MATH97131– Machine Learning
Coursework 2 — Spring 2024

Due Wednesday the 15th of May at 1pm

The final report must be typed up and should be a properly structured document in PDF, with a suggested length of **approximately 15 pages** although there is no strict page limit. Your report should be uploaded to Turnitin by the deadline stated above. Once the report is uploaded there is no option for re-uploading so you should **upload your final version only**. If possible, avoid last minute uploads as the system can crash if it simultaneously receives too many requests.

Submit your assessments with your CID and do not include your name anywhere on your submission. Please add a good academic practice statement to your coursework such as:

I, [insert CID], *certify that this assessed coursework is my own work, unless otherwise acknowledged, and includes no plagiarism. I have not discussed my coursework with anyone else except when seeking clarification with the module lecturer via email or on MS Teams. I have not shared any code underlying my coursework with anyone else prior to submission.*

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 course via Blackboard announcements.

Please note the following:

- Considerable emphasis will be put on **clarity of expression and a clean presentation**. Only detailed, well-written answers that clearly explain your reasoning will score highly. In particular, avoid using the code in place of written sentences and mathematics. The quality and suitability of the chosen methods will be taken into account in the distribution of marks in all questions.
- Marks will be allocated to the quality of the presentation including the quality of the figures.
- Your report should clearly describe any statistical assumptions or choices you are making. The report should contain all the necessary information for a statistician to be able to reproduce your analysis without looking at your code.
- Any figures or tables should be included in your report for a good reason, and this reason should be described in your report – all tables and figures should be referenced in the main body of your text. All axes should be appropriately labelled.
- Please clearly state in your report the name of the `Python` (or `R`) packages and functions you are using. Provide your commented code in appendix – do not use any code in your essay.

Question 1

In this question you will consider a spectroscopy dataset coming from the chemometric literature. Spectroscopy is the study of how light interacts with different materials at different wavelengths. By measuring the amount of absorption and transmission of electromagnetic waves through a material, it is possible to study the materials' properties. Here you will investigate the impact of temperature on the spectra of chemical samples.

The dataset provided on Blackboard consists of the spectra of 108 chemical samples. Each chemical sample consists of a mixture of three components: water, ethanol and 2-propanol. The dataset contains the following information:

- The relative concentrations of each component (water, ethanol, 2-propanol) in sample $i = 1, \dots 108$, also called the *composition* of sample i and denoted by $w_i = (w_{i,1}, w_{i,2}, w_{i,3})$, are recorded in the file `components.csv`. Note that w_i belongs to the 2-simplex, i.e. $w_{i,j} \geq 0$ for $j = 1, 2, 3$ and $\sum_{j=1}^{3} w_{i,j} = 1$.
- The temperature t_i of each chemical sample i is recorded in the file `temperatures.csv`.
- The file `spectra.csv` contains the spectra of the chemical samples. The spectra of sample i is a 199-dimensional vector $y_i = \{y_i(\lambda_j)\}_{j=1}^{199}$ of the recorded amount of absorption for a grid of wavelengths $\lambda_j \in \{800, 801, 802, \dots 998\}$.

Part 1 – 20% of total mark

1. *[5 marks]* Describe the dataset, providing some exploratory plots.
2. *[8 marks]* Use Hierarchical Clustering on the spectra to divide the chemical samples into groups. Do not use any information related to the temperature or the composition of the chemical samples during the clustering procedure.

   Discuss the effect of the choice of hyperparameters on the produced dendrograms and clustering. Discuss whether the produced clusters are related to the temperature or the composition of the chemical samples.
3. *[7 marks]* Perform a principal component analysis of the spectra and identify the smallest number (n_0) of principal components needed to explain 99.99% of the variance.

   Project the spectra data into the space spanned by these n_0 principal components and use Hierarchical Clustering on these projected spectra to divide the chemical samples into groups. Compare the produced dendograms and clustering to the one obtained in the previous question

Part 2 – 40% of total mark

In this second part, you will consider a subset of this dataset containing noisy measurements of the spectra with a large proportion of missing values.

1. *[5 marks]* Generate your individual dataset by following the instructions below:

   - Start by setting the seed of the random generator using your CID number and ensure that the sampling procedures below are run using this seed.
   - Randomly sample 40 chemical samples (among the 108). Denote by $\mathcal{I}$ the set of indexes of these 40 samples.
   - Consider the spectra measurements $\{y_i\}_{i \in \mathcal{I}}$ of these 40 chemical samples.
     - Add randomly generated noise to all these measurements. The noise should be sampled from a normal distribution with mean 0 and standard deviation 0.05.
     - Randomly replace 80% of the 40 × 199 measurements with missing values. I suggest you replace them with `NaN`.

   Denote by $\tilde{y}_i$ the noisy spectra with missing values for the i-th chemical sample.

   Your new individual dataset $\mathcal{D} = \{t_i, w_i, \tilde{y}_i\}_{i \in \mathcal{I}}$ consists of the temperature and component information for the 40 chemical samples along with their noisy spectra measurements, containing many missing values. Ensure that your code in the appendix clearly shows how your individual dataset was produced.

   Tip: You might want to save your dataset into a csv file. You might also want to produce some exploratory plots of your new dataset.
2. *[15 marks]* Select one of the 40 chemical samples of your choice in your dataset; denote by i^* the index of this sample. Using a Gaussian process, construct a predictive model for the amount of absorption (y) in this sample i^* as a function of the wavelength λ. Precisely describe the model with equations and your procedure to choose the kernel hyperparameters. Provide plots to demonstrate the fit of the constructed predictive model to the observed spectra in y_{i^*}.

3. *[13 marks]* Consider now the entire dataset $\mathcal{D}$. Propose a Gaussian Process to model the amount of absorption (y) as a function of the wavelength (λ) as well as the temperature (t) and the composition of the sample (w). Precisely describe the model with equations and your procedure to choose the kernel hyperparameters.

   Demonstrate the fit of the constructed predictive model to the observed spectra in y_{i^*} of the i^*-th sample as a function of λ given the known temperature t_{i^*} and the sample composition w_{i^*}. Compare to the fit obtained in question 1.2.2.

4. *[7 marks]* We are now interested in using the Gaussian Process that models the amount of absorption (y) as a function of λ, t, and w to infer the temperature of the sample composition for a new spectra. For this, you will use a test dataset recorded in the files `spectra_test.csv`, `components_test.csv` and `temperatures_test.csv` available on Blackboard. These files contain the spectra of two chemical samples along with their temperatures and their composition (i.e. relative concentration of water, ethanol and 2-propanol)

   (a) Consider the first observation in this test dataset for which the spectra is recorded at every 199 wavelengths and the sample composition is known but the temperature is missing. Define a grid of temperatures and plot the negative log predictive density (NLPD) for this observation for each values of the temperature in the grid. What can you say about the temperature for this chemical sample?

   (b) Consider the second observation in the test dataset for which the spectra is recorded at every 199 wavelengths and the temperature is known but the sample composition is missing. Using the negative log predictive likelihood, estimate and comment on the likely composition of the sample.

Question 2– 40% of total 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 concentration of various proteins of interest in different regions of the cell.

Your individual dataset records the value of 50 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 ($X_1, \dots X_{49}$). You will be allowed to use any of the supervised and unsupervised learning techniques covered during this module.

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

1. *[5 marks]* A basic summary of the data, noting any interesting features and reporting any useful exploratory data analysis.

2. *[21 marks]* Appropriately divide your dataset into a training and test set. Using the breadth of the approaches covered in this module (except Gaussian Processes), propose four distinct machine learning models to predict the output of interest Y given the other biological features. Precisely describe your procedures for selecting any hyperparameters and justify any statistical choices you are making.

3. *[8 marks]* Compute the predictions of the four models on the test set.

4. *[6 marks]* Recommend a model for the prediction task and make note of any concerns or issues related to the recommendation.