

Tut 6: LDA (Bayes' Classifier)

May/June 2022

The general mathematical formulation of a generative model:

$$\begin{aligned} h_D(\mathbf{x}) &= \operatorname{argmax}_{j \in \{1, \dots, K\}} \mathbb{P}(\mathbf{X} = \mathbf{x} | Y = j) \mathbb{P}(Y = j) \\ &= \operatorname{argmax}_{j \in \{1, \dots, K\}} [\ln \mathbb{P}(\mathbf{X} = \mathbf{x} | Y = j) + \ln \mathbb{P}(Y = j)] \end{aligned} \quad (6.1)$$

QDA (only works for numeric inputs which follows the normal distribution):

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = j) \approx \frac{1}{(2\pi)^{p/2} \sqrt{|\mathbf{C}_j|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{C}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}.$$

LDA (only works for numeric inputs which follows the normal distribution):

$$\begin{aligned} \mathbb{P}(\mathbf{X} = \mathbf{x} | Y = j) &\approx \frac{1}{(2\pi)^{p/2} \sqrt{|\mathbf{C}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}. \\ \Rightarrow h_D(\mathbf{x}) &= \operatorname{argmax}_{j \in \{1, \dots, K\}} \left\{ \ln \mathbb{P}(Y = j) + \tilde{\boldsymbol{\mu}}_j^T \mathbf{C}^{-1} \left[\mathbf{x} - \frac{1}{2} \tilde{\boldsymbol{\mu}}_j \right] \right\}. \end{aligned}$$

Naive Bayes:

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = j) \approx \prod_{i=1}^p \mathbb{P}(X_i = x_i | Y = j)$$

1. Factory XYZ produces very expensive and high quality golf balls that their qualities are measured in term of curvature and diameter. Result of quality control by experts is given in the table below:

Curvature	Diameter	Result
2.95	6.63	Passed
2.53	7.79	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	4.46	Not Passed
2.16	6.22	Not Passed
3.27	3.52	Not Passed

As a consultant to the factory, you get a task to set up the criteria for automatic quality control using LDA model. Then, the manager of the factory also wants to test your criteria upon a new type of golf ball which have curvature 2.81 and diameter 5.46.

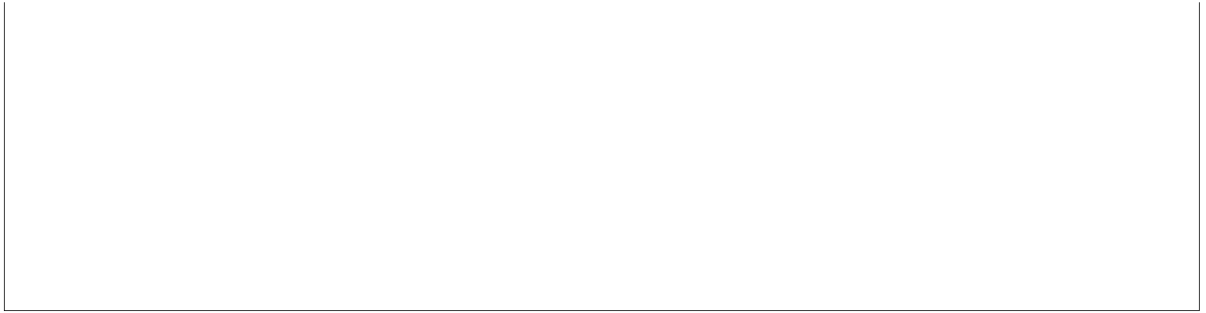
- (a) Plot the data with axes of curvature and diameter. Comment on the plot.

- (b) Write the data into matrix form by separating into “Passed” and “Not Passed”.

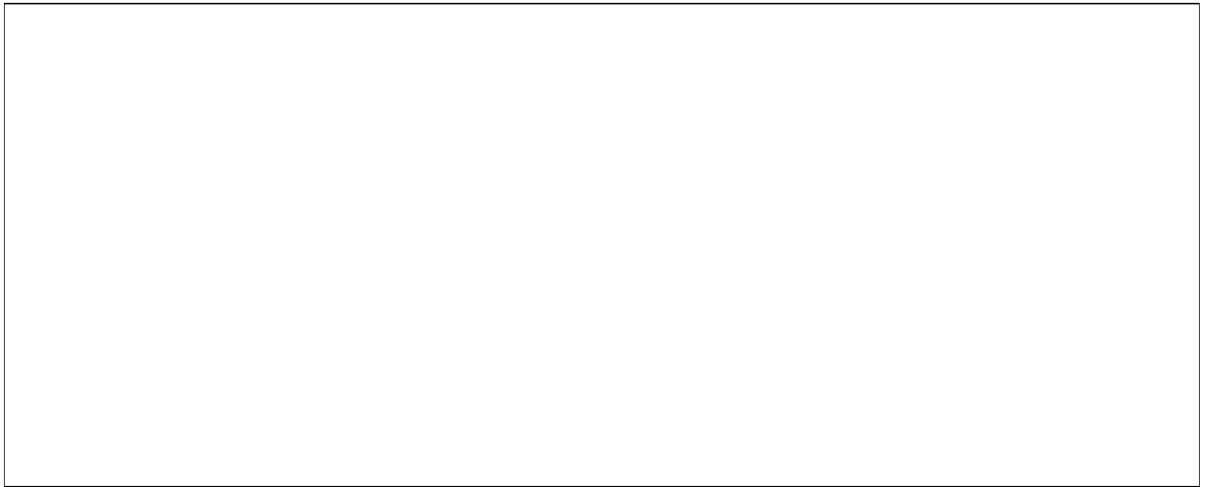
- (c) Compute the prior probability for both classes.

- (d) Compute the mean vectors for both classes.

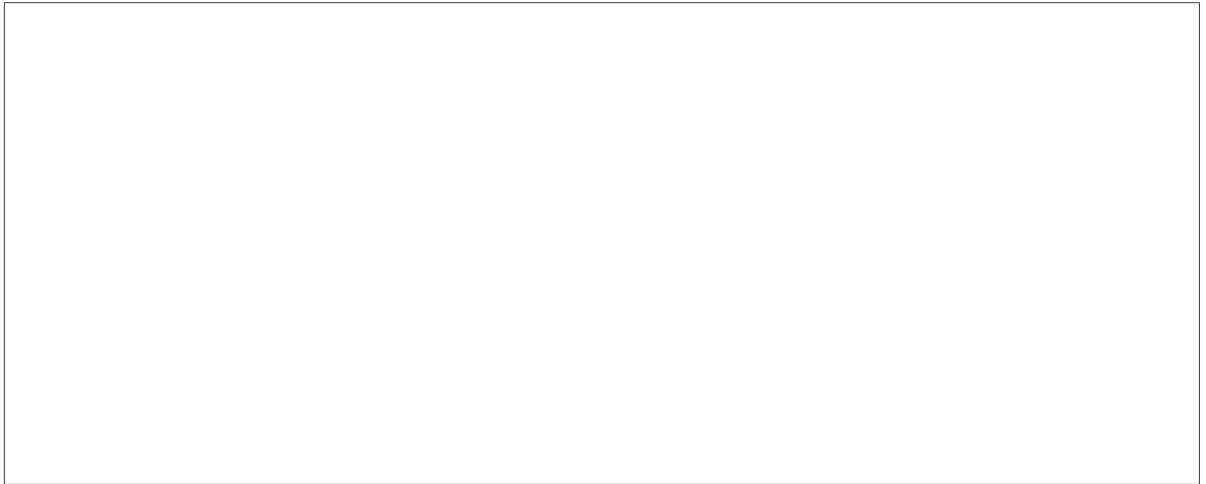
- (e) Compute the group covariance matrix.



(f) Write down the discriminant functions for both classes.



(g) Transform all the given data into discriminant functions.



(h) Locate the new golf ball in the plot as well as the functions to classify it.



(i) Plot the discriminant line into plot of $\delta_1(X)$ versus $\delta_2(X)$.

2. The data in Table 5.1 contains size measurements for two penguin species, i.e. Adelie and Gentoo, observed on three islands in the Palmer Archipelago, Antarctica.

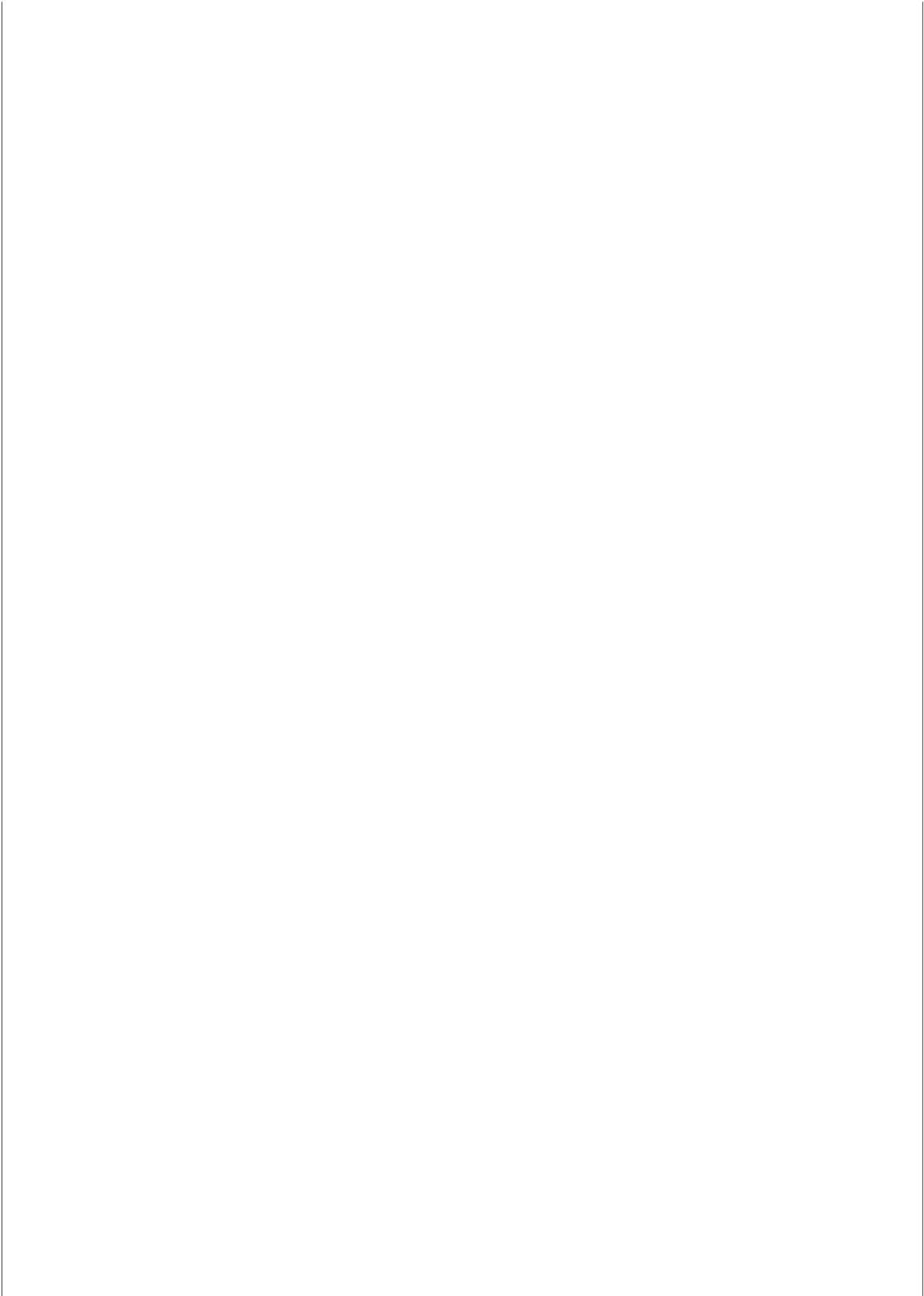
Table 5.1: Palmer Archipelago Penguin data

Flipper length (mm)	Body mass (gram)	Species
196	4400	Adelie
188	3050	Adelie
219	5250	Gentoo
193	4200	Adelie
208	4200	Gentoo
215	5000	Gentoo
197	4775	Adelie

Construct a linear discriminant analysis (LDA) model for the given data in Table 5.1 by following the following steps.

- (a) Write down the general mathematical formula of the LDA model. (2 marks)

- (b) Write down the estimates of all the parameters of discriminant functions of the LDA model in part (i). (11 marks)



- (c) Use the discriminant functions of the LDA model in part (ii) to determine the species of a penguin with a flipper length of 218 mm and a body mass of 4590 gram. (4 marks)

