Introduction

- Linearity: The decision boundary is linear

- Decision Boundaries:

  - Indicator Matrix: $\{x : (\hat{\beta}_{k0} - \hat{\beta}_{l0}) + (\hat{\beta}_k - \hat{\beta}_l)^T x = 0\}$

  - LDA: $\{x : \delta_k(x) = \delta_l(x)\}$

  - Logistic Regression: $\{x \mid \beta_0 + \beta^T x = 0\}$

$$Pr(G=1|X=x) = \frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)}$$

$$Pr(G=2|X=x) = \frac{1}{1+\exp(\beta_0+\beta^T x)}$$

$$\log \frac{Pr(G=1|X=x)}{Pr(G=2|X=x)} = \beta_0+\beta^T x$$

Linear Regression of an Indicator Matrix

- Model : $Y = (Y_1, \cdots, Y_K)^T$, $Y_K = \begin{cases} 1 & G = k \\ 0 & G \neq K \end{cases}$
- $\hat{Y} = x^T (X^T X)^{-1} X^T Y$

  $\hat{B} = (X^T X)^{-1} X^T Y$

- $\hat{G}(x) = \text{argmax}_{k \in G} (1 \quad x^T) \hat{B}$
- $E(Y_K \mid X = x) = Pr(G(x) = k \mid X = x)$

Linear Discriminant Analysis

- π_k — prior probability of class k, $\sum_{k=1}^{K} \pi_k = 1$

$$Pr(G=k|X=x) = \frac{f_x(x)\pi_k}{\sum_{k=1}^{K} f_x(x)\pi_k}$$

- $$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

$$\log \frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l)$$

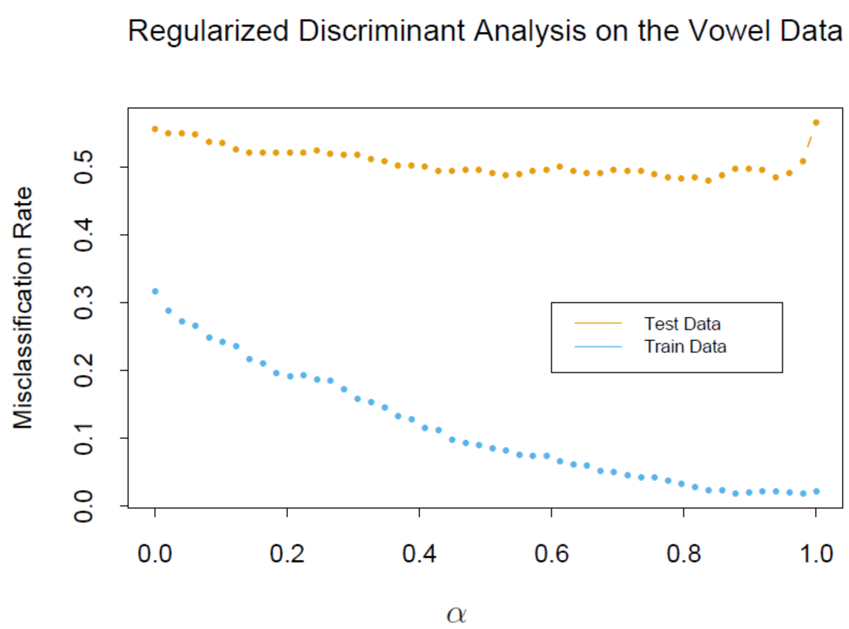
- $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$ (LDA)

where $\hat{\pi}_k = N_k/N$ $\quad \hat{\mu}_k = \sum_{g_i=k} x_i / N_k$

$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i-\mu_k)(x_i-\mu_k)^T / (N-K)$$

- If Σ_k are different, the boundary is quadratic

$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$ (QDA)



- Regularized discriminant analysis: $\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$ (RDA)

Logistic Regression

- $$\log \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{10}+\beta_1^T X$$
$$\log \frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{20}+\beta_2^T X$$
$$\vdots$$
$$\log \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0}+\beta_{K-1}^T X$$

$$\Rightarrow \quad Pr(G=k|X=x) = \frac{\exp(\beta_{k0}+\beta_k^T X)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T X)}$$
$$Pr(G=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T X)}$$

- Denote $Pr(G=k|X=x) = p_k(x;\theta)$

- log-likelihood: $l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i;\theta)$

- for 2-class case, $y_i=1$ for $g_i=1$, $y_i=0$ for $g_i=2$

$$p_1(x;\theta) = p(x;\theta),\quad p_2(x;\theta) = 1-p(x;\theta)$$

$$l(\beta) = \sum_{i=1}^{N}\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\}$$
$$= \sum_{i=1}^{N}\{y_i\beta^T x_i - \log(1+e^{\beta^T x_i})\}$$

$$\therefore \frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\beta)) = 0$$

and since $x_{i0}=1$, $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} p(x_i;\beta)$

Expect # of class one

observed # of class one

$$\frac{\partial l}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1-p(x_i;\beta))$$

- Optimization $\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)^{-1}\frac{\partial l(\beta)}{\partial\beta}$

- Notations $X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $p = \begin{pmatrix} p(x_1;\theta) \\ p(x_2;\theta) \\ \vdots \\ p(x_n;\theta) \end{pmatrix}$, $W = \mathrm{diag}\{p(x_i;\theta)(1-p(x_i;\theta))\}$

$$\frac{\partial l}{\partial\beta} = X^T(y-p) \qquad \frac{\partial^2 l}{\partial\beta\partial\beta^T} = -X^T W X$$

$$\therefore \beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y-p)$$

- Fisher Information:

Fisher Information Matrix

$$Var(\frac{\partial l}{\partial \beta}) = E((\frac{\partial l}{\partial \beta})^2) - (E(\frac{\partial l}{\partial \beta}))^2 = E(\frac{\partial^2 l}{\partial \beta \partial \beta^T}) = X^T W X$$

$\hat{\beta} \sim N(\beta, (X^T W X)^{-1})$? why $Var(\frac{\partial l}{\partial \beta}) = Var(\hat{\beta})$

- Confidence interval: $\hat{\beta}_j \pm 2\hat{se}_j$ where $\hat{se}_j = \sqrt{((X^T W X)^{-1})_{jj}}$

- Exponential family: $f(y, \theta, \phi) = \exp\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\}$

$E(Y) = b'(\theta)$ and $Var(Y) = \phi b''(\theta)$

- Generalized Linear Models:

  - linear predictor: $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$
  - link function: $E(Y_i) = \mu_i$, $g(\mu_i) = \eta_i$ $\boxed{\text{map from } (-\infty, +\infty)}$
  - Variance function: $var(Y_i) = \phi V(\mu)$

- Binomial Data:

$Y_i \sim Binom(n_i, p_i)$, $E(Y_i/n_i) = p_i$, $Var(Y_i/n_i) = \frac{1}{n_i} p_i(1-p_i)$

  - variance function: $V(\mu_i) = \mu_i(1-\mu_i)$
  - link function: $g(\mu_i) = logit(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$

- Poisson Data:

$Y_i \sim Poisson(\lambda_i)$, $E(Y_i) = \lambda_i$, $Var(Y_i) = \lambda_i$

  - Variance function: $V(\mu_i) = \mu_i$
  - link function: $g(\mu_i) = \log(\mu_i)$