High-dimensional classification by sparse logistic regression

Felix Abramovich

Tel Aviv University

(based on joint work with Vadim Grinshtein, The Open University of Israel and Tomer Levy, Tel Aviv University)

Outline

1. Review on (binary) classification
2. High-dimensional (binary) classification by sparse logistic regression
   - model, feature selection by penalized maximum likelihood
   - theory: misclassification excess bounds, adaptive minimax classifiers
   - computational issues: logistic Lasso and Slope
3. Multiclass extensions
   - sparse multinomial logistic regression
   - theory
   - multinomial logistic group Lasso and Slope

Binary Classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})),\ \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- Classifier $\eta : \mathbb{R}^d \to \{0, 1\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg\min_\eta R(\eta)$

$$\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}, \quad R(\eta^*) = E_{\mathbf{X}}\left(\min(p(\mathbf{X}), 1 - p(\mathbf{X}))\right)$$

Binary Classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})),\ \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$

- Classifier $\eta : \mathbb{R}^d \to \{0, 1\}$

- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$

- Bayes classifier $\eta^*(\mathbf{x}) = \arg\min_\eta R(\eta)$

$$\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}, \quad R(\eta^*) = E_{\mathbf{X}}\left(\min(p(\mathbf{X}), 1 - p(\mathbf{X}))\right)$$

- Data $D = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)) \sim \mathcal{F}$

  (conditional) Missclassification error $R(\hat{\eta}) = P(Y \neq \hat{\eta}(\mathbf{x})|D)$

  Misclassification excess risk $\mathcal{E}(\hat{\eta}, \eta^*) = ER(\hat{\eta}) - R(\eta^*)$

Vapnik-Chervonenkis (VC) dimension

Definition

Let $\mathcal{C}$ be a set of classifiers. $VC(\mathcal{C})$ is the maximal number of points in $\mathcal{X}$ that can be arbitrarily classified by classifiers in $\mathcal{C}$.

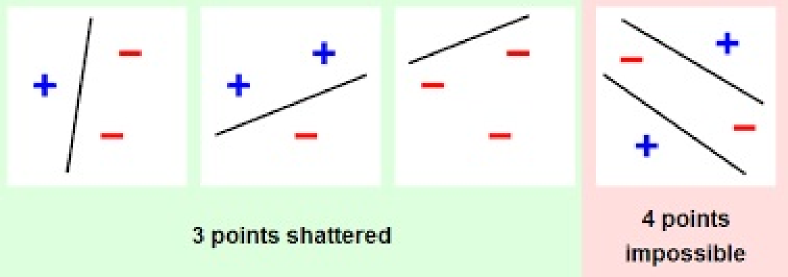
Vapnik-Chervonenkis (VC) dimension

Definition

Let $\mathcal{C}$ be a set of classifiers. $VC(\mathcal{C})$ is the maximal number of points in $\mathcal{X}$ that can be arbitrarily classified by classifiers in $\mathcal{C}$.

Example: VC of linear classifiers $\mathcal{C} = \{\eta(\mathbf{x}) = I\{\boldsymbol{\beta}^t \mathbf{x} \geq 0\},\ \boldsymbol{\beta} \in \mathbb{R}^d\}$

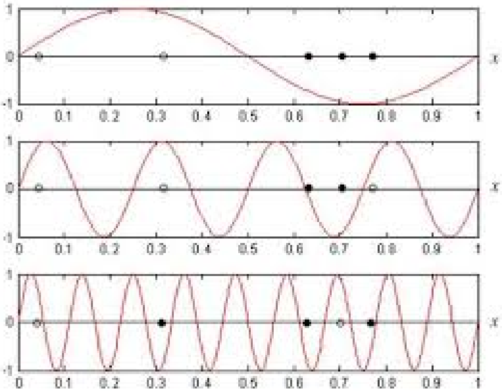
- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{C} = \{\eta(\mathbf{x}) = I\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0\}$ $VC(\mathcal{C}) = 3\ (= d)$



- $\mathcal{X} = \mathbb{R}^{d-1}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ $(x_0 = 1)$ $VC(\mathcal{C}) = d$

Example: VC of sine classifiers: $\mathcal{X} = \mathbb{R}$,
$\mathcal{C} = \{\eta(x) = I\{x \geq \sin(\theta x), \ \theta > 0\}$

Can classify any finite subset of points, $VC(\mathcal{C}) = \infty$



Minimax lower bound

Minimax lower bound. Let $2 \leq VC(\mathcal{C}) < \infty$, $n \geq VC(\mathcal{C})$ and $R(\eta^*) > 0$. Then,

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}, f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{VC(\mathcal{C})}{n}}$$

(e.g., Devroye, Györfi and Lugosi, '96).

In particular, for linear classifiers

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}, f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{d}{n}}$$

Two main approaches

1. Empirical Risk Minimization (ERM)

$$\hat{\eta} = \arg\min_{\eta \in \mathcal{C}} \hat{R}(\eta) = \arg\min_{\eta \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq \eta(\mathbf{x}_i))$$

- well-developed theory (Devroye, Györfi and Lugosi '96; Vapnik '00; see also Boucheron, Bousquet and Lugosi '05 for review)

$$\sup_{\eta^* \in \mathcal{C}} \mathcal{E}(\hat{\eta}, \eta^*) \leq C\sqrt{\frac{VC(\mathcal{C})}{n}} \quad \text{(optimal order)}$$

- computationally infeasible, various convex surrogates (e.g., SVM)

2. Plug-in Classifiers

- estimate $p(\mathbf{x})$ from the data
  (e.g, (parametric) logistic regression: $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$ or nonparametric: Yang '99, Koltchinskii and Beznosova '05, Audibert and Tsybakov '07)

- plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2)$

2. Plug-in Classifiers

- estimate $p(\mathbf{x})$ from the data
  (e.g, (parametric) logistic regression: $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$ or nonparametic: Yang '99, Koltchinskii and Beznosova '05, Audibert and Tsybakov '07)

- plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2)$

Logistic regression classifier

1. $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$
2. estimate $\boldsymbol{\beta}$ by MLE
3. plug-in $\hat{\eta} = I(\hat{p}(\mathbf{x}) \geq 1/2) = I(\hat{\boldsymbol{\beta}}^t \mathbf{x} \geq 0)$ – linear classifier

Big Data era – curse of dimensionality

For large d classification without feature (model) selection *is as bad as just pure random guessing* (e.g., Bickel and Levina '04; Fan and Fan '08)

Big Data era – curse of dimensionality

For large d classification without feature (model) selection *is as bad as just pure random guessing* (e.g., Bickel and Levina '04; Fan and Fan '08)

Sparse logistic regression classifier

1. model/feature selection – $\widehat{M}$
2. plug-in $\hat{\eta}_{\widehat{M}} = I(\hat{\boldsymbol{\beta}}^t_{\widehat{M}} \mathbf{x} \geq 0)$

Sparse logistic regression

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})),\ \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- $logit(p(\mathbf{x})) = \ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$
- sparsity assumption: $||\boldsymbol{\beta}||_0 \leq d_0$

Sparse logistic regression

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})),\ \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- $logit(p(\mathbf{x})) = \ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$
- sparsity assumption: $||\boldsymbol{\beta}||_0 \leq d_0$

Lemma (thanks to Noga Alon)

Let $\mathcal{C}(d_0) = \{\eta(\mathbf{x}) = I\{\boldsymbol{\beta}^t \mathbf{x} \geq 0\} : \boldsymbol{\beta} \in \mathbb{R}^d,\ ||\boldsymbol{\beta}||_0 \leq d_0\}$.

$$d_0 \log_2\left(\frac{2d}{d_0}\right) \leq VC(\mathcal{C}(d_0)) \leq 2d_0 \log_2\left(\frac{de}{d_0}\right),\ i.e.$$

$$VC(\mathcal{C}(d_0)) \sim d_0 \ln\left(\frac{de}{d_0}\right)$$

Model/feature selection by penalized MLE

- For a given model $M \subseteq \{1, \ldots, d\}$, MLE:

$$\widehat{\boldsymbol{\beta}}_M = \arg \max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \widetilde{\boldsymbol{\beta}}_M^t \mathbf{x}_i Y_i - \ln \left(1 + \exp(\widetilde{\boldsymbol{\beta}}_M)^t \mathbf{x}_i \right) \right\},$$

where $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_j = 0 \text{ iff } j \notin M\}$

Model/feature selection by penalized MLE

- For a given model $M \subseteq \{1,\ldots,d\}$, MLE:

$$\widehat{\beta}_M = \arg\max_{\widetilde{\beta}\in\mathcal{B}_M} \sum_{i=1}^n \left\{\widetilde{\beta}_M^t \mathbf{x}_i Y_i - \ln\left(1+\exp(\widetilde{\beta}_M)^t \mathbf{x}_i\right)\right\},$$

where $\mathcal{B}_M = \{\beta \in \mathbb{R}^d : \beta_j = 0 \text{ iff } j \notin M\}$

- $\widehat{M} = \arg\min_M \left\{\sum_{i=1}^n \left(\ln\left(1+\exp(\widehat{\beta}_M^t \mathbf{x}_i)\right) - \widehat{\beta}_M^t \mathbf{x}_i Y_i\right) + Pen(|M|)\right\}$

$$\widehat{p}_{\widehat{M}}(\mathbf{x}) = \frac{\exp(\widehat{\beta}_{\widehat{M}}^t \mathbf{x})}{1+\exp(\widehat{\beta}_{\widehat{M}}^t \mathbf{x})}$$

$$\widehat{\eta}_{\widehat{M}}(\mathbf{x}) = I(\widehat{p}_{\widehat{M}}(\mathbf{x}) \geq 1/2) = I(\widehat{\beta}_{\widehat{M}}^t \mathbf{x} \geq 0)$$

Complexity Penalties

- linear-type penalties $Pen(|M|) = \lambda|M|$

$\lambda = 1$	AIC (Akaike, '73)
$\lambda = \ln(n)/2$	BIC (Schwarz, '78)
$\lambda = \ln d$	RIC (Foster and George, '94)

Complexity Penalties

- linear-type penalties $Pen(|M|) = \lambda|M|$

  $\lambda = 1$ AIC (Akaike, '73)

  $\lambda = \ln(n)/2$ BIC (Schwarz, '78)

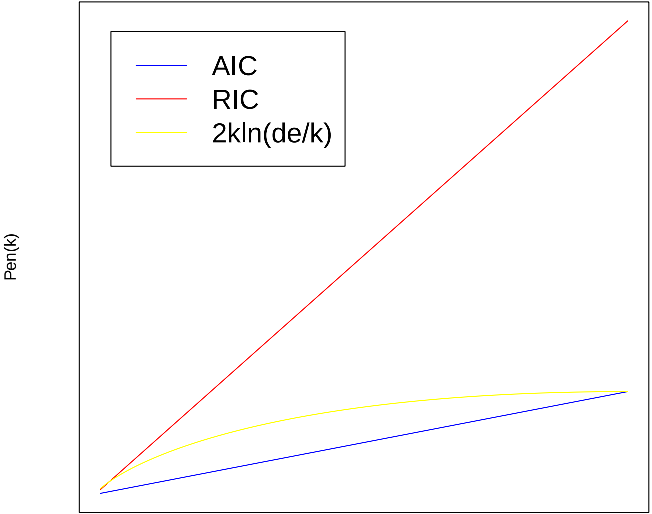
  $\lambda = \ln d$ RIC (Foster and George, '94)

- $k\ln(d/k)$-type nonlinear penalties $Pen(|M|) \sim C|M|\ln(de/|M|)$ (Birgé and Massart, '01, '07; Bunea *et al.* '07; AG '10 for Gaussian regression; AG '16 for GLM)

$$k\ln(d/k) \sim \ln\binom{d}{k} \; - \; \log(\text{number of models of size } k)$$

In addition, for classification, $k\ln(d/k) \sim VC(\mathcal{C}(k))$ (recall Lemma)

Various complexity penalties



Let $supp(f(\mathbf{x}))$ be bounded, w.l.o.g. $||\mathbf{x}||_2 \leq 1$ for all $x \in \mathcal{X}$

Assumption (boundedness)

There exists $0 < \delta < 1/2$ such that $\delta < p(\mathbf{x}) < 1 - \delta$ or, equivalently, there exists $C_0 > 0$ such that $|\boldsymbol{\beta}^t \mathbf{x}| < C_0$ for all $\mathbf{x} \in \mathcal{X}$.

The assumption prevents the variance $Var(Y) = p(\mathbf{x})(1 - p(\mathbf{x}))$ to be infinitely close to zero.

Excess risk bounds

Theorem (upper bound)

Under the boundedness assumption, for $Pen(|M|) = C|M| \ln\left(\frac{de}{|M|}\right)$,

$$\sup_{\eta \in \mathcal{C}(d_0)} \mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

The idea of the proof:

1. $\mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leq \sqrt{2\ EKL(p^*, \hat{p}_{\widehat{M}})}$ (Zhang '04; Bartlett *et al.* '06)
2. $\sup_{\beta \in \mathcal{B}(d_0)} EKL(p^*, \widehat{p}_{\widehat{M}}) = O\left(\frac{d_0 \ln \frac{de}{d_0}}{n}\right)$ (AG '16)

Excess risk bounds

Theorem (upper bound)

Under the boundedness assumption, for $Pen(|M|) = C|M| \ln\left(\frac{de}{|M|}\right)$,

$$\sup_{\eta \in \mathcal{C}(d_0)} \mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

The idea of the proof:

1. $\mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leq \sqrt{2\ EKL(p^*, \hat{p}_{\widehat{M}})}$ (Zhang '04; Bartlett *et al.* '06)
2. $\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(p^*, \widehat{p}_{\widehat{M}}) = O\left(\frac{d_0 \ln \frac{de}{d_0}}{n}\right)$ (AG '16)

Recall the lower bound for $2 \leq d_0 \ln\left(\frac{de}{d_0}\right) \leq n$:

$$\inf_{\tilde{\eta}} \sup_{\eta* \in \mathcal{C}(d_0), f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{VC(\mathcal{C}(d_0))}{n}} \geq C\sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

Tighter bounds under the additional low-noise condition

The main challenges are near the hyperplane $\boldsymbol{\beta}^t\mathbf{x} = 0$, where $p(\mathbf{x}) = 1/2$.

Assumption (low-noise condition)

$$P(|p(\mathbf{X}) - 1/2| \leq h) \leq Ch^{\alpha}, \quad \alpha \geq 0 \quad (\text{Tsybakov } '04)$$

- $\alpha = 0$ – no assumptions on the noise (as previously)
- $\alpha = \infty$ – there exists a "corridor" of width $2\ln\frac{1+2h}{1-2h}$ that separates the sets $\{\mathbf{x} : \boldsymbol{\beta}^t\mathbf{x} > 0\}$ and $\{\mathbf{x} : \boldsymbol{\beta}^t\mathbf{x} < 0\}$

Tighter bounds under the additional low-noise condition

The main challenges are near the hyperplane $\boldsymbol{\beta}^t\mathbf{x} = 0$, where $p(\mathbf{x}) = 1/2$.

Assumption (low-noise condition)

$$P(|p(\mathbf{X}) - 1/2| \leq h) \leq Ch^{\alpha}, \quad \alpha \geq 0 \quad \text{(Tsybakov '04)}$$

- $\alpha = 0$ – no assumptions on the noise (as previously)
- $\alpha = \infty$ – there exists a "corridor" of width $2\ln\frac{1+2h}{1-2h}$ that separates the sets $\{\mathbf{x} : \boldsymbol{\beta}^t\mathbf{x} > 0\}$ and $\{\mathbf{x} : \boldsymbol{\beta}^t\mathbf{x} < 0\}$

Under the low-noise assumption, for all $1 \leq d_0 \leq \min(d, n)$ and all $\alpha \geq 0$,

$$\sup_{\eta \in \mathcal{C}(d_0)} \mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) \leq \left(C(\delta)\, \frac{d_0 \ln \frac{de}{d_0}}{n} \right)^{\frac{\alpha+1}{\alpha+2}}$$

$\widehat{\eta}_{\widehat{M}}$ is rate-optimal and adaptive to both d_0 and α.

Computational aspects

$$\widehat{M} = \arg\min_{M} \{-\ell(M) + Pen(|M|)\}$$

combinatorial search over 2^d models (NP problem)

Computational aspects

$$\widehat{M} = \arg\min_M \{-\ell(M) + Pen(|M|)\}$$

combinatorial search over 2^d models (NP problem)

- Greedy algorithms (e.g., forward selection) – approximate the global solution by a stepwise sequence of local ones

  (require strong constraints on design)

Computational aspects

$$\widehat{M} = \arg\min_M \{-\ell(M) + Pen(|M|)\}$$

combinatorial search over 2^d models (NP problem)

- Greedy algorithms (e.g., forward selection) – approximate the global solution by a stepwise sequence of local ones

  (require strong constraints on design)

- Convex relaxation methods – replace the original combinatorial problem by some convex surrogate

Convex relaxation methods

Recall that $||\mathbf{x}||_2 \leq 1$.

- logistic Lasso (for linear penalties): $||\hat{\boldsymbol{\beta}}||_0 \to ||\hat{\boldsymbol{\beta}}||_1$
$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \left\{-\frac{1}{n}\ell(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1\right\}$$
  - fixed $\lambda \propto \sqrt{\frac{\ln d}{n}}$: rate-suboptimal (up to an extra log-factor: $O(\sqrt{\frac{d_0 \ln d}{n}})$) (van de Geer '08, Bellec *et al.* '16)
  - adaptively chosen λ : rate-optimal ($O(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}$) (Bellec *et al.* '16 for Gaussian regression; conjecture for classification)

Convex relaxation methods

Recall that $||\mathbf{x}||_2 \leq 1$.

- logistic Lasso (for linear penalties): $||\hat{\boldsymbol\beta}||_0 \to ||\hat{\boldsymbol\beta}||_1$

$$\hat{\boldsymbol\beta}_{Lasso} = \arg\min_{\boldsymbol\beta} \left\{-\frac{1}{n}\ell(\boldsymbol\beta) + \lambda||\boldsymbol\beta||_1\right\}$$

  - fixed $\lambda \propto \sqrt{\frac{\ln d}{n}}$: rate-suboptimal (up to an extra log-factor: $O(\sqrt{\frac{d_0 \ln d}{n}})$) (van de Geer '08, Bellec *et al.* '16)
  - adaptively chosen λ : rate-optimal ($O(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}$) (Bellec *et al.* '16 for Gaussian regression; conjecture for classification)

- logistic Slope: $k \ln(2d/k) \sim \sum_{j=1}^k \ln(2d/j)$

$$\hat{\boldsymbol\beta}_{Slope} = \arg\min_{\boldsymbol\beta} \left\{-\frac{1}{n}\ell(\boldsymbol\beta) + \sum_{j=1}^d \lambda_j |\beta|_{(j)}\right\}, \quad \lambda_1 \geq \ldots \geq \lambda_d > 0$$

$\lambda_j \propto \sqrt{\frac{\ln(2d/j)}{n}}$: rate-optimal ($O(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}$) (AG '19)

Multiclass classification

- appears in a variety applications, a lot of methods
- much less theory behind

Multiclass classification

- appears in a variety applications, a lot of methods
- much less theory behind

Main approaches :

1. reduction to a series of binary classifications
    - One-vs-All – each class is compared against all others
    - One-vs-One – all pairs of classes are compared to each other
2. extensions of binary classification approaches

Multiclass classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim Mult(p_1(\mathbf{x}), \ldots, p_L(\mathbf{x})),\ \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- Classifier $\eta : \mathbb{R}^d \to \{1, \ldots, L\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg\max_{1 \leq j \leq L}\ p_j(\mathbf{x})$,
  $R(\eta^*) = 1 - E_{\mathbf{X}}\left(\max_{1 \leq j \leq L}\ p_j(\mathbf{X})\right)$

  Misclassification excess risk $\mathcal{E}(\hat{\eta}, \eta^*) = ER(\hat{\eta}) - R(\eta^*)$

Multinomial logistic regression

$$Y \sim Mult(p_1(\mathbf{x}),\dots,p_L(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^d,\ \sum_{j=1}^{L} p_j(\mathbf{x}) = 1$$

$$\theta_j = \ln\frac{p_j(\mathbf{x})}{p_L(\mathbf{x})} = \boldsymbol{\beta}_j^t\mathbf{x}, \quad p_j(\mathbf{x}) = \frac{\exp\left(\boldsymbol{\beta}_j^t\mathbf{x}\right)}{\sum_{k=1}^{L}\exp\left(\boldsymbol{\beta}_k^t\mathbf{x}\right)}, \quad j = 1,\dots,L;\ \boldsymbol{\beta}_L = \mathbf{0}$$

(the choice of the reference class is arbitrary)

To each Y assign the corresponding indicator vector $\boldsymbol{\xi} \in \{0,1\}^L$

MLE: $B \in \mathbb{R}^{d\times L}$ – matrix of regression coefficients ($B_{\cdot L} = \mathbf{0}$)

$$\ell(B) = \sum_{i=1}^{n}\left\{\mathbf{x}_i^t B\boldsymbol{\xi}_i - \ln\sum_{l=1}^{L}\exp(\boldsymbol{\beta}_l^t\mathbf{x}_i)\right\} \to \max_B$$

Sparse multinomial logistic regression

- for multiclass setup there are various ways to define sparsity
- global sparsity: part of features do not have any impact on classification at all, i.e. $B_{j\cdot} = \mathbf{0}$
- for a given model $M \subseteq \{1, \ldots, d\}$
  - $|M| = \#\{\mathrm{non-zero\ rows\ of\ } B\} = r_B$
  - $$\widehat{B}_M = \arg\max_{\widetilde{B} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \mathbf{X}_i^t \widetilde{B} \boldsymbol{\xi}_i - \ln \sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{X}_i) \right\},$$

    where $\mathcal{B}_M = \{B \in \mathbb{R}^{d \times L} : B_{\cdot L} = \mathbf{0},\ \mathrm{and}\ B_{j\cdot} = \mathbf{0}\ \mathrm{iff}\ j \notin M\}$
- $\widehat{M} = \arg\min_M \{-\ell(\widehat{B}_M) + Pen(|M|)\}$
- $\widehat{\eta}_{\widehat{M}} = \arg\max_{1 \le l \le L} \widehat{\boldsymbol{\beta}}_{\widehat{M}l}^t \mathbf{x}$

$$\mathcal{C}_L(d_0) = \{\eta(\mathbf{x}) = \arg\max_{1 \le l \le L} \beta_l^t \mathbf{x} : B \in \mathbb{R}^{d \times L},\ B_{\cdot L} = \mathbf{0} \text{ and } r_B \le d_0\}$$

Assumption (boundedness)

There exists $0 < \delta < 1/2$ such that $\delta \le p_l(\mathbf{x}) \le 1 - \delta$ or, equivalently, $|\beta_l^t \mathbf{x}| < C_0$ with $C_0 = \ln \frac{1-\delta}{\delta}$ for all $l = 1, \ldots, L$ and $\mathbf{x} \in \mathcal{X}$.

Consider the complexity penalty

$$Pen(|M|) = C_1 \underbrace{|M|(L-1)}_{\#\text{ parameters,AIC}} + C_2 \underbrace{|M| \ln\left(\frac{de}{|M|}\right)}_{\log(\#\text{ models of size }|\mathrm{M}|)}$$

Theorem (upper bound)

Assume d_0-sparse multinomial logistic regression model. Under the boundedness assumption,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) \le C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

Excess risk bounds

Theorem (lower bound)

Let $2 \le d_0 \ln\left(\frac{de}{d_0}\right) \le n,\; d_0(L-1) \le n$ *and* $R(\eta^*) > 0$. *Then,*

$$\inf_{\widetilde{\eta}} \sup_{\eta^* \in \mathcal{C}_L(d_0), f(\mathbf{x})} \mathcal{E}(\widetilde{\eta}, \eta^*) \ge C\sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

The idea of the proof:

1. the error cannot be smaller than that for binary classification :
$$Error \ge C\sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n}} \quad \text{(see above)}$$
2. for a given true (oracle) model with $|M_0| = d_0$:

$Error \ge C\sqrt{\frac{d_0(L-1)}{n}}$ – via multiclass extension of VC (Daniely *et al.*, '12, '15)

Two regimes

1. Small number of classes: $L \leq 2 + \ln \frac{d}{d_0}$
   - $Pen(|M|) \sim |M| \ln \frac{de}{|M|}$
   - the error is of the order $\sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$ (does not depend on L, binary case)

2. Large number of classes: $2 + \ln \frac{d}{d_0} < L < \frac{n}{d_0}$
   - $Pen(|M|) \sim |M|(L-1)$ (AIC)
   - the error is of the order $\sqrt{\frac{d_0(L-1)}{n}}$ (regardless of d)

3. $L > \frac{n}{d_0}$ – consistent classification is impossible

As before, the rates can be improved under the additional low-noise condition $P\left(p_{(1)}(\mathbf{X}) - p_{(2)}(\mathbf{X}) \leq h\right) \leq Ch^{\alpha}$

Multinomial logistic group Lasso

B has a a row-wise sparsity. Let $|B|_j = ||B_{j\cdot}||_2$, $\quad ||\mathbf{x}||_2 \leq 1$

$$\widehat{B}_{gL} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{x}_i) \right) - \mathbf{x}_i^t \widetilde{B} \xi_i \right) + \lambda \sum_{j=1}^{d} |\widetilde{B}|_j \right\}$$

with $\lambda \sim \sqrt{\frac{L + \ln d}{n}}$

Under the boundedness assumption,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gL}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln d}{n}}$$

(sub-optimal)

Multinomial logistic group Slope

$$\widehat{B}_{gS} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{x}_i) \right) - \mathbf{x}_i^t \widetilde{B} \xi_i \right) + \sum_{j=1}^{d} \lambda_j |\widetilde{B}|_{(j)} \right\}$$

with $\lambda_j \sim \sqrt{\frac{L+\ln(d/j)}{n}}$

Under the boundedness assumption,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gS}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln \frac{de}{d_0}}{n}}$$

(optimal)

Future work/extensions

- different types of sparsity (e.g., double sparsity: nonzero rows are also sparse – multinomial logistic sparse group Slope

$$\widehat{B}_{sgS} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{x}_i) \right) - \mathbf{x}_i^t \widetilde{B} \xi_i \right) + \sum_{j=1}^{d} \lambda_j |\widetilde{B}|_{(j)} + \sum_{j=1}^{d} \sum_{l=1}^{L} \alpha_l |\widetilde{B}_{j(l)}| \right\}$$

Future work/extensions

- different types of sparsity (e.g., double sparsity: nonzero rows are also sparse – multinomial logistic sparse group Slope

$$\widehat{B}_{sgS} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{x}_i) \right) - \mathbf{x}_i^t \widetilde{B} \xi_i \right) + \sum_{j=1}^{d} \lambda_j |\widetilde{B}|_{(j)} + \sum_{j=1}^{d} \sum_{l=1}^{L} \alpha_l |\widetilde{B}_{j(l)}| \right\}$$

- different types of design (e.g., Gaussian, sub-Gaussian)

Future work/extensions

- different types of sparsity (e.g., double sparsity: nonzero rows are also sparse – multinomial logistic sparse group Slope

$$\widehat{B}_{sgS} = \arg\min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln \left(\sum_{l=1}^{L} \exp(\widetilde{\beta}_l^t \mathbf{x}_i) \right) - \mathbf{x}_i^t \widetilde{B} \xi_i \right) + \sum_{j=1}^{d} \lambda_j |\widetilde{B}|_{(j)} + \sum_{j=1}^{d} \sum_{l=1}^{L} \alpha_l |\widetilde{B}_{j(l)}| \right\}$$

- different types of design (e.g., Gaussian, sub-Gaussian)

- cost-sensitive classification

Thank You!