

Hastie, Tibshirani & Friedman: Elements of Statistical Learning Chapter 4.1-4.3

Linear Methods for Classification

CN700/January 31, 2008

Satyavarta

sat@cns.bu.edu

Auditory Neuroscience Laboratory,

Department of Cognitive and Neural Systems,

Boston University, Boston MA 02215



Outline: Linear Methods for Classification

- ⇒ ● Linear Methods for Classification
 - Linear Regression for Classification
 - Linear Discriminant Analysis
 - Dimensionality Reduction

Linear Methods for Classification



Figure 1: 2.1, Hastie et al.

Discriminant Functions

Discriminant Functions

* Two classes x and o

Discriminant Functions

* Two classes x and o

* *Training*: determine *discriminant functions* $\delta_x(X)$ and $\delta_o(X_i)$

Discriminant Functions

- ❄ Two classes x and o
- ❄ *Training*: determine *discriminant functions* $\delta_x(X)$ and $\delta_o(X_i)$
- ❄ *Testing*: Classify a test data point X

Discriminant Functions

✧ Two classes x and o

✧ *Training*: determine *discriminant functions* $\delta_x(X)$ and $\delta_o(X_i)$

✧ *Testing*: Classify a test data point X

★ $\delta_x(X) > \delta_o(X) \Rightarrow X \in x$

★ $\delta_x(X) < \delta_o(X) \Rightarrow X \in o$

Discriminant Functions

- ❄ Two classes x and o
- ❄ *Training*: determine *discriminant functions* $\delta_x(X)$ and $\delta_o(X_i)$
- ❄ *Testing*: Classify a test data point X
 - ★ $\delta_x(X) > \delta_o(X) \Rightarrow X \in x$
 - ★ $\delta_x(X) < \delta_o(X) \Rightarrow X \in o$
- ❄ δ_x estimates in some fashion $Pr(G = x|X = x)$

Linear Discriminant Functions

Linear Discriminant Functions

✻ “Decision boundary is a line.”

Linear Discriminant Functions

✿ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

Linear Discriminant Functions

✱ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

Linear Discriminant Functions

❄ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

❄ Example

Linear Discriminant Functions

✿ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

✿ Example

$$\star \delta_x(X) = \frac{\exp(X)}{1+\exp(X)} \quad \delta_o(X) = \frac{1}{1+\exp(X)}$$

Linear Discriminant Functions

❄ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

❄ Example

$$\star \delta_x(X) = \frac{\exp(X)}{1+\exp(X)} \quad \delta_o(X) = \frac{1}{1+\exp(X)}$$

★ At decision boundary

$$\delta_x(X_d) = \delta_o(X_d) \Rightarrow \frac{\exp(X_d)}{1+\exp(X_d)} = \frac{1}{1+\exp(X_d)}$$

Linear Discriminant Functions

✿ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

✿ Example

$$\star \delta_x(X) = \frac{\exp(X)}{1+\exp(X)} \quad \delta_o(X) = \frac{1}{1+\exp(X)}$$

★ At decision boundary

$$\delta_x(X_d) = \delta_o(X_d) \Rightarrow \frac{\exp(X_d)}{1+\exp(X_d)} = \frac{1}{1+\exp(X_d)}$$

★ Linearizing transformation

$$f(\cdot) = \log\left(\frac{p}{1-p}\right)$$

$$\star f(\delta_x(X)) = X \quad f(\delta_o(X)) = 1$$

Linear Discriminant Functions

❄ “Decision boundary is a line.”

★ If $\delta_x(\cdot)$ and $\delta_o(\cdot)$ are linear,

$$\delta_x(X) = \delta_o(X) \text{ looks like } \beta_0 + \beta^T x = c$$

★ If $f(\cdot)$ is monotonous, and $f(\delta(\cdot))$ are linear,

$$f(\delta_x(X)) = f(\delta_o(X)), \text{ also looks like } \beta_0 + \beta^T x = c$$

❄ Example

$$\star \delta_x(X) = \frac{\exp(X)}{1+\exp(X)} \quad \delta_o(X) = \frac{1}{1+\exp(X)}$$

★ At decision boundary

$$\delta_x(X_d) = \delta_o(X_d) \Rightarrow \frac{\exp(X_d)}{1+\exp(X_d)} = \frac{1}{1+\exp(X_d)}$$

★ Linearizing transformation

$$f(\cdot) = \log\left(\frac{p}{1-p}\right)$$

$$\star f(\delta_x(X)) = X \quad f(\delta_o(X)) = 1$$

★ At decision boundary

$$f(\delta_x(X)) = f(\delta_o(X_d)) \Rightarrow X_d = 1$$

4.1 Notes

* Perceptron and optimally separating hyperplane

* Augment dimensions:

$$X = X_1, X_2, \dots, X_p \rightarrow X_1, X_2, \dots, X_p, X_1^2, X_2^2, \dots, X_p^2, X_1 X_2, \dots, X_i X_j, \dots, X_p X_p$$

* Generally: $X \rightarrow h(X) : \mathbb{R}^p \mapsto \mathbb{R}^q$

\Rightarrow Linear boundary in augmented space \mathbb{R}^q is non-linear in original space \mathbb{R}^p .

Outline: Linear Methods for Classification

- ✓ ● Linear Methods for Classification
- ⇒ ● Linear Regression for Classification
 - Linear Discriminant Analysis
 - Dimensionality Reduction

Linear Regression for Classification

Regress X to indicator matrix Y hoping that Y_i gives same classification as $Pr(G = i/X = x)$

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

Linear Regression for Classification

Regress X to indicator matrix Y hoping that Y_i gives same classification as $Pr(G = i/X = x)$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

For example: Given X , regress onto

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & & & \end{pmatrix}$$

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

✧ *Training*: Simultaneous regression of K functions

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

- ❄ *Training*: Simultaneous regression of K functions
- ❄ *Testing*: For a new X , get $f_1(X), \dots, f_K(X)$

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

- ❄ *Training*: Simultaneous regression of K functions
- ❄ *Testing*: For a new X , get $f_1(X), \dots, f_K(X)$
- ❄ Pick class $k = \arg \max_{k \in \mathcal{G}} \hat{f}_k(X)$

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

- ❄ *Training*: Simultaneous regression of K functions
- ❄ *Testing*: For a new X , get $f_1(X), \dots, f_K(X)$
- ❄ Pick class $k = \arg \max_{k \in \mathcal{G}} \hat{f}_k(X)$
- ❄ *Training*: Regression on K -dimensional outputs (with unit L_1 -norm)

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

❄ *Training:* Simultaneous regression of K functions

❄ *Testing:* For a new X , get $f_1(X), \dots, f_K(X)$

❄ Pick class $k = \arg \max_{k \in \mathcal{G}} \hat{f}_k(X)$

❄ *Training:* Regression on K -dimensional outputs (with unit L_1 -norm)

❄ *Testing:* Each X_i produces a K -dimensional output vector Y_i

Linear Regression of an Indicator Matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1K} \\ I_{21} & I_{22} & \dots & I_{2K} \\ \vdots & & & \\ I_{N1} & I_{N2} & \dots & I_{NK} \end{pmatrix}$$

❄ *Training:* Simultaneous regression of K functions

❄ *Testing:* For a new X , get $f_1(X), \dots, f_K(X)$

❄ Pick class $k = \arg \max_{k \in \mathcal{G}} \hat{f}_k(X)$

❄ *Training:* Regression on K -dimensional outputs (with unit L_1 -norm)

❄ *Testing:* Each X_i produces a K -dimensional output vector Y_i

❄ Pick class whose t_k is closest to Y_i

Masking

Figure 2: 4.3, Hastie et al.

Masking

- ❄ Gauss Markov Theorem: OLS provides best linear *unbiased* estimator

Figure 2: 4.3, Hastie et al.

Masking

- ❄ Gauss Markov Theorem: OLS provides best linear *unbiased* estimator
- ❄ Expected value of each indicator dimension $1 \dots K$ is the same

Figure 2: 4.3, Hastie et al.

Masking

- ❄ Gauss Markov Theorem: OLS provides best linear *unbiased* estimator
- ❄ Expected value of each indicator dimension $1 \dots K$ is the same
- ❄ All K linear approximations pass through the expected value, i.e. have varying slopes about the same point. **Only extreme values are ever selected**

Figure 2: 4.3, Hastie et al.

Masking

- ❄ Gauss Markov Theorem: OLS provides best linear *unbiased* estimator
- ❄ Expected value of each indicator dimension $1 \dots K$ is the same
- ❄ All K linear approximations pass through the expected value, i.e. have varying slopes about the same point. **Only extreme values are ever selected**
- ❄ All other values are masked. Need augmentation of dimensions to reach them.

Figure 2: 4.3, Hastie et al.

Masking

- ❄ Gauss Markov Theorem: OLS provides best linear *unbiased* estimator
- ❄ Expected value of each indicator dimension $1 \dots K$ is the same
- ❄ All K linear approximations pass through the expected value, i.e. have varying slopes about the same point. **Only extreme values are ever selected**
- ❄ All other values are masked. Need augmentation of dimensions to reach them.
- ❄ In general, to resolve K classes, augment dimensions upto degree $K - 1$, to have $O(p^{K-1})$ terms in all.

Figure 2: 4.3, Hastie et al.

Masking Hurts

Figure 3: 4.3, Hastie et al.

Masking Hurts

❄ For this example, Quadratic polynomials overcome masking.

Figure 3: 4.3, Hastie et al.

Masking Hurts

- ❄ For this example, Quadratic polynomials overcome masking.
- ❄ On vowel dataset, with $K=11$ classes in $N=10$ dimensions,

Technique	Err_{train}	Err_{test}
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

Figure 3: 4.3, Hastie et al.

Outline: Linear Methods for Classification

- ✓ • Linear Methods for Classification
- ✓ • Linear Regression for Classification
- ⇒ • Linear Discriminant Analysis
 - Dimensionality Reduction

Linear Discriminant Analysis

Linear Discriminant Analysis

* In Linear Regression for classification: Y_i stands in for $Pr(G = i/X = x)$

Linear Discriminant Analysis

- * In Linear Regression for classification: Y_i stands in for $Pr(G = i/X = x)$
- * Can $Pr(G = i/X = x)$ be modeled directly?

Linear Discriminant Analysis

❄ In Linear Regression for classification: Y_i stands in for $Pr(G = i/X = x)$

❄ Can $Pr(G = i/X = x)$ be modeled directly?

❄ Bayes Theorem:

$$\begin{aligned} Pr(G = k|X = x) &= \frac{Pr(G=k, X=x)}{Pr(X=x)} \\ &= \frac{Pr(G=k, X=x)}{\sum_k Pr(G=k, X=x)} \\ &= \frac{Pr(G=k)Pr(X=x|G=k)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &= \frac{\pi_k f_k(X=x)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &\propto \pi_k f_k(X = x) \end{aligned}$$

Linear Discriminant Analysis

❄ In Linear Regression for classification: Y_i stands in for $Pr(G = i|X = x)$

❄ Can $Pr(G = i|X = x)$ be modeled directly?

❄ Bayes Theorem:

$$\begin{aligned} Pr(G = k|X = x) &= \frac{Pr(G=k, X=x)}{Pr(X=x)} \\ &= \frac{Pr(G=k, X=x)}{\sum_k Pr(G=k, X=x)} \\ &= \frac{Pr(G=k)Pr(X=x|G=k)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &= \frac{\pi_k f_k(X=x)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &\propto \pi_k f_k(X = x) \end{aligned}$$

❄ Modeling class-conditional density $f_k(X) \equiv Pr(G = k|X = x)$ for classification

Linear Discriminant Analysis

❄ In Linear Regression for classification: Y_i stands in for $Pr(G = i|X = x)$

❄ Can $Pr(G = i|X = x)$ be modeled directly?

❄ Bayes Theorem:

$$\begin{aligned} Pr(G = k|X = x) &= \frac{Pr(G=k, X=x)}{Pr(X=x)} \\ &= \frac{Pr(G=k, X=x)}{\sum_k Pr(G=k, X=x)} \\ &= \frac{Pr(G=k)Pr(X=x|G=k)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &= \frac{\pi_k f_k(X=x)}{\sum_k Pr(G=k)Pr(X=x|G=k)} \\ &\propto \pi_k f_k(X = x) \end{aligned}$$

❄ Modeling class–conditional density $f_k(X) \equiv Pr(G = k|X = x)$ for classification

❄ Model $f_k(\cdot)$ as:

★ Linear and Quadratic discriminant analysis: $f_k(\cdot) =$ Gaussians

★ $f_k(\cdot) =$ Mixture of Gaussians

★ $f_k(\cdot) =$ Non–parametric density estimate from training data

★ Naïve Bayes: $f_k(X) = f_k(X_1)f_k(X_2) \dots f_k(X_p)$

Geometry of $f_k(\cdot)$

Figure 4: 4.5, Hastie et al.

Geometry of $f_k(\cdot)$

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

Figure 4: 4.5, Hastie et al.

Geometry of $f_k(\cdot)$

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$\ast f_{red}(X)$ is probability that a point X belongs to the red class, and it is distributed in p dimensions $f_{red}(X) \sim \mathcal{N}(\mu_{red}, \Sigma_{red})$

Figure 4: 4.5, Hastie et al.

Geometry of $f_k(\cdot)$

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$\ast f_{red}(X)$ is probability that a point X belongs to the red class, and it is distributed in p dimensions $f_{red}(X) \sim \mathcal{N}(\mu_{red}, \Sigma_{red})$

\ast Given a test data point X at μ_{red} , perform $K - 1$ comparisons:

★ f_{blue} vs. $f_{green} \Rightarrow X \in blue$

★ f_{blue} vs. $f_{red} \Rightarrow X \in red$

Figure 4: 4.5, Hastie et al.

Geometry: Linear/Quadratic Discriminant _____

Figure 4.5(a), Hastie et al.

Equal covariances for all groups $\Sigma_k = \Sigma \forall k$

\Rightarrow Linear boundaries

Textbook figure 4.6(a), Hastie et al.

Different covariances for each group Σ_k

\Rightarrow Quadratic boundaries

Mathematics: Linear/Quadratic Discriminant _____

Mathematics: Linear/Quadratic Discriminant _____

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

Mathematics: Linear/Quadratic Discriminant _____

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$$\ast Pr(G = k|X = x) = f_k(x)\pi_k \quad Pr(G = l|X = x) = f_l(x)\pi_l$$

Mathematics: Linear/Quadratic Discriminant ---

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$$\ast Pr(G = k|X = x) = f_k(x)\pi_k \quad Pr(G = l|X = x) = f_l(x)\pi_l$$

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

Mathematics: Linear/Quadratic Discriminant

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$$\ast Pr(G = k|X = x) = f_k(x)\pi_k \quad Pr(G = l|X = x) = f_l(x)\pi_l$$

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

\ast Linear Discriminant:

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

Mathematics: Linear/Quadratic Discriminant

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$$\ast Pr(G = k|X = x) = f_k(x)\pi_k \quad Pr(G = l|X = x) = f_l(x)\pi_l$$

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

\ast Linear Discriminant:

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

\ast Quadratic Discriminant:

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

Mathematics: Linear/Quadratic Discriminant

$$\ast f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

$$\ast Pr(G = k|X = x) = f_k(x)\pi_k \quad Pr(G = l|X = x) = f_l(x)\pi_l$$

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

\ast Linear Discriminant:

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

\ast Quadratic Discriminant:

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

\ast Estimates

★ $\mu_k \approx$ sample mean of $x \in$ Group k

★ $\hat{\Sigma} \approx$ within-group sample variance of $x \in$ Group k

$$\sum_{k \in K} \sum_{x_i \in G_k} (x_i - \mu_k)^T (x_i - \mu_k)$$

★ $\hat{\pi}_k \approx N_k/N$, where $N_k = |G_k|$ and $N = \#$ of training points.

4.3.0 Notes

❄ QDA on given space, and LDA on variable space augmented with polynomial terms gives similar results.

❄ Number of parameters:

★ LDA : $(K - 1)(p + 1)$ parameters ($\frac{p(p+1)}{2} + (K - 1)(p + 1)$?)

★ QDA : $(K - 1)(\frac{p(p+3)}{2} + 1)$ parameters ($(K - 1)(\frac{p(p+1)}{2} + (p + 1)$?)

❄ LDA and QDA are simple and work well.

❄ Continuum between LDA and QDA: per-group covariance approaches global covariance:

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma}$$

Outline: Linear Methods for Classification

- ✓ • Linear Methods for Classification
- ✓ • Linear Regression for Classification
- ✓ • Linear Discriminant Analysis
- ⇒ • Dimensionality Reduction

Dimensionality Reduction

Figure 4.9, Hastie et al.

Projection from 2D to 1D, in a particular direction and in direction of the maximal linear discriminant.

Dimensionality Reduction: from 10D to 2D, various _____



Figure 4.8, Hastie et al.

Dimensionality Reduction: from 10D to 2D, LDA _____



Figure 4.11, Hastie et al.

Dimensionality Reduction: performance _____



Figure 4.10, Hastie et al.

Dimensionality Reduction: Mechanics

- ❄ Compute centroids M (a $K \times p$ matrix) and the common covariance W
- ❄ Sphere the centroids: $M^* = MW^{\frac{1}{2}}$
- ❄ Compute B^* , the covariance matrix of M^* . This is the between-class covariance matrix.
- ❄ Eigen-decompose $B^* = V^*D_BV^{*T}$
- ❄ Columns of V^* are coordinates of optimal subspaces in order of discriminability.

Dimensionality Reduction: Fisher's Linear Discriminant _____

✿ "Find the linear combination $Z = a^T X$ such that the between-class variance is maximized relative to the within-class variance"

✿ Maximize the *Rayleigh coefficient*

$$\max_a \frac{a^T B a}{a^T W a}$$

Outline: Linear Methods for Classification

- ✓ • Linear Methods for Classification
- ✓ • Linear Regression for Classification
- ✓ • Linear Discriminant Analysis
- ✓ • Dimensionality Reduction