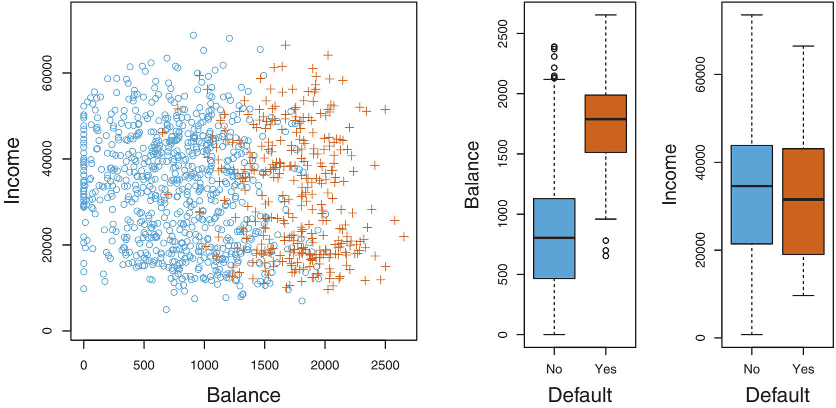
Data Analysis

Methods of classification

National Research University Higher School of Economics
Master's Program "Big Data Systems"

Fall 2018

Classification example



- 10000 observations
- We want to classify a new customer as defaulted (=1) or not defaulted (=0) based on his/her balance and income

Logistic regression

- Binary response:

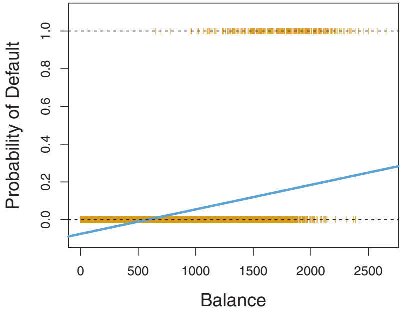
$$Y = \begin{cases} 1, & \texttt{default = Yes}, \\ 0, & \texttt{default = No}. \end{cases}$$

- Logistic regression models the **probability** of default:

$$\Pr(\texttt{default=Yes}\,|\,\texttt{balance}) \equiv p(\texttt{balance})$$

- $X = \texttt{balance}$

Logistic regression (2)



- Linear regression (bad!):

$$p(X) = \beta_0 + \beta_1 X$$

- Logistic regression (good):

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Logistic regression (3)

- Logit, or log-odds:

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- If $\beta_1 > 0$ then increase of X leads to increase of $p(X)$ and vice versa
- β_0, β_1 are estimated via maximum likelihood technique => $\hat{\beta}_0$, $\hat{\beta}_1$
- Prediction:

$$p(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

- A threshold for classification should be set. For example, $p(x) > 0.5$ => $y = 1$, i.e., default.
- Multiple logistic regression including dummy variables:

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

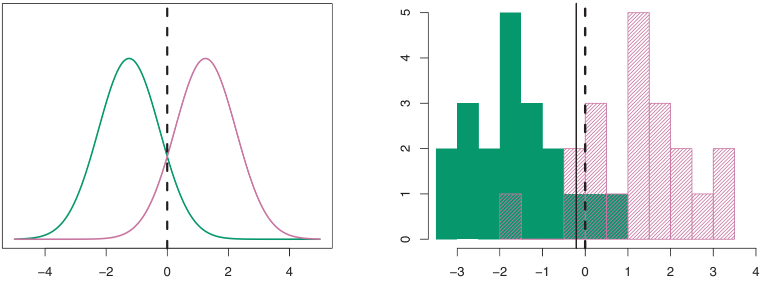
Linear discriminant analysis

- We wish to classify an observation into one of $K \geq 2$ classes
- Let π_k denote prior probability that a randomly chosen observation comes from the kth class
- Let $f_k(x) = \Pr(X = x | Y = k)$
- Bayes' theorem:

$$\Pr(Y = k | X = x) = \frac{\pi_k \Pr(X = x | Y = k)}{\sum_k \pi_k \Pr(X = x | Y = k)} \tag{1}$$

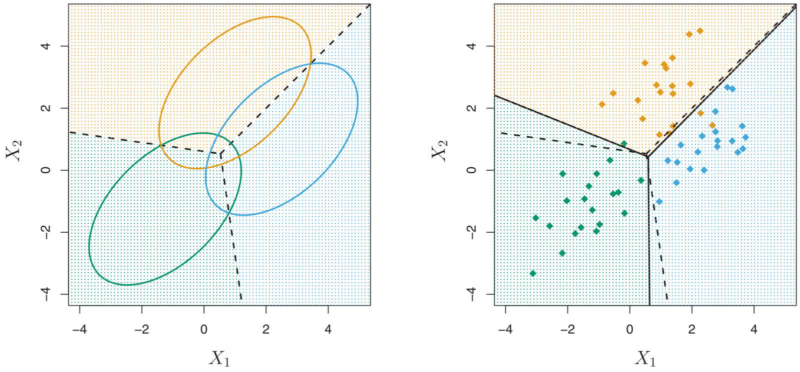
- For given predictor $X = x$, assign the observation to class k such that (1) has maximal value.
- $f_k(x)$ is often a Gaussian with parameters μ_k, σ_k.

Linear discriminant analysis (2)



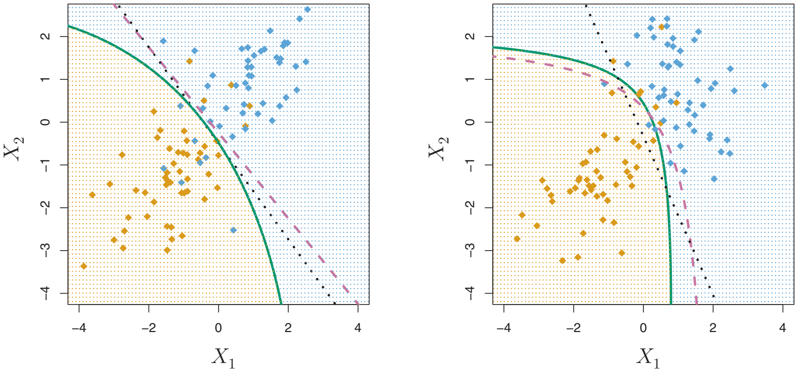
- Estimates $\hat{\pi}_i$, $\hat{\mu}_i$, $\hat{\sigma}_i$, $i = 1, 2$ are used

Linear discriminant analysis for multiple predictors



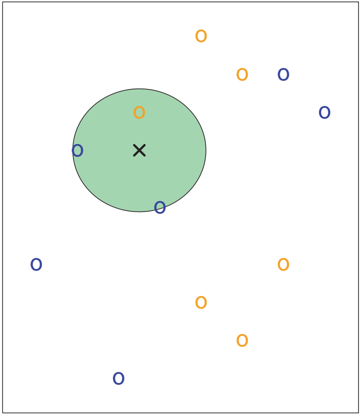
- Estimates $\hat{\pi}_i$, $\hat{\mu}_i$, $\hat{\Sigma}_i$, $i = 1, 2, 3$ are used

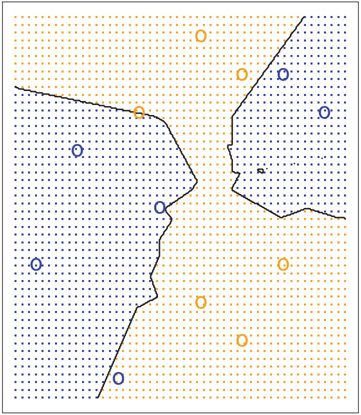
Quadratic discriminant analysis



- Unlike the LDA, no assumption $\Sigma_1 = \Sigma_2$ is used

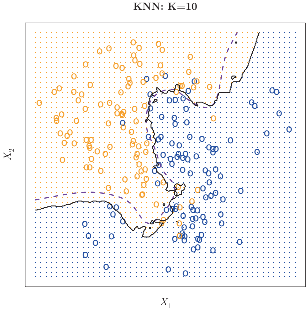
K-nearest neighbors classifier

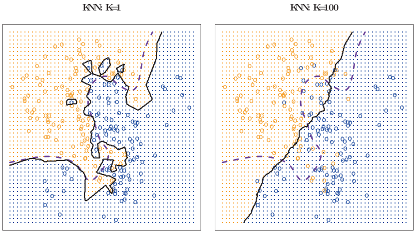




- Training set of six blue and six orange circles
- Cross denotes the observation to be classified
- $K = 3$ => consider three nearest neighbors
- Two of those are blue (2/3) => the cross belongs to the blue class

K-nearest neighbors classifier (2)





References

See Chapters 2,4 of [1] for more.

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: With Applications in R.
Springer Publishing Company, Incorporated, 2014.