MATH97131– Machine Learning
Coursework 1 — Spring 2021

Submit by 5pm Wednesday the 10th of February 2021. Upload your final version as a PDF file with at most 10 pages (excluding the appendix). The report should not include your name anywhere (to allow for anonymous marking). 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.
- 10 marks will be allocated to the quality of the presentation including the quality of the figures.
- Provide your code in appendix – do not use any code in your essay.
- Report results rounded with 4 digits.

Question 1 – 35 marks

This question should be solved without using any statistical or machine learning packages.

Please state your CID at the beginning of the report. Download your individual dataset $\mathcal{D} = \{X_{1,i}, X_{2,i}, Y_i\}_{i=1}^{100}$ at

`http://wwwf.imperial.ac.uk/~sfilippi/DataQ1/XXXXXXXX.csv`

where XXXXXXXX designates your 8 digits CID. Consider the following model

$$Y_i = c + \sum_{j=1}^{2} \left(a_j \cos(X_{j,i}) + \sum_{p=1}^{4} b_{j,p} X_{j,i}^p \right) + \epsilon_i$$

where ϵ_i are independent random variables such that $\mathbb{E}(\epsilon_i) = 0$.

1. Infer the parameters of the model using ridge regression. Precisely describe the fitting procedures including the choice of the regularization parameter. Report the inferred parameters. Predict the value of Y for the following four observations using the inferred model:

X_1	-2.1	2.9	-0.9	3.8
X_2	4.4	-4.5	0.3	3.9

2. Suppose now that $\epsilon \sim \mathcal{N}(0, 4)$. Assuming that the parameters are *a priori* independent and considering a Gaussian prior distribution with mean 0 and variance α for each parameter, perform a Bayesian inference of the model parameters. Provide the equation of the joint posterior distribution of the 11 model parameters. Plot the posterior mean as a function of α. Discuss the results. Compare the posterior mean obtained to the parameter inferred in the previous question. Report the posterior predictive distribution for the four observations in the previous table for $\alpha = 1$.

Question 2 – 55 marks

If you chose to use a R statistical package, please clearly state the name of the package and function.

Consider a binary classification problem where, under class $\mathcal{C}_1$, X is uniformly distributed on the p-dimensional unit cube $[0, 1]^p$ while, under class $\mathcal{C}_2$, X is uniformly distributed on the p-dimensional cube $[0, 1 + \epsilon]^p$ with $\epsilon > 0$. Assume equal prior probabilities for the two classes.

1. (a) Describe the rule followed by the Bayes classifier for this example.
   (b) Compute the Bayes error rate for the Bayes classifier as a function of p. Plot this for $\epsilon = 0.1$, 1 and 2. Discuss the results and comment on the scaling of this specific classification problem with p.
2. Setting the seed of the random generator equal to your CID, generate a random dataset of $n = 1000$ observations for $p = 4$ and $\epsilon = 0.5$. This dataset will be used for training, validation and testing.
   (a) Using cross validation, compute the error rate, with standard error, of both the logistic regression and QDA classifiers. Precisely describe the cross validation procedure in this context (with mathematical equations and/or pseudo-code).

(b) Using the so-called *nested cross-validation*, compute the error rate, with standard error, of the k nearest neighbours classifier. Precisely describe the nested cross validation procedure in this context (with mathematical equations and/or pseudo-code).

(c) Given the three estimates of performance, which of these three classifiers (logistic, QDA and k nearest neighbours) would you recommend? Why? You might want to consider the performance, their variabilities as well as computational costs.

(d) Use the cross-validation procedure to compute the error rate, with standard error, of the Bayes classifier. Comment on this result in comparison to the error rates of the three other classifiers as well as to the Bayes error rate derived in question 1b.

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. In particular, letting somebody else copy your work constitutes 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.