MATH38161 Coursework

Korbinian Strimmer

23 November 2020, last update 25 November 2020

Overview

This coursework is about analysing a biological data set containing size and weight measurements for 333 adult penguins near Palmer Station, Antarctica. See further below for the detailed task description and the marking scheme.

You have ~~14 days~~ 21 days to complete the coursework. The intended total workload is 10 hours, to be used for:

- writing and testing R code
- running the actual data analysis
- interpreting the results
- writing the report

Submission process and deadline

- The deadline for submission is **Monday ~~7 December~~ 14 December 2020, 12 noon**.
- Submission is online on Blackboard.

Note: the deadline has been extended on 25 November to accommodate returning home from campus.

Format

- You may use any document preparation system of your choice but the final document must be a PDF in A4 format. It is highly recommended to use R Markdown.
- Your report must include the complete analysis in a reproducible way: include the full R code, figures, text etc. in one single document.
- Recommended length: 4-6 pages (single sided).
- Submissions longer than 8 pages will be be marked down for excess length.
- Put your full name and your University ID in the header/title of your report.

Copying and plagiarism

This is **individual coursework** — all text and analyses need to be done and written independently by yourself and in your own words! Write everything yourself, including computer code! However, you may freely adopt the example R code presented in the worksheets of the module. Cite and acknowledge all your sources (except the worksheets and the lecture notes).

Copying and plagiarism (=passing off someone else's work as your own) is a very serious offence and will be strictly prosecuted. There will be a zero tolerance approach to plagiarism. As a minimum you will fail the coursework and receive zero marks. As 3rd year students you will also be referred to a Faculty level scientific misconduct panel and you may be subject to additional penalties.

For more details see the "Guidance to students on plagiarism" available at http://documents.manchester.ac.uk/display.aspx?DocID=2870

A word of caution: in every single year this course has run there were instances of plagiarism. All of them were prosecuted and students convicted.

Coursework description:

Your task is to analyse an Antarctic penguin data set using unsupervised learning.

The data

The data are contained in the file "penguins.rda" available on Blackboard.

Load the data in R:

```
load("penguins.rda")
```

The data is on 333 penguins for 4 measured variables (bill length and depth, flipper length, weight):

```
dim(X.penguins)
```

```
## [1] 333   4
```

```
colnames(X.penguins)
```

```
## [1] "bill_length_mm"    "bill_depth_mm"     "flipper_length_mm"
## [4] "body_mass_g"
```

There are 165 female and 168 male penguins:

```
levels(L.sex)
```

```
## [1] "female" "male"
```

```
table(L.sex)
```

```
## L.sex
## female   male
##    165    168
```

The data set contains three species of penguins (Adelie, Chinstrap and Gentoo):

```
levels(L.species)
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

```
table(L.species)
```

```
## L.species
##    Adelie Chinstrap    Gentoo
##       146        68       119
```

They live on three islands:

```
levels(L.islands)
```

```
## [1] "Biscoe"    "Dream"     "Torgersen"
```

```
table(L.islands)
```

```
## L.islands
##    Biscoe     Dream Torgersen
##       163       123        47
```

You can find more information about the Palmer station penguin data set at the web page:

https://allisonhorst.github.io/palmerpenguins/#meet-the-palmer-penguins

and in the following publication:

K. B. Gorman, T. D. Williams, and W. R. Fraser. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081

Data analysis

Perform clustering analysis of the 333 penguins based on the 4 measured variables.

- Decide whether to analyse male and female penguins separately, or combined in one data set.
- Use PCA, hierarchical clustering, K-means and Gaussian mixture models as you see fit.
- As a minimum, use two different methods to obtain the clusterings.
- Visualise the results from the cluster analysis.
- Compare with the known clusters given by the labels L.species and L.island.

Structure of the report

The report should be structured into three sections:

1. Dataset
2. Methods
3. Results and Discussion

In Section 1 provide some background and describe the data set. In this section present summary statistics and decide whether to split the data into male and female birds. In Section 2 briefly introduce the multivariate methods you are using to analyse the data. In Section 3 run the analyses and present and interpret the results. Show all your R code so that your results are fully reproducible.

Marking scheme

The assignment will be marked out of 20 as follows:

- Description of the data: 4 marks
- Description of the methods: 4 marks
- Results and Discussion section: 8 marks
- Overall presentation of report: 4 marks

Penalties:

- If the report exceeds 8 pages (A4 PDF) you will loose 2 marks for each extra page.
- In the case of plagiarism the whole coursework is void. There will be no partial credit for non-plagiarised parts. In addition, there will be further penalities imposed by the University.