SML Week 4-6: Linear Methods for Classification

C. Andy Tsao

Dept. of Applied Math
National Dong Hwa University

September 30, 2013

Outline

Reference: §3.4, Chapter 4 of HTF

Why is LSE not satisfactory?

- Prediction can be improved by shrinking or zeroing some coeficients.
- Large number of predictors is not helpful in interpretation nor computation.

Approaches to var. selection and coef. shrinkage

- Subset selection: Best, forward (stepwise), backward (stepwise)
- Shrinkage methods
  1. Bridge regression includes *Lasso* [q=1] and *ridge regression* [q=2] in which $\hat{\beta}$ minimizes $L(\beta) = ||Y - \beta_0 - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$ where $\lambda, q > 0$.
  2. For $q = 2$, $\hat{\beta}^{ridge}$ minimizes equivalently $||Y - \beta_0 - X\beta||^2$ subject to $\sum_{j=1}^{p} |\beta_j|^2 \leq s$. One-to-one correspondence between λ and s.
  3. Standardization: $x'_{ij} = (x_{ij} - x_{.j})/s_j$ where $x_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ and $s_j = \frac{\sum_{i=1}^{n} (x_{ij} - x_{.j})^2}{n-1}$. And $\hat{\beta}_0 = \bar{y}$.

Other Methods

- Methods using derived inputs: *Principal Component Regression* (PCA), *Partial Least Squares* (PLS)
- Bayesian shrinkage/variable selection

Discrete Y classification

- Closer look at the task of classification: Find $F : \mathcal{X} \to \mathcal{Y}$
- Linear Methods: Linear Decision Boundaries.
- There is nothing *completely* new under the sun.
- Naive approach: regression. What's wrong and what can be right?
- Hypothesis Testing

Discrete Y classification

- Closer look at the task of classification: Find $F : \mathcal{X} \to \mathcal{Y}$
- There is nothing *completely* new under the sun.
- Regression
- Hypothesis Testing

Regression for classification

- Naive approach: regression. $Y = X\beta + \epsilon$.
- What's wrong and what can be right?

Regression for classification

- Naive approach: regression. $Y = X\beta + \epsilon$.
- What's wrong and what can be right?
- Point Estimation $\hat{Y} = X(X'X)^{-1}X'Y$.
- Inference on Y: Confidence Interval for Y?

Regression for classification

- Naive approach: regression. $Y = X\beta + \epsilon$.
- What's wrong and what can be right?
- Point Estimation $\hat{Y} = X(X'X)^{-1}X'Y$.
- Inference on Y: Confidence Interval for Y?
- Gauss-Markov Condition
- Normality assumption

Regression for classification

- Naive approach: regression. $Y = X\beta + \epsilon$.
- What's wrong and what can be right?
- Point Estimation $\hat{Y} = X(X'X)^{-1}X'Y$.
- Inference on Y: Confidence Interval for Y?
- Gauss-Markov Condition
- Normality assumption
- Is it a linear method?

Testing Hypotheses as a classification problem

- Recall hypotheses testing problem
  Formulation, Criteria/gurantee,
  (recommended) procedure, Supplemental measures
- Two classes $\approx$ two subspaces
- Matching criteria: TE/GE vs power function
- Accuracy estimation approach. (cf. Hwang, Casella, Robert, Wells and Farrel (1992, Annals of Stat))
- Is it a linear method?

Logistic Regression

For classifying K classes Estimation of $P[G = k|X = x]$

- For two classes $G = \{1, 2\}$.
- $P(G = 1|X = x) = \frac{exp(\beta_0 + x'\beta)}{1 + exp(\beta_0 + x'\beta)}$,
- $P(G = 2|X = x) = \frac{1}{1 + exp(\beta_0 + x'\beta)}$,
- $g(E(Y|x)) = \beta_0 + x'\beta$ where $g(p) = log[p/(1 - p)]$, the *logit* transformation

Logistic Regression

For classifying K classes Estimation of $P[G = k|X = x]$

- For two classes $G = \{1, 2\}$.
- $P(G = 1|X = x) = \frac{exp(\beta_0+x'\beta)}{1+exp(\beta_0+x'\beta)}$,
- $P(G = 2|X = x) = \frac{1}{1+exp(\beta_0+x'\beta)}$,
- $g(E(Y|x)) = \beta_0 + x'\beta$ where $g(p) = log[p/(1-p)]$, the *logit* transformation
- For usual GLM, Y is normally distributed and g is identity transformation.
- Strength and weakness

LR of an Indicator Matrix

For classifying K classes, coding response via an indicator variables.

- If $\#\mathcal{G} = K$, define $Y_k = 1_{[G=k]}$, $Y = (Y_1, \cdots, Y_K)$ and $\mathbf{Y}$ is a NxK indicator response matrix.
- $\hat{\mathbf{Y}} = X(X'X)^{-1}X'\mathbf{Y}$
- $\hat{\mathbf{B}} = (X'X)^{-1}X'\mathbf{Y}$
- A new observation with input x is classified
  - $\hat{f}(x) = [(1, x)\hat{\mathbf{B}}]'$
  - $\hat{G}(x) = argmax_{k\in\mathcal{G}}\hat{f}_k(x)$.
- Strength and weakness

Blackboard & Chalk: Special Case $K = 2$

$P(G = k|X = x)$ $(E(G|x))$

- Let π_k be prior probability of class k, $\sum_k \pi_k = 1$
- $P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$
- Various techniques are based on models of f's:
  - Gaussian: linear and quardratic discriminant analysis
  - mixture of Gaussians, nonparametirc density
  - *Naive Bayes*: a variant of the previous models assuming the conditional independence among explanatory variables

Gaussian class densities: LDA

- $f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} exp(\frac{-1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k))$.
- Linear Discriminant Analysis (LDA): $\Sigma_k = \Sigma$ for all k
- $$\log \frac{P(Y=k|X=x)}{P(Y=l|X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$
  $$= \log \pi_k \pi_l - \frac{1}{2}(\mu_k+\mu_l)'\Sigma^{-1}(\mu_k-\mu_l) + x'\Sigma^{-1}(\mu_k+\mu_l) \tag{1}$$
- $\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + log\pi_k$
- $G(x) = argmax_k \delta_k(x)$

$\hat{\pi}_k = N_k/N$ where N_k is the number of class-k obs.
$\hat{\mu_k} = \sum_{g_i=k} x_i/N_k$
$\hat{\Sigma} = \sum_k \sum_{g_i=k}(x_i-\hat{\mu_k})(x_i-\hat{\mu_k})'/(N-K)$.

LDA vs. LSE

- For $K = 2$, the LDA classification rule is

$$x'\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2'\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1'\hat{\Sigma}^{-1}\hat{\mu}_1 + log(N_1/N) - log(N_2/N)$$

Quardratic distriminant Analysis

QDA

- Non-equal Σ_k
- $\delta_k(x) = \frac{1}{2} log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + log\pi_k$

Interpretation of FDA

For easy exposition, let $K = 2$

- For LDA, assuming $\pi_2 = \pi_1$, decision rule $G(x) = arg\max_k f_k(x)$
- Note:$f_k(x) \propto (x - \mu_k)'\Sigma^{-1}(x - \mu_k)$
- Connection with logistic regression

Model and Estimation

- $log\left(\frac{P(G=1|X=x)}{P(G=K|X=x)}\right) = \beta_{1,0} + \beta_1' x$
- $log\left(\frac{P(G=k|X=x)}{P(G=K|X=x)}\right) = \beta_{k,0} + \beta_k' x, g = 1, \cdots, K-1.$

That is

- $P(G=k|X=x) = \frac{\exp(\beta_{k,0}+\beta_k' x)}{1+\sum_{k=1}^{K-1}\exp(\beta_{k,0}+\beta_k' x)}$
- $P(G=K|X=x) = \frac{1}{1+\sum_{k=1}^{K-1}\exp(\beta_{k,0}+\beta_k' x)}$
- Newton-Raphson Method (R)
- Large $\#x$ demands heavy computation and might not converge at all.

Logistic Regression or LDA?

- For logistic regression

$$\log \frac{P(Y=k|X=x)}{P(Y=l|X=x)} = \beta_{k0} + \beta_k' x.$$

- Recall (1) in LDA

$$\begin{aligned}
&\log \frac{P(Y=k|X=x)}{P(Y=l|X=x)} \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)'\Sigma^{-1}(\mu_k - \mu_l) + x'\Sigma^{-1}(\mu_k + \mu_l) \\
&= \alpha_{k0} + \alpha_k' x.
\end{aligned}$$

Compare and contrast

- Same log conditional probability ratio $log\left(\frac{P(G=k|X=x)}{P(G=K|X=x)}\right)$
- Different ways in estimating the unknown parameters: Logistic: Max conditional likelihood vs. LDA: full likelihood $P(X, Y = k) = f_k(x)\pi_k$

Variables: Continuous or Categorical

- R summary of data set (data.frame)
- Continuous or categorical variable, a practical understanding
- Continuous: detailed information, effective yet sensitive modelling
- Categorical: rough information, complex (when the level number is large) yet robust modelling

Hyperplane

A *hyperplane* or *affine set* L is defined by the equation $f(\mathbf{x}) = \beta_0 + \beta'\mathbf{x}$. For $\mathcal{R}^2$ this is a line. Perceptrons (Rosenblatt 1958)

- For any $\mathbf{x}_1, \mathbf{x}_2 \in L$, $\beta'(\mathbf{x}_1 - \mathbf{x}_2) = 0$.
- $\beta^* = \beta/||\beta||$ is the (unit) normal vector of L
- For any $\mathbf{x}_0$ in L, $\beta'\mathbf{x}_0 = -\beta_0$
- The signed distance of any $\mathbf{x}$ to L is $\beta^*(\mathbf{x} - \mathbf{x}_0) = \frac{1}{||\beta||}(\beta'\mathbf{x} + \beta_0) = \frac{f(\mathbf{x})}{||f'(\mathbf{x})||}$
- Figure 4.14.
- $f(\mathbf{x})$ is proportional to the signed distance from $\mathbf{x}$ to the hyperplane: $f(\mathbf{x}) = 0$.

Bisection Method

Given $f(x)$ is continuous on $[a_0, b_0]$ and $f(a_0)f(b_0) < 0$.

- For $n = 0, 1, 2, \ldots$ until satisfied, do
- Set $m = (a_n + b_n)/2$
- If $f(a_n)f(m) < 0$, set $a_{n+1} = a_n, b_{n+1} = m$
- Otherwise set $a_{n+1} = m, b_{n+1} = b_n$
- Then $f(x)$ has a root in $[a_n + 1, b_{n+1}]$.

Conte and de Boor (1980). Elementary numerical analysis: an algorithmic approach. 3rd Edition.

Perceptron Learning Algorithm–1

Goal: Minimize the distance of misclassified points to the decision boundary. Consider the binary problem when $Y = \pm 1$.

- Min $D(\beta, \beta_0) = -\sum_{i \in M} y_i(\beta_0 + \mathbf{x}_i'\beta)$ (4.37)
- $D \geq 0$ and D is proportional to the distances of the misclassified points to the decision boundary defined by $\beta_0 + \mathbf{x}_i'\beta = 0$. Assuming M is fixed, the gradient is
- $\frac{\partial D(\beta,\beta_0)}{\partial \beta} = -\sum_{i \in M} y_i \mathbf{x}_i$
- $\frac{\partial D(\beta,\beta_0)}{\partial \beta_0} = -\sum_{i \in M} y_i$

Perceptron Learning Algorithm–2

- A step is taken after each obs rather than the sum. This is a *stochastic gradient decent* minimizes a piecewise linear criterion.
- $\beta = \beta + \rho y_i \mathbf{x}_i$
- $\beta_0 = \beta_0 + \rho y_i$ with ρ is the learning rate. WLOG $\rho = 1$.
- When the cases are linearly seperable, this algorithm will converge to a seperating hyperplane in finite number of steps.
- Remark: Contrast to population version

Perceptron Learning Algorithm–3

Problems with the algorithm, Ripley (1996)

- When data are separable, the solutions are non-unique and depend on starting points.
- "Finite" can be very large.
- When the cases are *Not* linearly separable, this algorithm can develop long hard-to-detect cycles.

Optimal Separating Hyperplane

The *Optimal Separating Hyperplane* separates the two classes and maximizes the distance to the closest point from either class, Vapnik (1996).

- $\max_{\beta_0,\beta,||\beta||=1} C$
- subject to $y_i(\beta_0 + x_i'\beta) \geq C, i = 1, \cdots, N$. Equivalently
- subject to $\frac{1}{||\beta||} y_i(\beta_0 + x_i'\beta) \geq C$ (redefining β_0)
- subject to $\leftrightarrow y_i(\beta_0 + x_i'\beta) \geq C||\beta||$

WLOG, rescale and set $||\beta|| = 1/C$, the optimization problem becomes

- $\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2$
- subject to $y_i(\beta_0 + x_i'\beta) \geq 1$

Further transform the optimization problem $\rightarrow$ SVM.