍早い

NIPS論文紹介:
Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (坪井)

まとめ

- スパースなロジスティック回帰の学習
- 注目：ハイパーパラメータ無しモデル
  - 解析的に正則化項の（ハイパー）パラメータを積分消去
  - クロスバリデーションでモデル選択する手法と同等の分類性能で、5-100倍速い！
- 手法は先行研究(Williams 1995)のまま
  - 自分達の提案法のように誤解しがちでずるい…
- ベンチマークデータで有効性を示した