

Chapter 4 - classification from

Hastie et al. (2009) and James et al. (2013)

Lecturers: Geir Storvik and Aliaksandr Hubin

Classification

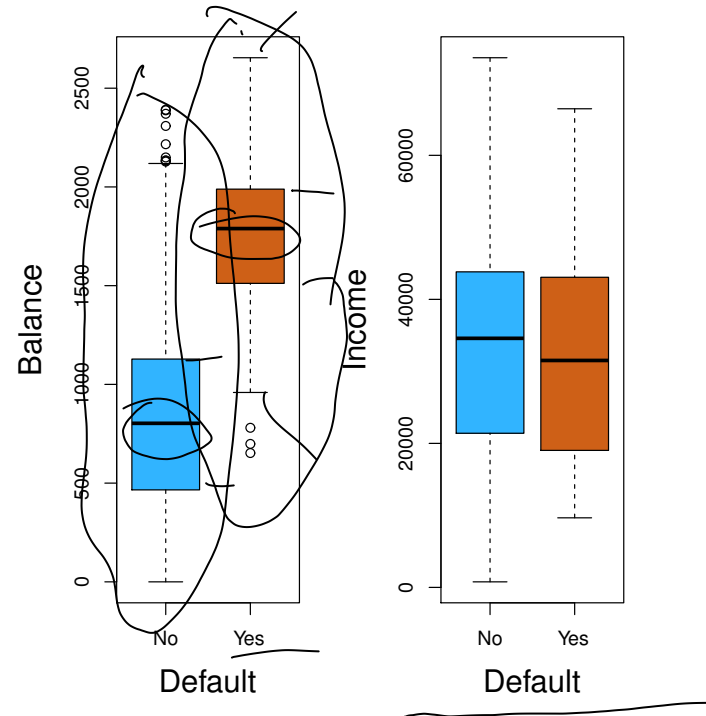
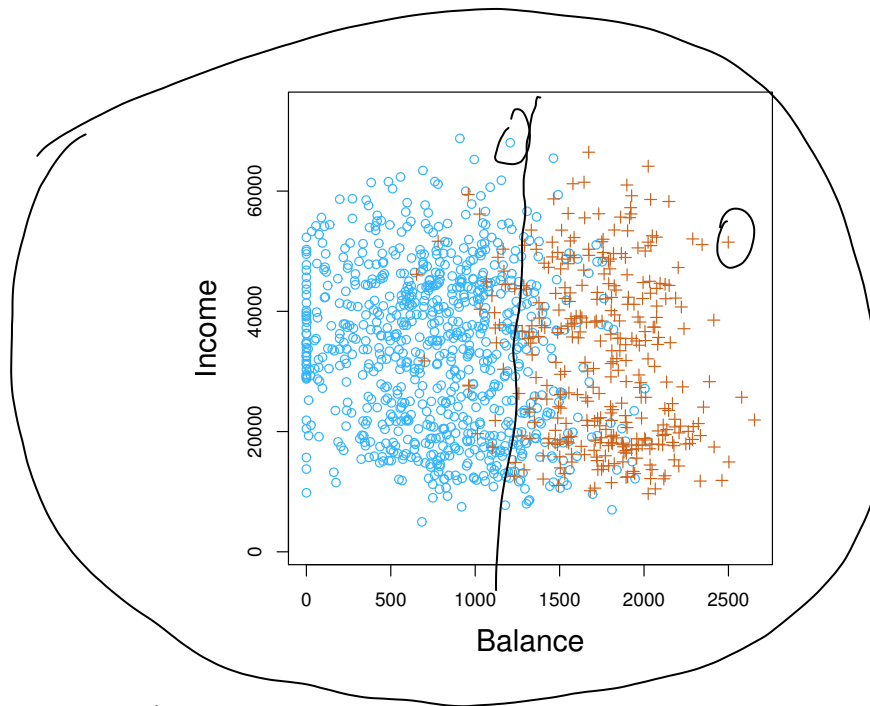
- ▶ If response is qualitative / categorical: Classification
 - ▶ Pictures of faces: Man or woman?
 - ▶ Increase / decrease in share value
 - ▶ Color of eyes
 - ▶ Diagnosis of various diseases
 - ▶ Is a transaction normal or fraud?

Classification

- ▶ If response is qualitative / categorical: Classification
 - ▶ Pictures of faces: Man or woman?
 - ▶ Increase / decrease in share value
 - ▶ Color of eyes
 - ▶ Diagnosis of various diseases
 - ▶ Is a transaction normal or fraud?
- ▶ Bayes classifier: $\hat{Y} = \arg \max_j \Pr(Y = j | \mathbf{X} = \mathbf{x})$
- ▶ Probability an **quantitative** measure
- ▶ Methods for analyzing probabilities very similar to regression.
- ▶ Chapter 4:
 - ▶ Logistic regression
 - ▶ Linear and quadratic discriminant analysis
 - ▶ K-nearest neighbors
 - ▶ Separating hyper-planes
- ▶ Later: Trees, random forest, boosting, support vector machines

Classification - overview

- ▶ Data $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, $y_i \in \{1, \dots, K\}$
- ▶ Default dataset
 - ▶ Response Y : Will customer end up with a credit default (No / Yes)
 - ▶ Explanatory variables: x_1 income (Income) and x_2 credit card limit (Balance)



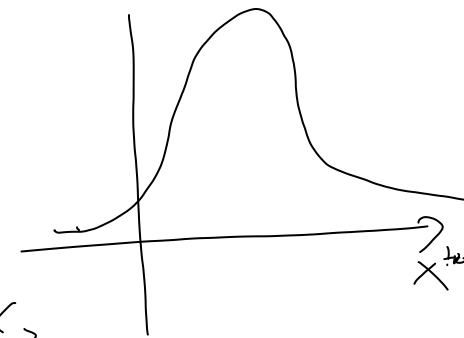
¹ Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

- ▶ One wants to **predict** Y from x_1 and x_2 .

Why not linear regression?

$$Z \sim \mathcal{N}(\mu(x), \sigma^2)$$

$$\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



- ▶ Default dataset

- Can code Yes as $Z = 1$ and No as $Z = 0$ and use linear model

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- ▶ Prediction: $\hat{z} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- ▶ Classification:

$$\hat{y} = \begin{cases} \text{Yes} & \text{if } \hat{z} > \epsilon; \\ \text{No} & \text{else} \end{cases}$$

$$\hat{z} \in (-\infty; +\infty)$$

- ▶ A bit ad hoc?
- ▶ How to interpret $\hat{z} < 0$ or $\hat{z} > 1$?
- ▶ Not obvious how to generalize to multiple classes

Example with multiple classes

$$\hat{y} \in (-\infty; a^*] \Rightarrow 1$$

$$\hat{y} \in (a^*, b^*] \Rightarrow 2$$

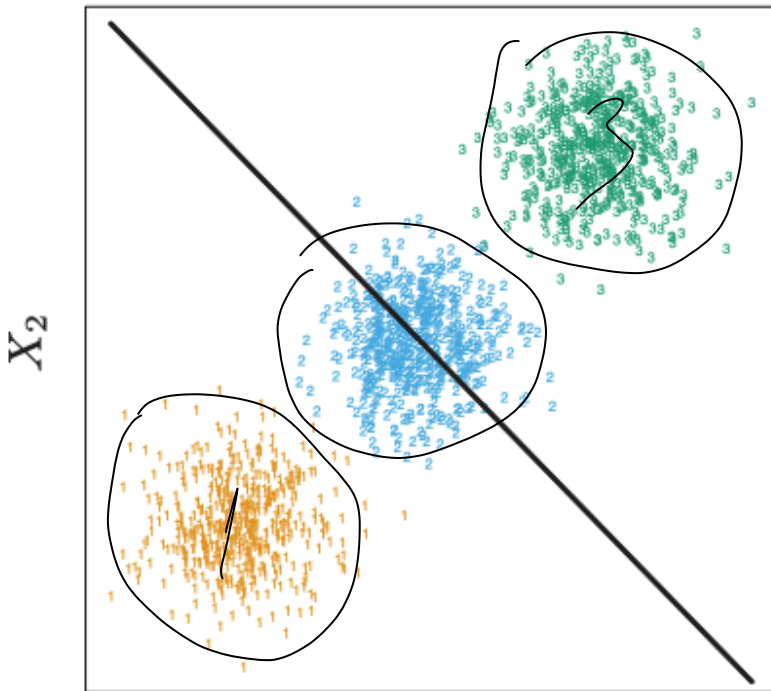
$$\hat{y} \in (b^*, +\infty) \Rightarrow 3$$

- ▶ Consider $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{If overdose} \\ 3 & \text{If epileptic seizures} \end{cases}$ Can also here use linear regression with numeric response.
- ▶ Problem: Other answer if we code $Y = \begin{cases} 1 & \text{If epileptic seizures} \\ 2 & \text{if stroke} \\ 3 & \text{If overdose} \end{cases}$
- ▶ Difficult to convert a categorical variable without scheme to a numeric variable.

Example with multiple classes

- ▶ Consider $\begin{cases} Y_1 = I(\text{if stroke}) \in \{0, 1\} \\ Y_2 = I(\text{If overdose}) \in \{0, 1\} \\ Y_3 = I(\text{If epileptic seizures}) \in \{0, 1\} \end{cases}$
- ▶ Regress $\mathbf{x}$ on Y_1, Y_2 and Y_3
- ▶ A new observation is classified by computing its fitted vector and choosing the class with the largest expectation
 $\arg \max_{i \in \{1, 2, 3\}} (\hat{\mu}_1(\mathbf{x}), \hat{\mu}_2(\mathbf{x}), \hat{\mu}_3(\mathbf{x}))$ across the classes
- ▶ Masking can appear where we can ignore a particular class (i.e.
 $\hat{\mu}_2(\mathbf{x}) < \hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_2(\mathbf{x}) < \hat{\mu}_3(\mathbf{x})$ for all $\mathbf{x}$)

Linear Regression



Logistic regression

$$Y \sim \mathcal{N}(\underbrace{\mu(x)}, \sigma^2) \quad \mu(x) \in (-\infty, +\infty)$$
$$\mu(x) = \underbrace{x^T \beta}_{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$$

- ▶ Method for **classification** with **two** possible outcomes
- ▶ Model ($\mathbf{x}^T \beta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$) and assume:

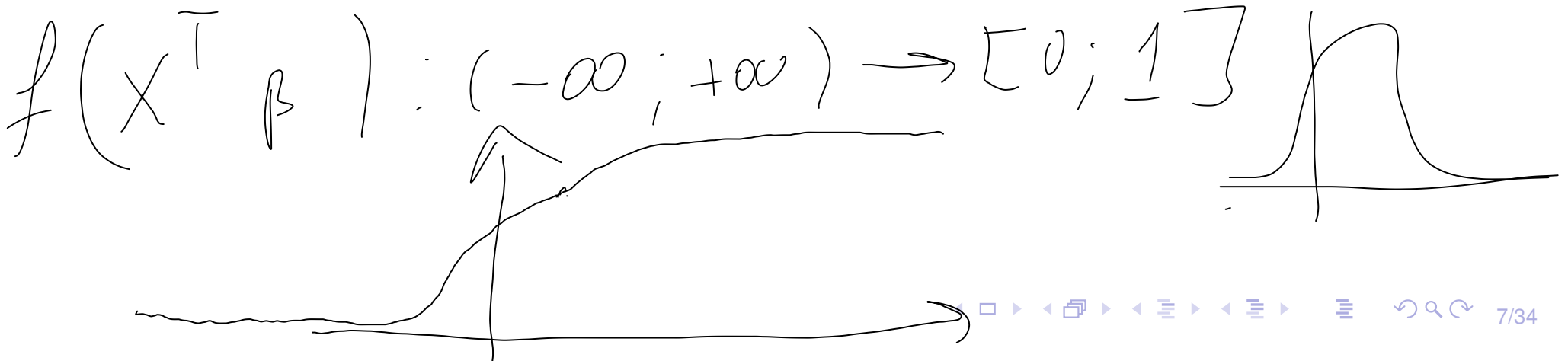
$$Y \sim \text{Bernoulli}(p(\mathbf{x})) = p(x)^Y (1 - p(x))^{1-Y} \quad ; \quad p(x) \in [0, 1]$$

$$p(\mathbf{x}) = \Pr(Y = 1 | \mathbf{x}) = \frac{e^{\mathbf{x}^T \beta}}{1 + e^{\mathbf{x}^T \beta}} = 1 - \Pr(Y = 0 | \mathbf{x})$$

Note:

$$\log \left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} \right) = \underbrace{\mathbf{x}_{1 \times p}^T \beta_{p \times 1}}_{\beta_{1 \times p}^T \mathbf{x}_{p \times 1}} \text{ just a scalar } 1 \times 1$$

Part of **Generalized linear models** (GLM)



Logistic regression

- ▶ Method for **classification** with **two** possible outcomes
- ▶ Model $(\mathbf{x}^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$ and assume:

$$Y \sim \text{Bernoulli}(p(\mathbf{x})) = p(x)^Y (1 - p(x))^{1-Y}$$

$$p(\mathbf{x}) = \Pr(Y = 1|\mathbf{x}) = \frac{e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}} = 1 - \Pr(Y = 0|\mathbf{x})$$

Note:

$$\log \left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} \right) = \mathbf{x}_{1 \times p}^T \boldsymbol{\beta}_{p \times 1} = \boldsymbol{\beta}_{1 \times p}^T \mathbf{x}_{p \times 1} \text{ just a scalar } 1 \times 1$$

Part of **Generalized linear models** (GLM)

- ▶ Data: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ (assumed independent)

- ▶ Likelihood $L(\boldsymbol{\beta}) = \prod_{i=1}^n p(\mathbf{x}_i)^{y_i} (1 - p(\mathbf{x}_i))^{1-y_i}$

$$= \prod_{i=1}^n \left[\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right]^{y_i} \left[\frac{1}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right]^{1-y_i}$$

- ▶ ML estimate: Numeric optimization

Interpretation and significance

- For logistic regression

$$\log \left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} \right) = \log \frac{\Pr(Y = 1|\mathbf{x})}{\Pr(Y = 0|\mathbf{x})} = \eta(\mathbf{x}) = \beta_0 + \sum_{j=1}^p \beta_j x_j$$

$\frac{\Pr(Y=1|\mathbf{x})}{\Pr(Y=0|\mathbf{x})}$ called **odds**

default data set - confounding

```
> fit1 <- glm(default~student, family="binomial", data=Default)
> summary(fit1)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.5041278	0.07071301	-49.554219	0.0000000000
studentYes	0.4048871	0.11501883	3.520181	0.0004312529

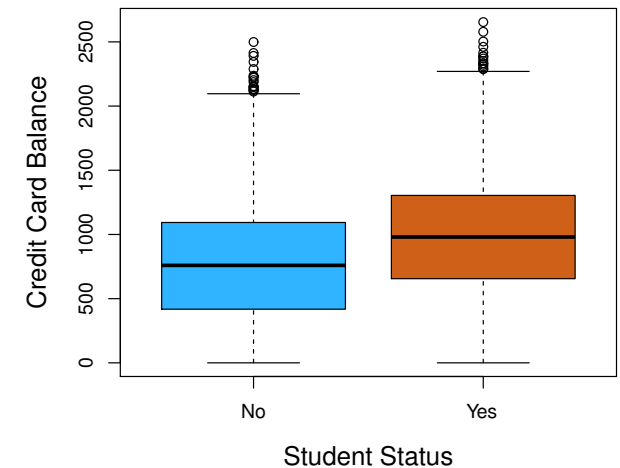
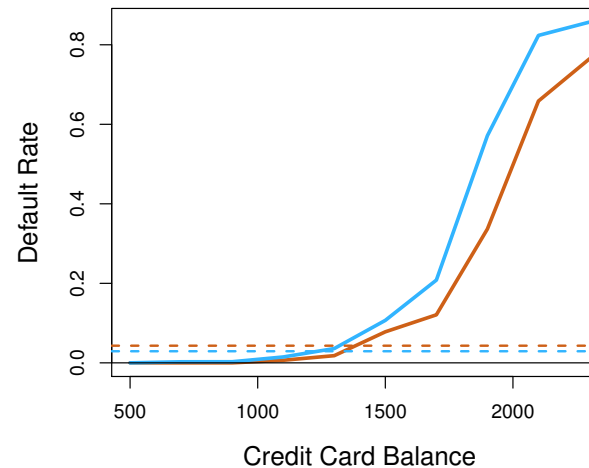
```
> fit <- glm(default~student+balance+income, family="binomial", data=Default)
> summary(fit)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.086905e+01	4.922555e-01	-22.080088	4.911280e-108
studentYes	-6.467758e-01	2.362525e-01	-2.737646	6.188063e-03
balance	5.736505e-03	2.318945e-04	24.737563	4.219578e-135
income	3.033450e-06	8.202615e-06	0.369815	7.115203e-01

default data set - confounding

```
> fit1 <- glm(default~student, family="binomial", data=Default)
> summary(fit1)$coef
              Estimate Std. Error   z value    Pr(>|z|)
(Intercept) -3.5041278  0.07071301 -49.554219 0.0000000000
studentYes   0.4048871  0.11501883   3.520181 0.0004312529>
fit <- glm(default~student+balance+income, family="binomial", data=Default)
> summary(fit)$coef
              Estimate Std. Error   z value    Pr(>|z|)
(Intercept) -1.086905e+01 4.922555e-01 -22.080088 4.911280e-108
studentYes   -6.467758e-01 2.362525e-01  -2.737646 6.188063e-03
balance       5.736505e-03 2.318945e-04  24.737563 4.219578e-135
income        3.033450e-06 8.202615e-06   0.369815 7.115203e-01
```

- ▶ The sign of β_j for student **changes** when more covariates are added
- ▶ The result of confounding



Students - orange, non-students - blue.

Logistic regression and class boundary

- Reasonable classification rule:

$$\hat{y} = \begin{cases} 1 & \text{if } \Pr(Y = 1|\mathbf{x}) > 0.5; \\ 0 & \text{else} \end{cases}$$

$$\Pr(Y = 1|\mathbf{x}) > 0.5$$

 $\Updownarrow$

$$\frac{e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}} > 0.5$$

 $\Updownarrow$

$$e^{\mathbf{x}^T \boldsymbol{\beta}} > 0.5[1 + e^{\mathbf{x}^T \boldsymbol{\beta}}]$$

 $\Updownarrow$

$$e^{\mathbf{x}^T \boldsymbol{\beta}} > 1$$

 $\Updownarrow$

$$\mathbf{x}^T \boldsymbol{\beta} > 0$$

- I.e. **linear** boundary between classes

Linear Discriminant Analysis (LDA)

- Bayes theorem

$$\Pr(Y = k|\mathbf{x}) = \frac{\Pr(Y = k)p(\mathbf{x}|Y = k)}{p(\mathbf{x})} = \frac{\Pr(Y = k)p_k(\mathbf{x})}{p(\mathbf{x})}$$

- Linear discriminant analysis: Modelling of

$$\begin{array}{ll} \pi_k = \Pr(Y = k) & \text{Prior} \\ p_k(\mathbf{x}) = p(\mathbf{x}|Y = k) & \text{conditional density/probability for } \mathbf{x} \end{array}$$

Linear Discriminant Analysis (LDA)

- Bayes theorem

$$\Pr(Y = k|\mathbf{x}) = \frac{\Pr(Y = k)p(\mathbf{x}|Y = k)}{p(\mathbf{x})} = \frac{\Pr(Y = k)p_k(\mathbf{x})}{p(\mathbf{x})}$$

- Linear discriminant analysis: Modelling of

$$\begin{aligned}\pi_k &= \Pr(Y = k) && \text{Prior} \\ p_k(\mathbf{x}) &= p(\mathbf{x}|Y = k) && \text{conditional density/probability for } \mathbf{x}\end{aligned}$$

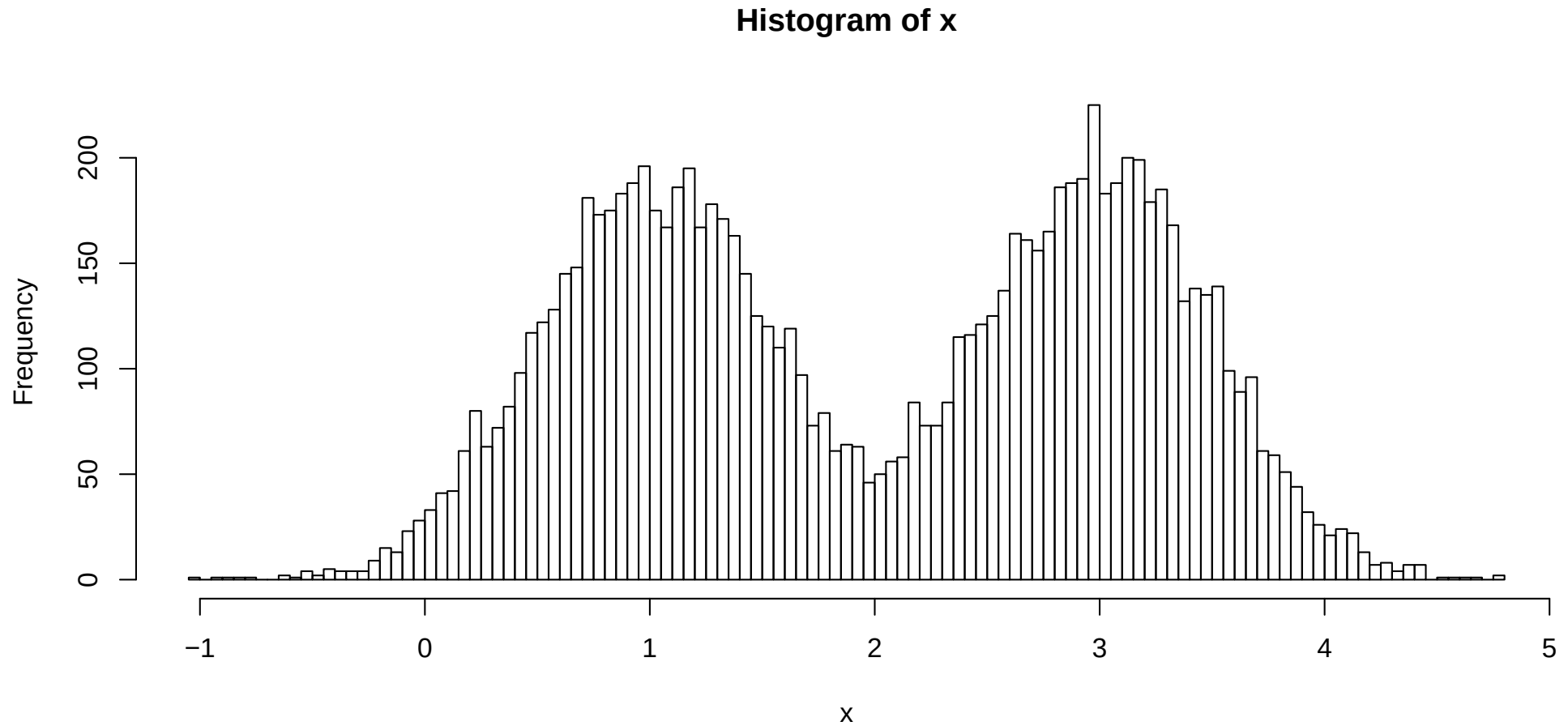
- Assume $p = 1$ and $p_k(x) = N(\mu_k, \sigma^2)$. Then

$$\Pr(Y = k|x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- Note: Marginally,

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} \pi_k p_k(\mathbf{x})$$

Illustration mixture distribution



- ▶ Mixture of two Gaussians
- ▶ Should be possible to guess/estimate centerpoints and variances even if class labels are unknown!

LDA - classification rule

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

$$\hat{y} = \arg \max_k p_k(x)$$

$$= \arg \max_k \left[\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right) \right]$$

$$= \arg \max_k \left[\log(\pi_k) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2}(x - \mu_k)^2 \right]$$

$$= \arg \max_k \left[\log(\pi_k) - \frac{1}{2\sigma^2} [x^2 - 2\mu_k x + \mu_k^2] \right]$$

$$= \arg \max_k \left[\log(\pi_k) + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \right] = \arg \max_k \delta_k(x)$$

LDA - classification rule

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

$$\hat{y} = \arg \max_k p_k(x)$$

$$= \arg \max_k \left[\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right) \right]$$

$$= \arg \max_k \left[\log(\pi_k) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2}(x - \mu_k)^2 \right]$$

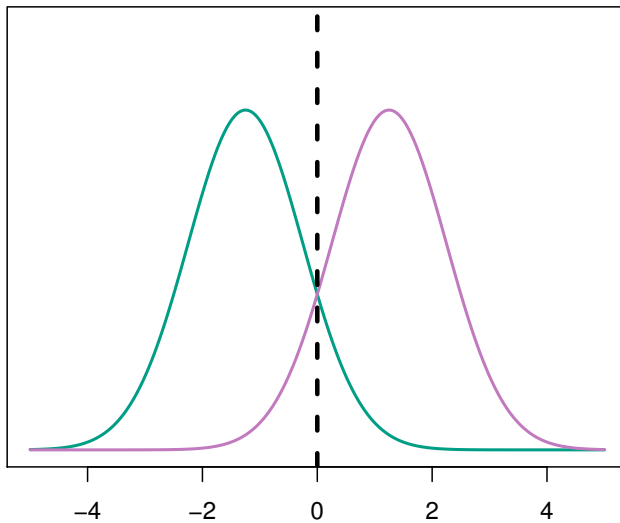
$$= \arg \max_k \left[\log(\pi_k) - \frac{1}{2\sigma^2} [x^2 - 2\mu_k x + \mu_k^2] \right]$$

$$= \arg \max_k \left[\log(\pi_k) + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \right] = \arg \max_k \delta_k(x)$$

Note also **linear** decision boundaries between classes!

Illustration

- ▶ $\arg \max_k \delta_k(x) = \arg \max_k \left[\log(\pi_k) + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \right]$
- ▶ $K = 2, \pi_k = 0.5$: Border for $x = \frac{\mu_1 + \mu_2}{2}$ is at 0
- ▶ $\mu_1 = -1.25, \mu_2 = 1.25, \sigma^2 = 1$



LDA - estimation

- ▶ Unknown parameters: $\{\pi_k, \mu_k, k = 1, \dots, K\}, \sigma$
- ▶ Assume **training data** $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

LDA - estimation

- ▶ Unknown parameters: $\{\pi_k, \mu_k, k = 1, \dots, K\}, \sigma$
- ▶ Assume **training data** $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$n_k = \sum_i I(y_i = k)$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^K \frac{n_k - 1}{n - K} \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^K \frac{n_k - 1}{n - K} \hat{\sigma}_k^2$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

LDA in practice

► We had

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

LDA in practice

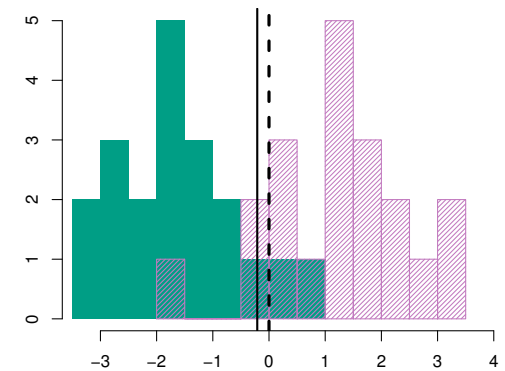
- We had

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

- LDA, calculate

$$\hat{\delta}_k(x) = \log(\hat{\pi}_k) - \frac{1}{2} \frac{\hat{\mu}_k^2}{\hat{\sigma}^2} + \frac{\hat{\mu}_k}{\hat{\sigma}^2} x$$

and classify to the largest **discriminant function**



LDA and $p > 1$

- ▶ $p = 1 : p_k(x) = N(\mu_k, \sigma^2)$
- ▶ $p > 1 : p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

- ▶ $x_j \sim N(\mu_{k,j}, \Sigma_{jj})$
- ▶ $\mathbf{a}^t \mathbf{x} \sim N(\mathbf{a}^t \boldsymbol{\mu}_k, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$
- ▶ estimation

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i:y_i=k} \mathbf{x}_i$$

$$n_k = \sum_i I(y_i = k)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\begin{aligned} \hat{\delta}_k(\mathbf{x}) &= \log(\hat{\pi}_k) - \frac{1}{2} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) \\ &= \log(\hat{\pi}_k) + \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k \end{aligned}$$

LDA and $p > 1$

- ▶ $p = 1 : p_k(x) = N(\mu_k, \sigma^2)$
- ▶ $p > 1 : p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

LDA and $p > 1$

- ▶ $p = 1 : p_k(x) = N(\mu_k, \sigma^2)$
- ▶ $p > 1 : p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

- ▶ $x_j | Y = k \sim N(\mu_{k,j}, \Sigma_{jj})$
- ▶ $\mathbf{a}^T \mathbf{x} | Y = k \sim N(\mathbf{a}^T \boldsymbol{\mu}_k, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$
- ▶ Classification rule

$$\begin{aligned} \hat{y} &= \arg \max_k \pi_k p_k(\mathbf{x}) \\ &= \arg \max_k [\log \pi_k + \log p_k(\mathbf{x})] \\ &= \arg \max_k \left[\log(\pi_k) - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right] \\ &= \arg \max_k [\log(\pi_k) - \frac{1}{2} [\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k]] \\ &= \arg \max_k [\log(\pi_k) - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x}] \\ &= \arg \max_k \delta_k(\mathbf{x}) \end{aligned}$$

Estimation for $p > 1$

► Estimation

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i: y_i = k} \mathbf{x}_i$$

$$n_k = \sum_i I(y_i = k)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i: y_i = k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

Estimation for $p > 1$

- ▶ Estimation

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i: y_i = k} \mathbf{x}_i$$

$$n_k = \sum_i I(y_i = k)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i: y_i = k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

- ▶ Classification rule based on

$$\begin{aligned} \hat{\delta}_k(\mathbf{x}) &= \log(\hat{\pi}_k) - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x} \\ &= a_k + \mathbf{b}_k^T \mathbf{x} \end{aligned}$$

- ▶ Note: $a_k, \mathbf{b}_k$ can be precomputed!

LDA and boundaries

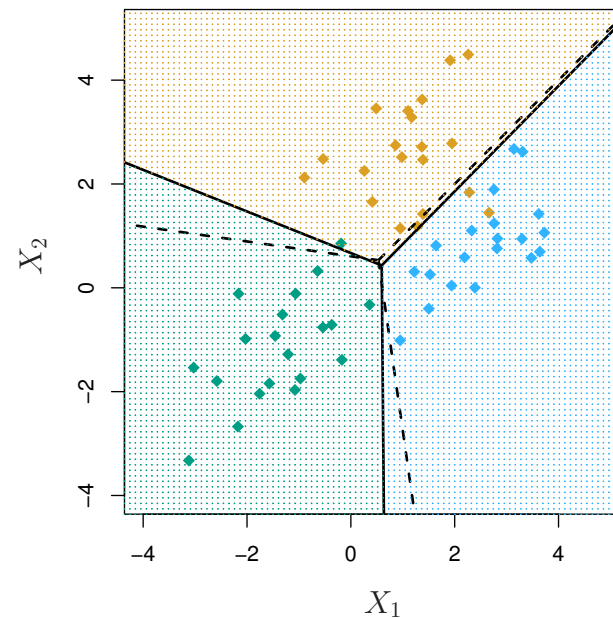
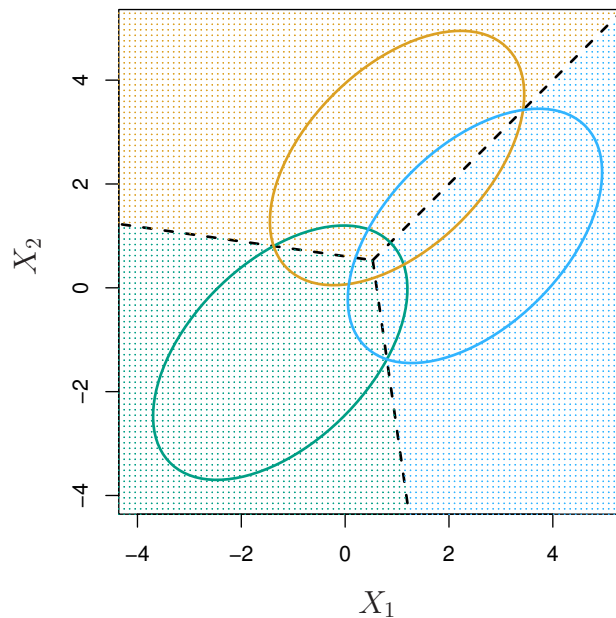
► $\hat{\delta}_k(\mathbf{x}) = \log(\hat{\pi}_k) + \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k$

$$\hat{\delta}_k(\mathbf{x}) = \hat{\delta}_l(\mathbf{x})$$

$\Updownarrow$

$$[\hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\mu}}_l^T \hat{\boldsymbol{\Sigma}}^{-1}] \mathbf{x} = \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_l^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_l - \log(\hat{\pi}_k) + \log(\hat{\pi}_l)$$

► Linear boundaries



LDA and boundaries

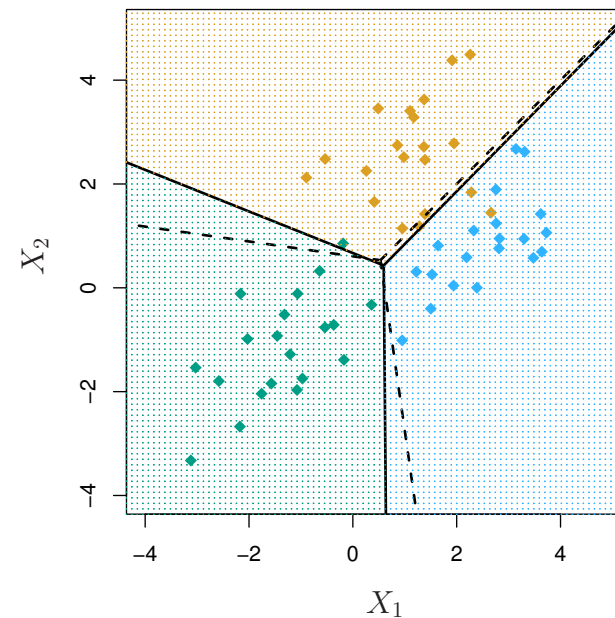
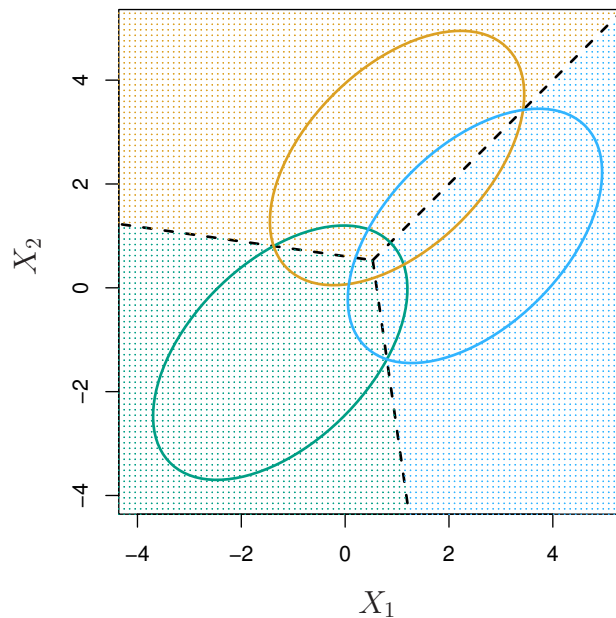
► $\hat{\delta}_k(\mathbf{x}) = \log(\hat{\pi}_k) + \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k$

$$\hat{\delta}_k(\mathbf{x}) = \hat{\delta}_l(\mathbf{x})$$

$\Updownarrow$

$$[\hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\mu}}_l^T \hat{\boldsymbol{\Sigma}}^{-1}] \mathbf{x} = \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_l^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_l - \log(\hat{\pi}_k) + \log(\hat{\pi}_l)$$

► Linear boundaries



LDA in R

```
> library(MASS)
> fit.Ida <- lda(default~balance+income+student, data=Default)
> pred.Ida <- predict(fit.Ida, Default)$class
> table(Default$default, pred.Ida)
```

	No	Yes
No	9645	22
Yes	254	79

#Which prediction method is used?

```
> class(fit.Ida)
[1] "lda"
> help(predict.lda)
```

Quadratic discriminant analysis (QDA)

- Bayes theorem

$$p_k(\mathbf{x}) = \Pr(Y = k|\mathbf{x}) = \frac{\Pr(Y = k)p(\mathbf{x}|Y = k)}{p(\mathbf{x})}$$

- Consider $p = 1$ and $p_k(x) = N(\mu_k, \sigma_k^2)$. Then we get

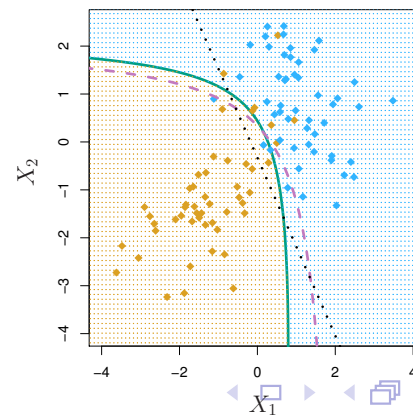
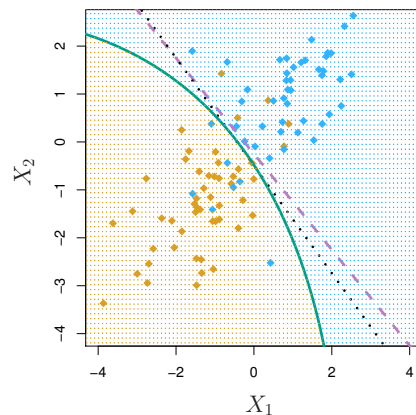
$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x - \mu_l)^2\right)}$$

QDA - classification rule

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x - \mu_l)^2\right)}$$

$$\begin{aligned}\hat{y} &= \arg \max_k p_k(x) = \arg \max_k \left[\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \right] \\ &= \arg \max_k \left[\log(\pi_k) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2\sigma_k^2}(x - \mu_k)^2 \right] \\ &= \arg \max_k \left[\log(\pi_k) - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2\sigma_k^2}[x^2 - 2\mu_k x + \mu_k^2] \right] \\ &= \arg \max_k \left[\log(\pi_k) - \frac{1}{2} \log(\sigma_k^2) - \frac{x^2}{2\sigma_k^2} + \frac{\mu_k x}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} \right] = \arg \max_k \delta_k(x)\end{aligned}$$

LDA - black dotted, QDA - green solid



QDA and $p > 1$

- ▶ $p = 1 : p_k(x) = N(\mu_k, \sigma_k^2)$
- ▶ $p > 1 : p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- ▶ Estimating

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i: y_i = k} \mathbf{x}_i, \quad n_k = \sum_i I(y_i = k)$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k - 1} \sum_{i: y_i = k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\begin{aligned} \hat{\delta}_k(\mathbf{x}) &= \log(\pi_k) - \frac{1}{2} \log(|\hat{\boldsymbol{\Sigma}}_k|) + (\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) \\ &= \log(\hat{\pi}_k) - \frac{1}{2} \log(|\hat{\boldsymbol{\Sigma}}_k|) - \frac{1}{2} \mathbf{x}^T \hat{\boldsymbol{\Sigma}}_k^{-1} \mathbf{x} + \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}_k^{-1} \mathbf{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}_k^{-1} \hat{\boldsymbol{\mu}}_k \end{aligned}$$

QDA in R

```
> library(MASS)
> fit.qda <- qda(default~balance+income+student , data=Default)
> pred.qda <- predict(fit.qda, Default)
> table(Default$default , pred.qda$class)
```

	No	Yes
No	9636	31
Yes	239	94

#Which prediction method is used?

```
> class(fit.qda)
[1] "qda"
> help(predict.qda)
```

K-nearest neighbors

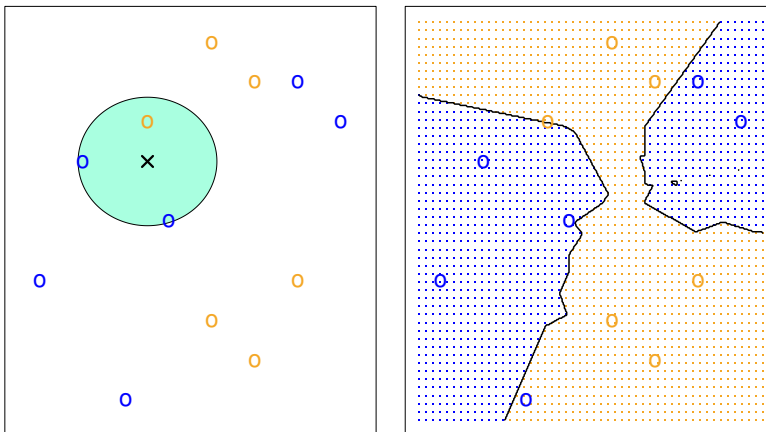
- ▶ k -nearest neighbor: Let $\mathcal{N}_0 \subset \{1, \dots, n\}$ be the k indexes with $\mathbf{x}_i$ nearest to $\mathbf{x}_0$.
- ▶ Estimation of $\Pr(Y = k | \mathbf{X} = \mathbf{x}_0) = E[I(Y = k | \mathbf{X} = \mathbf{x}_0)]$:

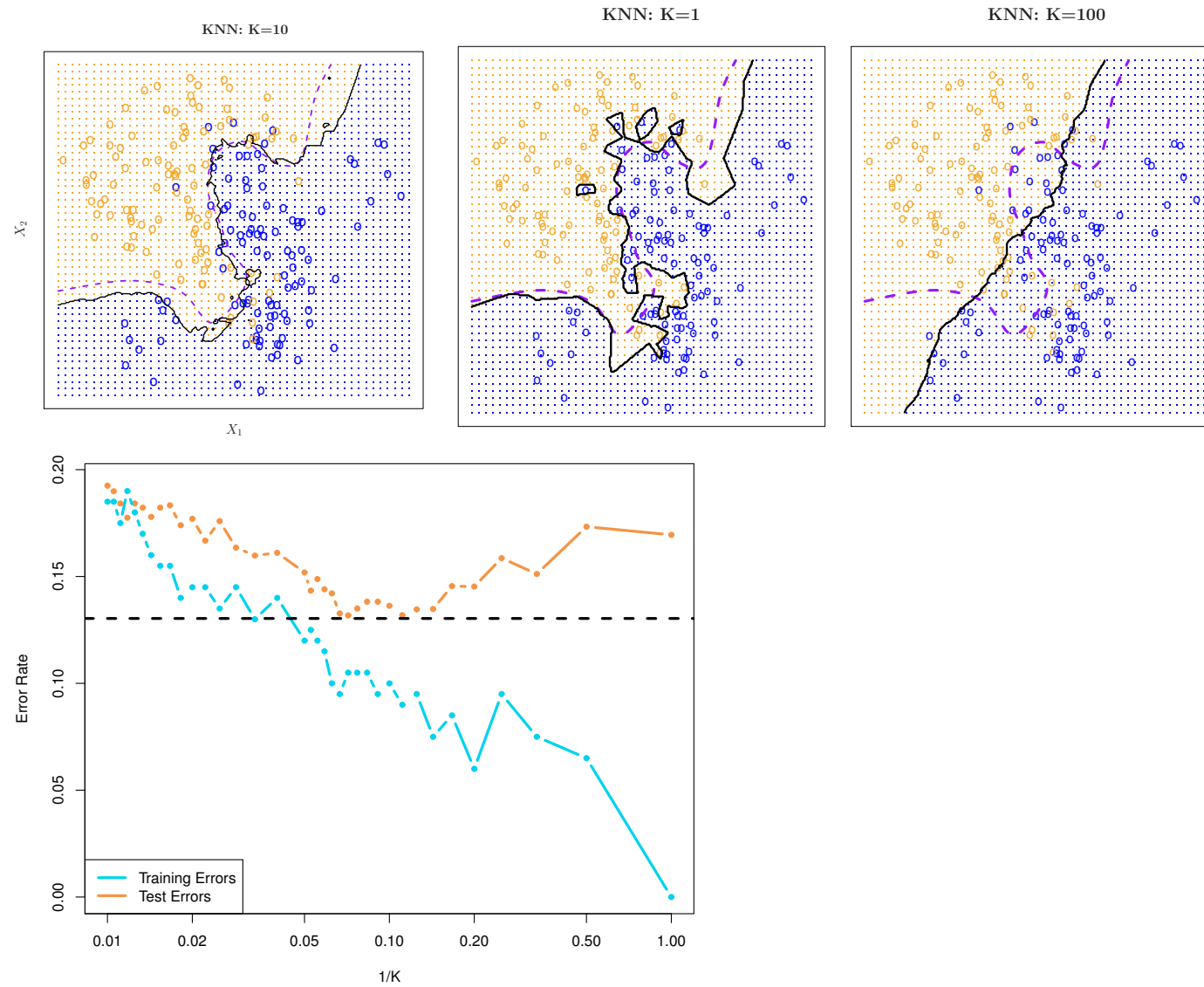
$$\hat{\Pr}(Y = k | \mathbf{X} = \mathbf{x}) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = k)$$

- ▶ Classification rule:

$$\hat{Y} = \arg \max_k \hat{\Pr}(Y = k | \mathbf{X} = \mathbf{x})$$

Classify test point with $\mathbf{X} = \mathbf{x}^*$ to the class with the highest number of cases among the K nearest points.



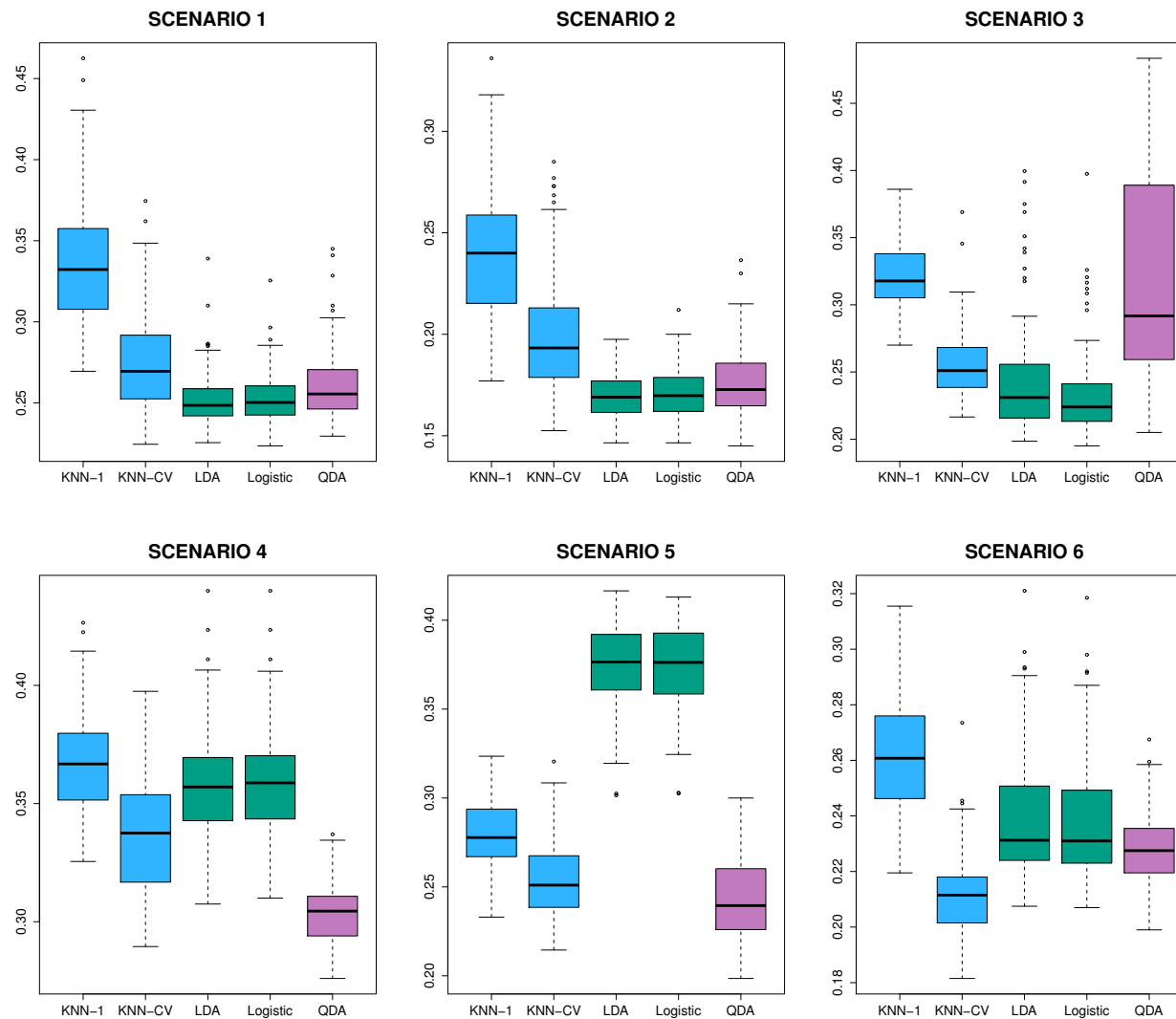


Comparisons

$p = 2, K = 2$

- ▶ Scenario 1: $n_k = 20, p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}), k = 1, 2$
- ▶ Scenario 2: $n_k = 20, p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), k = 1, 2, \text{Cor}(x_1, x_2) = -0.5$
- ▶ Scenario 3: Like Scenario 1 but t -distribution and $n_k = 50$.
- ▶ Scenario 4: $n_k = 20, p_k(\mathbf{x}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, 2, \text{Cor}(x_1, x_2) = -0.5$ for $k = 1$ and $\text{Cor}(x_1, x_2) = 0.5$ for $k = 2$
- ▶ Scenario 5: Logistic regression model with quadratic terms and all pairwise interactions
- ▶ Scenario 6: Logistic regression model with a complex non-linearity

Comparisons based on simulations



No uniformly best performing method!

Regularized discriminant analysis - prevent overfitting

- ▶ Allow to choose between the common and class specific covariance matrices creates a form of regularization:

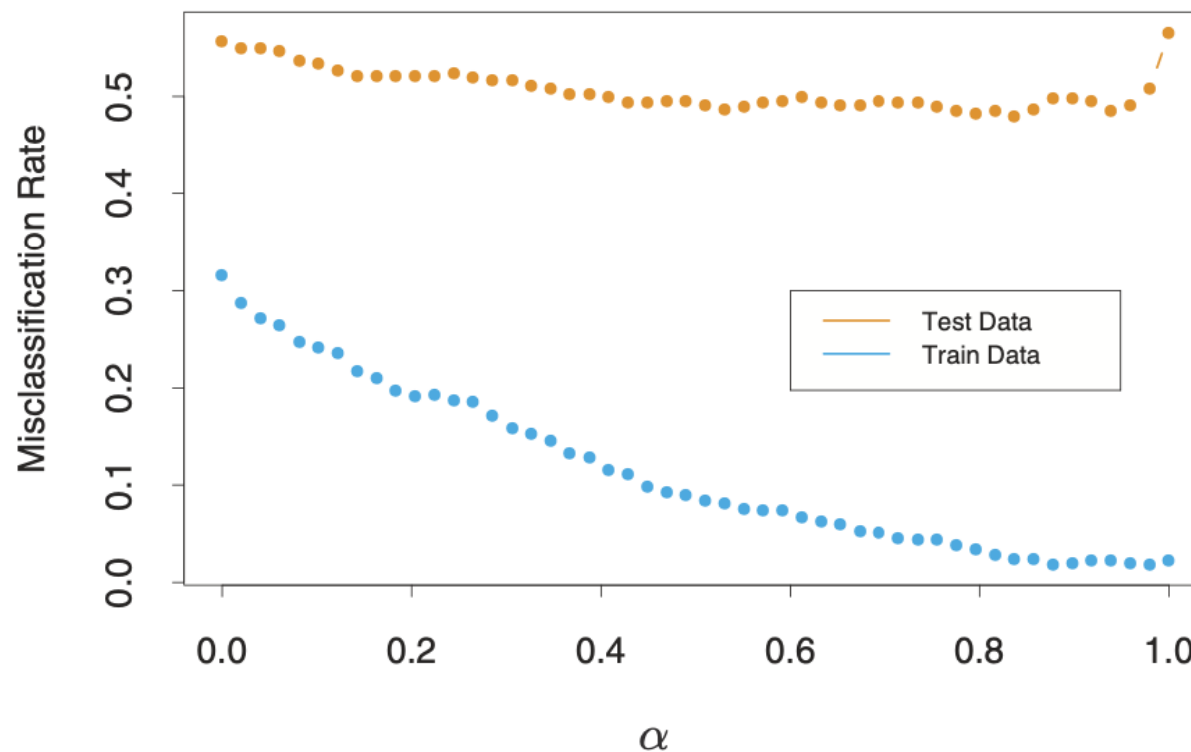
$$\hat{\Sigma}_k(a) = a\hat{\Sigma}_k + (1 - a)\hat{\Sigma}, a \in [0, 1]. \quad (1)$$

- ▶ Also possible to choose between the covariance structure or a shared common variance with no dependence in a similar way:

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1 - \gamma)\hat{\sigma}^2\mathbf{I}, a \in [0, 1]. \quad (2)$$

- ▶ Mixing the three variants of covariance could also be possible

Regularized Discriminant Analysis on the Vowel Data



Computations of LDA or QDA

- Compute eigen decomposition of each covariance matrix:

$$\hat{\Sigma}_k = U_k D_k U_k^T, \text{ where } U_k \text{ is orthonormal and } D_k \text{ is diagonal.} \quad (3)$$

- As Σ_k^{-1} decomposes as $U_k D_k^{-1} U_k^T$ we simplify computing $\delta_k(x)$ to

$$(x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) = [U_k^T (x - \hat{\mu}_k)]^T D_k^{-1} [U_k^T (x - \hat{\mu}_k)]. \quad (4)$$

$$\log |\hat{\Sigma}_k| = \sum_l \log d_{kl}, \quad (5)$$

where d_{kl} are eigen values

- This allows to avoid costly matrix inversion and gives matrix determinants for free.
- We can sphere the data with the common covariance prior to analysis by setting

$$X \longleftarrow D^{-0.5} U^T X$$

where $\hat{\Sigma} = U D U^T$ and get the common covariance as identity

References

- T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.