Classification

- ⋆ Situation:
  - observations $x_1, \ldots, x_n$ with class labels $y_1, \ldots, y_n$, where $y_i \in \{0, 1, \ldots, J-1\}$
  - new observation: x_0
  - want to classify x_0 to one of the J possible classes

Classification

- ⋆ Situation:
  - observations $x_1, \ldots, x_n$ with class labels $y_1, \ldots, y_n$, where $y_i \in \{0, 1, \ldots, J-1\}$
  - new observation: x_0
  - want to classify x_0 to one of the J possible classes
- ⋆ Model:

$$f(x|y = j) = f_j(x) \quad \text{and} \quad p_j = P(Y = j)$$

Classification

- ⋆ Situation:
  - observations $x_1, \ldots, x_n$ with class labels $y_1, \ldots, y_n$, where $y_i \in \{0, 1, \ldots, J-1\}$
  - new observation: x_0
  - want to classify x_0 to one of the J possible classes
- ⋆ Model:

$$f(x|y=j) = f_j(x) \quad \text{and} \quad p_j = P(Y=j)$$

- ⋆ Distribution of interest

$$\pi_j(x_0) = P(Y_0 = j | X = x_0) = \frac{p_j f_j(x_0)}{\sum_{i=0}^{J-1} p_i f_i(x_0)}$$

Classification

- ⋆ Situation:
  - observations $x_1, \ldots, x_n$ with class labels $y_1, \ldots, y_n$, where $y_i \in \{0, 1, \ldots, J-1\}$
  - new observation: x_0
  - want to classify x_0 to one of the J possible classes
- ⋆ Model:

$$f(x|y=j) = f_j(x) \quad \text{and} \quad p_j = P(Y=j)$$

- ⋆ Distribution of interest

$$\pi_j(x_0) = P(Y_0 = j | X = x_0) = \frac{p_j f_j(x_0)}{\sum_{i=0}^{J-1} p_i f_i(x_0)}$$

- ⋆ Cost function

$$c(i|j): \text{ cost of classifying to } i \text{ when true class is } j$$

  - $c(i|i) = 0$ and $c(i|j) > 0$ for $i \neq j$

Classification

- Expected cost of classify to class i:

$$\begin{aligned} \text{ECM}(i) &= \text{E}[c(i|Y)|x_0] = \sum_{j=0}^{J-1} c(i|j) P(Y_0 = j | X = x_0) \\ &= \sum_{j=0}^{J-1} c(i|j) \pi_j(x_0) = \frac{\sum_{j=0}^{J-1} c(i|j) p_j f_j(x_0)}{\sum_{k=0}^{J-1} p_k f_k(x_0)} \end{aligned}$$

- Classification rule:

$$\widehat{y}_0 = \underset{i}{\text{argmin}}\, \text{ECM}(i)$$

Classification

- Expected cost of classify to class i:

$$\begin{aligned} \text{ECM}(i) &= \text{E}[c(i|Y)|x_0] = \sum_{j=0}^{J-1} c(i|j) P(Y_0 = j | X = x_0) \\ &= \sum_{j=0}^{J-1} c(i|j) \pi_j(x_0) = \frac{\sum_{j=0}^{J-1} c(i|j) p_j f_j(x_0)}{\sum_{k=0}^{J-1} p_k f_k(x_0)} \end{aligned}$$

- Classification rule:

$$\widehat{y}_0 = \underset{i}{\text{argmin}}\, \text{ECM}(i)$$

- Zero-one loss:

$$c(i|j) = \begin{cases} 1 & \text{for } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

- Bayes classifier

$$\widehat{y}_0 = \underset{i}{\text{argmin}} \{p_i f_i(x_0)\}$$

Linear discriminant analysis (LDA)

- ⋆ Assume zero-one loss and $x|y = j \sim N(\mu_j, \Sigma)$
- ⋆ Classification rule becomes

$$\widehat{y}_0 = \underset{i}{\text{argmax}}\{\widehat{\delta}_i(x_0)\}$$

where

$$\widehat{\delta}_i(x_0) = x_0^T \widehat{\Sigma}^{-1} \widehat{\mu}_i - \frac{1}{2} \widehat{\mu}_i^T \widehat{\Sigma}^{-1} \widehat{\mu}_i + \ln(\widehat{p}_i)$$

Linear discriminant analysis (LDA)

- ⋆ Assume zero-one loss and $x|y = j \sim N(\mu_j, \Sigma)$
- ⋆ Classification rule becomes

$$\widehat{y}_0 = \underset{i}{\text{argmax}}\{\widehat{\delta}_i(x_0)\}$$

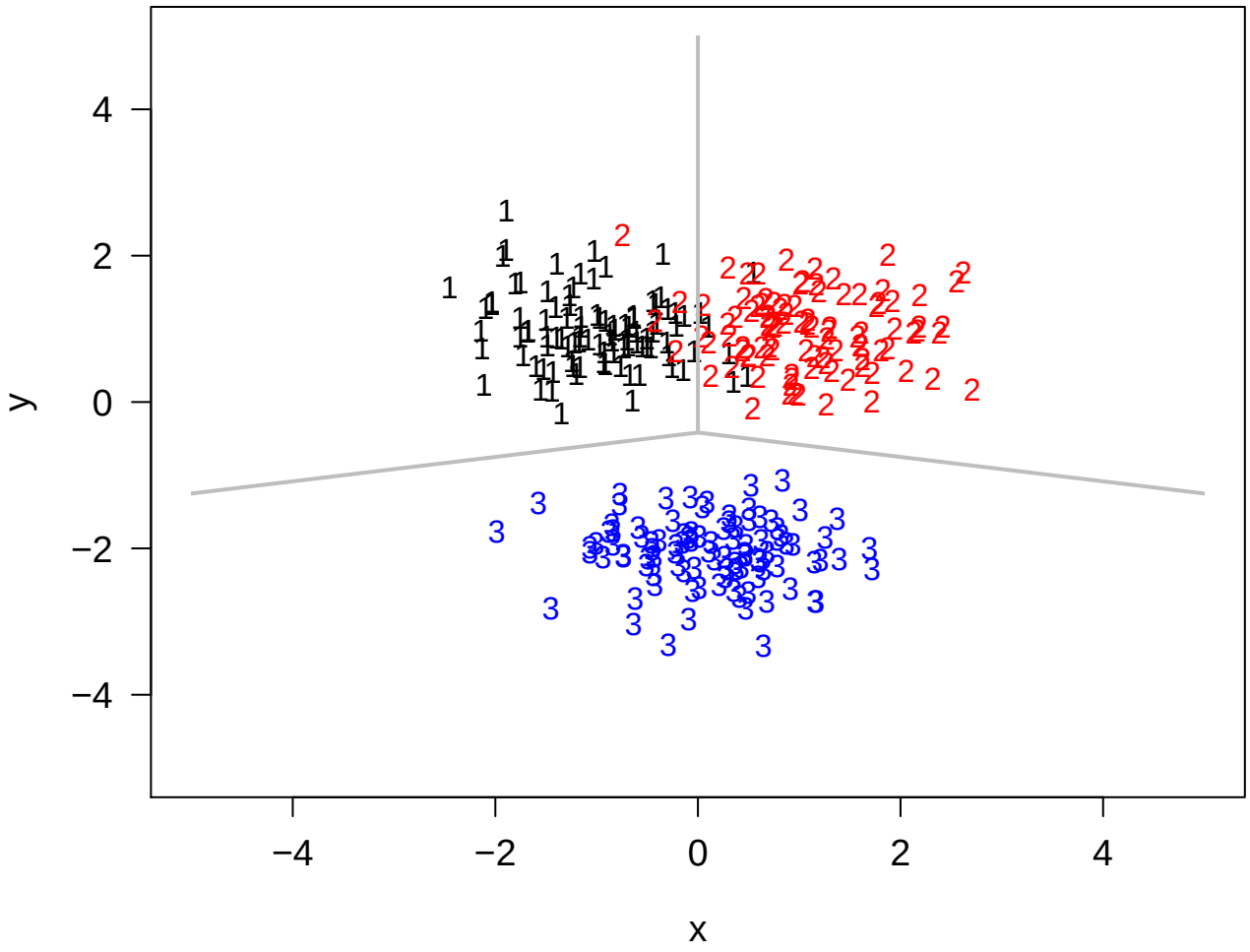
where

$$\widehat{\delta}_i(x_0) = x_0^T \widehat{\Sigma}^{-1} \widehat{\mu}_i - \frac{1}{2} \widehat{\mu}_i^T \widehat{\Sigma}^{-1} \widehat{\mu}_i + \ln(\widehat{p}_i)$$

- ⋆ $\widehat{\delta}_i(x_0)$ is linear in x_0. Borders between classification regions becomes lines/planes/hyer-planes

LDA

Note: $\hat{\delta}_i(x_0)$ is linear in x_0. Thus the Bayes decision borders between the classification regions become lines/hyper-planes.



Quadratic discriminant analysis (QDA)

- ⋆ Assume zero-one loss and $x|y = j \sim N(\mu_j, \Sigma_j)$
- ⋆ Classification rule becomes

$$\widehat{y}_0 = \underset{i}{\text{argmax}}\{\widehat{\delta}_i(x_0)\}$$

where

$$\widehat{\delta}_i(x_0) = -\frac{1}{2}(x_0 - \widehat{\mu}_i)^T \widehat{\Sigma}_i^{-1}(x_0 - \widehat{\mu}_i) - \frac{1}{2}\ln|\widehat{\Sigma}_i| + \ln(\widehat{p}_i)$$

Quadratic discriminant analysis (QDA)

- ⋆ Assume zero-one loss and $x|y = j \sim N(\mu_j, \Sigma_j)$
- ⋆ Classification rule becomes

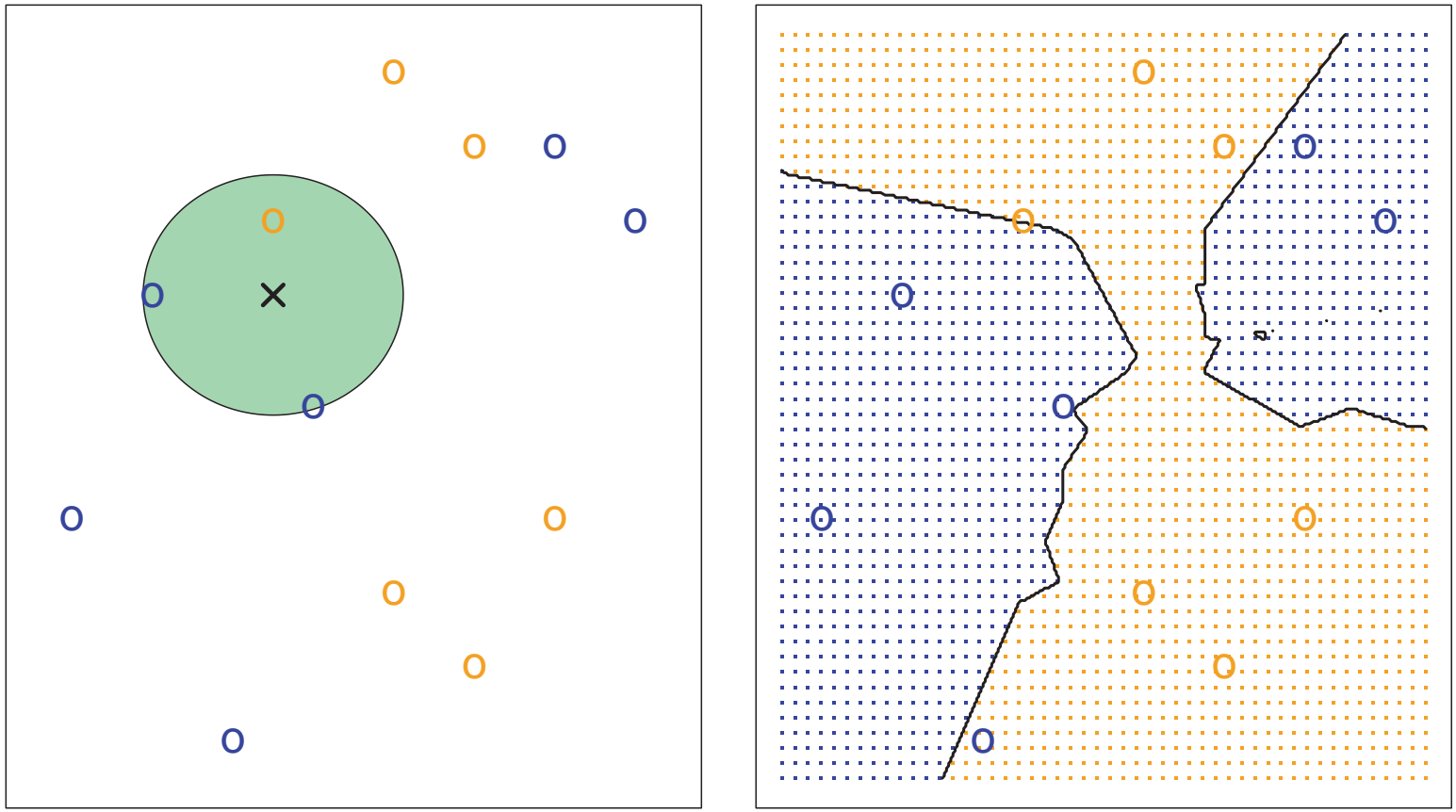
$$\widehat{y}_0 = \underset{i}{\text{argmax}}\{\widehat{\delta}_i(x_0)\}$$

where

$$\widehat{\delta}_i(x_0) = -\frac{1}{2}(x_0 - \widehat{\mu}_i)^T \widehat{\Sigma}_i^{-1}(x_0 - \widehat{\mu}_i) - \frac{1}{2}\ln|\widehat{\Sigma}_i| + \ln(\widehat{p}_i)$$

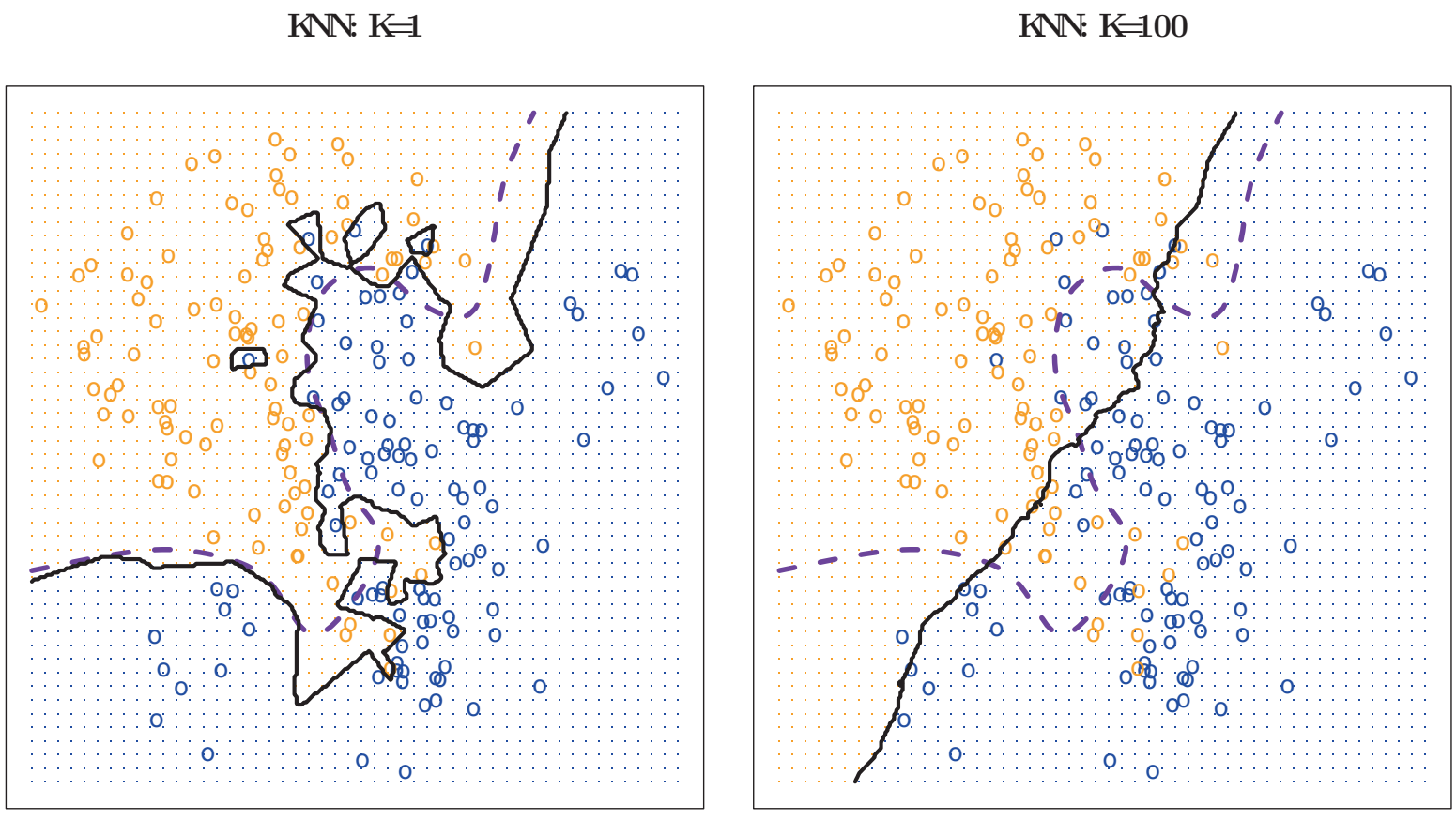
- ⋆ $\widehat{\delta}_i(x_0)$ is quadratic in x_0. Borders between classification regions becomes quadratic

KNN classification - Example K=3



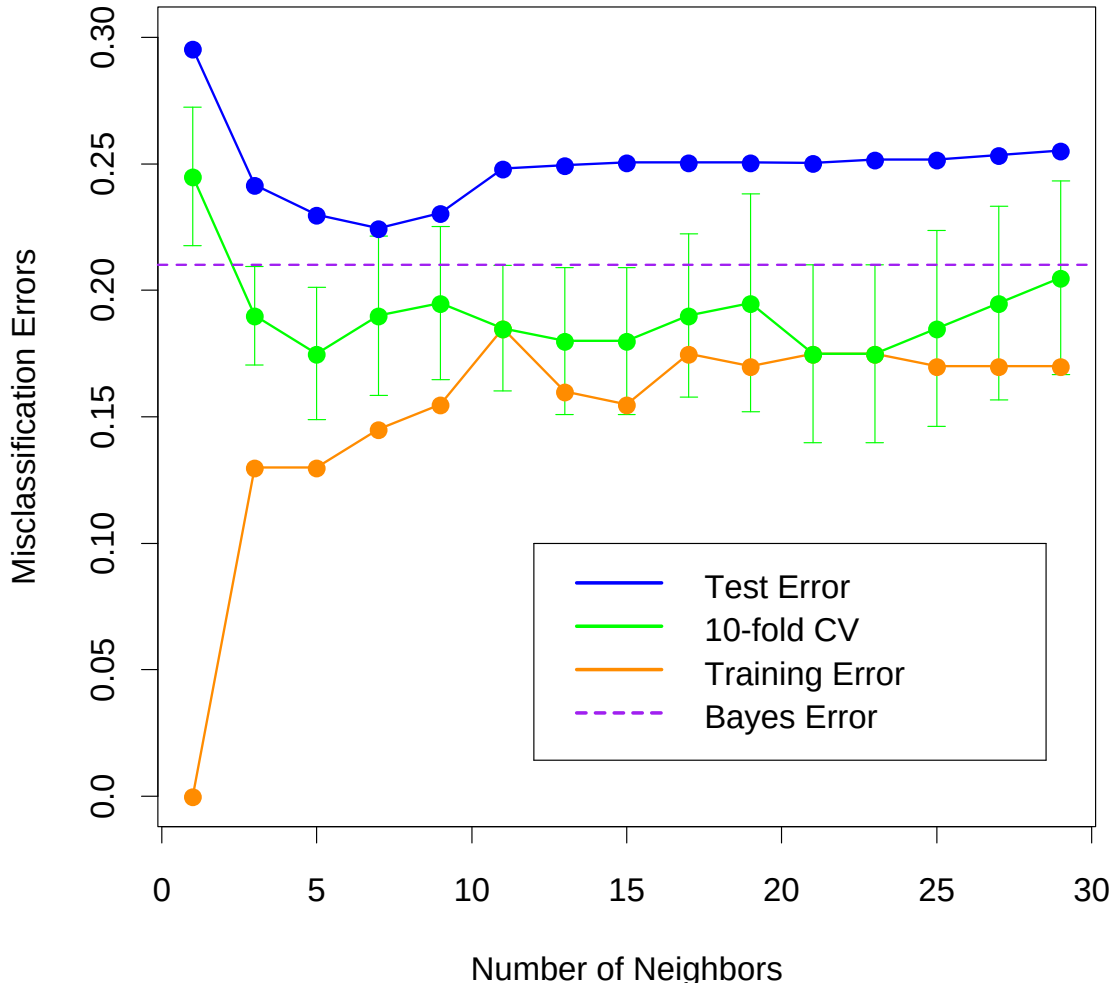
(James, Witten, Tibshirani, Hastie (2014), An Introduction to Statistical Learning, Springer, p.40)

k-nearest-neighbour classifiers



(James, Witten, Tibshirani, Hastie (2014), An Introduction to Statistical Learning, Springer, p.41)

Misclassification as function of k



Hastie, Tibshirani, Friedman, "The elements of statistical learning", 2nd ed., p. 467

Bootstrap

Bootstrap

http://tradingconsequences.blogs.edina.ac.uk/files/2013/10/Dr_Martens_black_old.jpg

. . . pull oneself up by one's bootstraps

> *To begin an enterprise or recover from a setback without any outside help; to succeed only on one's own effort or abilities.*
>
> Wiktionary

The term is sometimes attributed to Rudolf Erich Raspe's story "The Surprising Adventures of Baron Munchausen", where the main character pulls himself (and his horse) out of a swamp by his hair

http://redstateeclectic.typepad.com/redstate_commentary/2010/11/sustainability-isnt-sustainable.html

Bootstrap Bill Turner

"Bootstrap" Bill Turner from Pirates of the Caribbean.

> *. . . Barbossa tied Bill to a cannon by his bootstraps and sent him to the bottom of the sea.*

http://kidstvmovies.about.com/od/piratesofthecaribbean3/ig/Pirates-At-World-s-End/-Bootstrap-Bill.htm