**Exercise Sheet 2.**

**Classification. kNN and SVMs**

**Exercise 1.**

Using the Palmer Penguin dataset (with the species as output variable, and the flipper length, bill length, bill depth and body mass as predictors/ features), compare and contrast the outputs of classification tasks, performed using a cross-validation setting. The cross-validation can be based on a 5-fold setup; repeat the CV on 5 iterations.

The classification approaches to be compared are:

[a] Discriminative approaches

[a1] Logistic regression

[a2] Nearest neighbours

[a3] Support vector machines

[b] Generative approaches

[b1] Linear discriminant analysis

[b2] Quadratic discriminant analysis

For each approach, interpret the parameters i.e. coefficient for the logistic regression, hyper-parameters for the kNN and SVM. Also present the relevant confusion matrices and comment on the stability and robustness of the approaches.

Introduce 2 outliers for at least 2 predictors (these can be randomly selected).

Rerun the generative vs discriminative approaches, and comment on the influence of outliers on the stability of the model.

Exercise 2

[a] Using the optical recognition of handwritten digits dataset from scikit-learn (https://scikit-learn.org/stable/datasets.html), assess the parameters for the SVMs. Optimise both the kernel and the C parameter.

[b] Assess the “effect” of data size (number of points). Perform the analysis on 10%, 25%, 50%, 75% and 90% of the data. Compare and contrast your conclusions.

[c] Introduce 5 random, strong outliers in the data, and compare and contrast your conclusions for the subsamples (without replacement) mentioned in [b].