Math221 – Applied Statistics (Data Analysis)

Dr. Kamal Dingle

Project 1

Scenario: You are employed as a data scientist to help a Swiss bank with bank note forgeries (fake bank notes which look like real notes). The bank wants you to use 6 measurements from a sample of notes, which contains some real and some fake notes. The bank hopes that the fakes can be identified using just these measurements.

The six measurements are:

1: Length of the bank note,
2: Height of the bank note, measured on the left,
3: Height of the bank note, measured on the right,
4: Distance of inner frame to the lower border,
5: Distance of inner frame to the upper border,
6: Length of the diagonal.

There are 200 samples in total. The first 100 samples in the dataset are genuine, and the second 100 are fake bank notes.

You will investigate the bank notes as follows:

1) [2 marks] Download **Math221_swissbanknotes.txt** from MyGUST.

2) [3 marks] Use **np.loadtxt(…)** to load the data into Spyder.

3) [5 marks] Make a histogram of the length of all the notes (hint: Use X[:,0]). Comment on the distribution.

4) [5 marks] Make a boxplot of the distance of the inner frame to the upper border. Are there any points that look like outliers? Which point(s)?

5) [3 marks] Perform a principal component analysis (PCA) projection from the 6 dimensions to 2 dimensions. Do not plot.
(Comment: Usually, before doing a PCA projection, you should **scale the data**, by subtracting the mean and dividing by the standard deviation. In this project, we will not do this, for the sake of simplicity, and because the variables are not extremely different in terms of their scale.)

6) [2 marks] Write a sentence explaining the idea behind a PCA projection.

7) [3 marks] Print the first 5 rows of the projected data and save the projection of all the 200 samples into a .txt file, using **np.savetxt(…)**.

8) [6 marks] Make a 2D scatter plot of the PCA projected data, with different colours and symbols for the real and fake bank notes.

9) [3 marks] Are the real and fake notes different on the PCA plot? Are there any clusters in the data?

10) [5 marks] The bank gives you a new note, and asks if it is similar to the real notes, fake notes, or neither. Download your personalised note measurements from MyGUST, and try to classify your note as real, fake, or different from both groups. Explain your methods in detail.

11) [5 marks] Suggest other ways the bank could use data and statistics to reduce the production and use of fake bank notes.

12) [3 marks] Write your project neatly and clearly, and comment on the code For example `plt.figure() # makes an empty figure`

Notes:

(A) Please hand in a document with all your figures and answers to the above questions.

(B) You can use the Python example code on MyGUST.

(C) Please include the statement "**This project is all my work, and I wrote all the Python code myself**" and sign beneath it.

The Swiss note data came from Flury and Riedwyl (1988), Multivariate Statistics: A practical approach.