SML Week 4-5: Linear Methods for Classi cation

C. Andy Tsao

Dept. of Applied Math
National Dong Hwa University

October 12, 2010

Why is LSE not satisfactory?

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

Other Methods

Regression for classi cation

- 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: Con dence Interval for Y?

Regression for classi cation

- 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: Con dence Interval for Y?
- Gauss-Markov Condition
- Normality assumption
- Is it a linear method?

Testing Hypotheses as a classi cation problem

- Recall hypotheses testing problem
  Formulation, Criteria/gurantee,
  (recommended) procedure, Supplemental measures
- Two classes 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

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

- Let π_k be prior probability of class k, $\sum_k \pi_k = 1$
- $P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$
-

Gaussian class densities: LDA

- $f_k(x) = \frac{1}{(2\pi)^{p/2} j \quad _k j^{1/2}} exp(\frac{1}{2}(x \quad \mu_k)^{\prime} \quad _k{}^{1}(x \quad \mu_k))$.
- Linear Discriminant Analysis (LDA): $\quad _k = \quad$ for all k
-

$$\log \frac{P(Y = kjX = x)}{P(Y = ljX = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \pi_k \pi_l \quad \frac{1}{2}(\mu_k + \mu_l)^{\prime} \quad {}^{1}(\mu_k \quad \mu_l) + x^{\prime} \quad {}^{1}(\mu_k + \mu_l) \tag{1}$$

- $\delta_k(x) = x^{\prime} \quad {}^{1}\mu_k \quad \frac{1}{2}\mu_k^{\prime} \quad {}^{1}\mu_k + log\, \pi_k$
- $G(x) = argmax_k \delta_k(x)$

$\hat{\pi}_k = N_k/N$ where N_k is the number of class-k obs.

$\hat{\mu}_k = \sum_{g_i = k} x_i / N_k$

$\hat{} = \sum_k \sum_{g_i = k} (x_i \quad \hat{\mu}_k)(x_i \quad \hat{\mu}_k)^{\prime}/(N \quad K)$.

LDA vs. LSE

- For $K = 2$, the LDA classi cation rule is

$$x^{0}\hat{}^{1}(\hat{\mu}_2 \quad \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2{}^{0}\hat{}^{1}\hat{\mu}_2 \quad \frac{1}{2}\hat{\mu}_1{}^{0}\hat{}^{1}\hat{\mu}_1$$
$$+ \log(N_1/N) \quad \log(N_2/N)$$

Quardratic distriminant Analysis

QDA

- Non-equal $_k$
- $\delta_k(x) = \frac{1}{2} log j \ _k j \quad \frac{1}{2}(x \quad \mu_k)^{\emptyset} \ _k{}^1(x \quad \mu_k) + log \pi_k$

Logistic Regression or LDA?

- For logistic regression

$$\log \frac{P(Y = kjX = x)}{P(Y = IjX = x)} = \beta_{k0} + \beta_k^{\emptyset} x.$$

- Recall (1) in LDA

$$= \alpha^{k0} + \log \alpha_k^{\emptyset} x \frac{P(Y = kjX = x)}{P(Y = IjX = x)} \quad {}^{I}) + x^{\emptyset} \quad {}^{1}(\mu^{k} + \mu^{}$$

$$= \log \frac{\pi_k}{\pi_I} \quad \frac{1}{2}(\mu_k + \mu^{I})^{\emptyset} \quad {}^{1}(\mu^{k}$$

Compare and contrast

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

Hyperplane

A *hyperplane* or *a ne set L* is de ned by the equation
$f(\mathbf{x}) =$

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 nite number of steps.
- Remark: Contrast to population version