MATH97131– Machine Learning
Coursework 2 — Spring 2021

Submit by 9am Monday the 26th of April 2021. Upload your final version as a PDF file with at most 15 pages (excluding the appendix). Only submit your final report - there is no option for re-uploading - and avoid last minute uploads.

Please note the following:

- Considerable emphasis will be put on clarity of expression, quality of presentation and on the depth of understanding. Ensure that your answers are well written, organised and are in the form of properly written sentences that include your full statistical reasoning. Use mathematical equations to describe your reasoning.
- Please clearly state in your essay the name of the `R` packages and functions you are using. Provide your code in appendix – do not use any code in your essay.
- Report results rounded with 4 digits.

As this is assessed work you need to work on it INDIVIDUALLY. It must be your own and unaided work. You are not allowed to discuss the assessed coursework with your fellow students or anybody else. All rules regarding academic integrity and plagiarism apply. Violations of this will be treated as an examination offence. All questions that you may have concerning the coursework must be addressed to the lecturer via e-mail (marking the e-mail as high priority). Any resulting clarifications will be communicated to the entire year via Blackboard announcements.

Question 1 – 40% of the mark

This question is concerned with the data in

`http://stats.ma.ic.ac.uk/~nadams/MSC-ML-Data/CIDXXXXX.csv`

where XXX is your CID, with leading zeros removed.

The data arises from automated karyotyping, as discussed in an early video on classification. Recall that the normal human chromosome complement consists of 46 chromosomes. Cells at an appropriate stage of cell division, in which the chromosomes are manifest, are chemically stained to induce a banding pattern on the chromosomes. A sequence of image processing steps is then conducted to identify the individual chromosomes, and extract useful measurements. These measurements, which form our feature data are based on (i) the size and shape of the chromosome, and (ii) the banding pattern induced by staining. This process results in a set of feature vectors, one per chromosome per individual, that are subsequently subjected to a normalisation procedure such that chromosomes from different individuals are comparable.

The first column of the dataset, denoted `z`, is a class indicator. Here, the problem has been framed as binary classification, corresponding to different chromosomes types. The variables, `V2`, `V3`,... `V12` relate to the size and shape of the chromosome and the remainder relate to quantitative descriptions of the banding pattern. Note that there are 28 feature variables, with the numeric indexing starting from 2. Each row of the data set refers to a single chromosome for one individual randomly sampled from a large group of test subjects.

In this coursework, you will investigate the data set using

- Multilayer perceptrons (MLP)
- Random forests (RF)

The objective is to select the classifier that has best performance at minimising error rate. Split the data set into approximately equal size pieces, to act as *training* and *test* sets. The test set should **only** be used once, for the final out-of-sample performance comparison.

Write a short report, summarising your work and addressing the following elements:

1. A basic summary of the data, noting any interesting features and reporting any useful exploratory data analysis.
2. Using the training set, build an MLP and RF. Describe your procedures for selecting hyperparameters. Explore any means with which you are familiar to improve the models.
3. Compute the predictions of the selected MLP and RF on the test data. Compare the two sets of predictions, using McNemar's test, as described below.
4. Based on the conclusion of the test, or otherwise, recommend a classifier for your data set and make note of any concerns or issues related to the recommendation.

5. In real applications, it is sometimes convenient to use a "reject option". A reject option identifies certain points as having predicted class membership probabilities that are uncertain, in a specific sense. These points are identified as "rejected" and not considered for classification. Extra data may be sought for these rejected case. A short summary description of the reject option is presented below. Implement a reject option for your selected classifier on the test data, using $t = 0.4$. Report the final confusion matrix with the rejected points removed.

McNemar's test: Quoting directly from Ripley (1996, Section 2.7).

"More appropriate methods are available, such as McNemar's test (Fleiss, 1981). Let n_A and n_B be the number of errors made by method A and not method B, and *vice versa* .... Then McNemar's test (with continuity correction) refers

$$\frac{|n_A - n_B| - 1}{\sqrt{n_A + n_B}}$$

to a $N(0, 1)$ distribution, and an exact test refers n_A to a binomial $(n_A + n_b, 1/2)$ distribution."

The null hypothesis is that the predictions of A and B are the same.

Reject option: Quoting directly from Webb (2002, Section 1.5.1).

"... we partition the sample [feature] space into two complementary regions R, a *reject region*, and A, an *acceptance* or *classification region*. These are defined by

$$\begin{aligned} R &= \left\{\mathbf{x} | 1 - \max_i p(\omega_i|\mathbf{x}) > t\right\} \\ A &= \left\{\mathbf{x} | 1 - \max_i p(\omega_i|\mathbf{x}) \leq t\right\} \end{aligned}$$

where t is a threshold. ... Thus, if a pattern x lies in the region A we classify it according to the Bayes [decision] rule for minimum error ([in text cross reference]). However, if $\mathbf{x}$ lies in the region R, we reject $\mathbf{x}$.

The approach then is identify which test set observations fall in the "acceptance" region, and evaluate performance only on this subset."

In this author's notation, $p(\omega_i|\mathbf{x})$ refers to the posterior probability of class membership for class ω_i. That is, ω_i is the label for the ith class. Note that the reference to Bayes here is merely terminology for the decision rule for minimum error.

Question 2 – 30% of the mark

1. In simple words describe Hierarchical Clustering and its dependence on the choice of different measures.

   Using the dataset of question 1, divide the features into an appropriate number of groups using Hierarchical Clustering. Discuss the effect of the different measures on the produced clustering. Using the silhouette measure, identify the best choice of measures and present your final clustering of the features.

2. Assess the performance of the K-means algorithm in dividing the observations into the classes of the "z" indicator by computing the Rand Index.

   In your answer include the following:

   - Explain the Rand Index in your own words using the appropriate mathematical equations
   - Interpret the final Rand Index of the clustering.

   Discuss any problem(s) that you have encountered while applying the K-means algorithm to the data, and present your chosen solution.

   Present the clusters of the observations in a 2D graph.

Question 3 – 30% of the mark

For this question you will use the simulated dataset provided in the CSV files `train_dataQ3.csv` and `test_dataQ3.csv` and investigate various regression models for the independent variable x and dependent variable y.

1. The training data contains 2000 examples, but you should randomly sample a subset to work with in this project. You should use your CID as a random seed to randomly sample 200 examples (without replacement), and use this subsample as the training data. This will ensure your training dataset is unique to you. The test data will be the same for all students.

2. Use the polynomial kernel $K(x_i, x_j) = (x_i x_j + 1)^p, \quad p \in \mathbb{N}$ and kernel ridge regression to construct a linear model of the form $f(x; \boldsymbol{\theta}) = \sum_{i=1}^{200} K(x, x^{(i)})\theta_i$, where $\{x^{(i)}\}_{i=1}^{200}$ are the training data inputs, and $\theta_i \in \mathbb{R}$. You should describe the procedure followed to tune hyperparameters, and provide plots that demonstrate the obtained fit to the training and test sets. Evaluate your model on the test data, and comment on your results.

3. Using a Gaussian process, construct three predictive models using different kernels for the covariance function. Compare your models quantitatively, providing details of any assumptions made. Evaluate the predictions for your best model on the test data, and provide plots to demonstrate the model on the training and test sets.

References

- Fleiss, J.L. (1981) *Statistical methods for rates and proportions.* Second edition. Wiley.
- Ripley, B.D. (1996) *Pattern recognition and neural networks.* Cambridge.
- Webb, A. (2002) *Statistical pattern recognition.* Second edition. Wiley.