MATH97079– Machine Learning
Coursework 2 — Spring 2022

Submit by 9am Monday the 14th of March 2022. Upload your final version as a PDF file with at most 15 pages (excluding the appendix).

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.
- 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 – 70% of the mark

Download your individual dataset available on Blackboard. The dataset arises from a single-cell analysis of cells in the immune system. High-content automated image analysis was used to measure dozen of descriptors for thousands of cells. The biological features include measurements associated to cell morphology, cell cycle, cell proliferation as well as the the concentration of various proteins of interest in different regions of the cell.

Your individual dataset records the value of 40 biological features for 500 cells. The first column of the dataset (Y) corresponds to the log-ratio of the concentration of two proteins of interest. The objective is to build a regression model that best predicts Y given the other biological features ($X_1, \dots X_{39}$). In the following, you will investigate the dataset using unsupervised learning and two supervised learning approaches.

Start your report with a basic summary of the data, noting any interesting features and reporting any useful exploratory data analysis.

Part A

In this part, you will focus on the biological features ($X_1, \dots X_{39}$) and ignore the variable Y.

1. Use Hierarchical Clustering to divide the biological features into an appropriate number of groups. Discuss the effect of the choice of hyperparameters on the produced clustering. Using the silhouette measure, identify the best choice of measures and present your final clustering of the features.
2. Apply the K-means algorithm to cluster the cells into two groups. Present the clusters of the observations in a 2D graph. Then, apply the kernel K-means algorithm with your choice of kernel. Compare the clusters obtained from K-means and kernel K-means in terms of the number of observations clustered similarly.

Part B

This part focuses on comparing two regression models to predict Y given $X_1, \dots X_{39}$.

1. Divide the dataset into a training and a test set. Using the training set, build a LASSO and a Random Forest regression models for this prediction task. Describe your procedures for selecting the hyperparameters. Chose one procedure of your choice to evaluate the variable importance of the random forest regression model and compare the result to the variables selected by LASSO. Compare the performances of the inferred LASSO and Random Forest regression models in terms of prediction on the test dataset. Explore any means with which you are familiar to improve the models. Recommend a model for the prediction task and make note of any concerns or issues related to the recommendation.
2. The objective of this question is to evaluate how the conclusions from the previous question (B.1) regarding the two regression models - both in terms of prediction performances and variable importance – vary depending on the training/test split of the dataset. Precisely describe an appropriate cross-validation procedure for this task and discuss your results. You will need to use the so-called *nested cross-validation* procedure for LASSO.
To lower computational costs, you might want to limit the range of values in the hyperparameter search here.

Question 2 – **30% of the mark**

For this question you will use the simulated dataset provided in the CSV file `dataQ2.csv` containing the water temperature (`temp`) measured at different times (`t`) and at various distances (`d`) from the coast. The time is measured in days since the beginning of the study and the distances are measured in kilometres.

1. Suppose that we are interested in modelling the water temperature near the coast (i.e. when $d = 0$) as a function of time. Using a Gaussian process, construct three predictive models using different kernels for the covariance function. Precisely describe your choice of kernel hyperparameters. Provide plots to demonstrate the fit of the three models to the dataset. Compare your models quantitatively in terms of how well they describe the data. Use your best model to predict the water temperature near the coast at time $t = 35$ days. How likely is it to measure a temperature of 13 degrees that day?

2. Propose a Gaussian Process regression model to predict the water temperature as a function of both time (`t`) and the distance to the coast (`d`). Justify your choice of kernel. Provide plots to demonstrate the fit of the model to the dataset. Use the fitted model to predict the water temperature at day 55 as a function of d.