CCE2502: Pattern Recognition and Machine Learning - Assignment I:
Submission deadline: 14/04/2025

IMPORTANT Notes:

This assignment contains six tasks and the maximum score is 100.

The assignment contributes to 30% of the final study unit mark.

Submit the answers in a jupyter notebook (add explanations, discussion and diagrams in markdown cells within the jupyter notebook itself).

Cite any blogs, code repositories and/or generative AI tools (e.g chatgpt) used in completing this assignment. In the case of generative AI tools, explain how these tools were used, i.e submit a document with all the prompts and answers generated by the GenAI tools.

This work is to be attempted individually. It is essential that the work you eventually submit and present for your assignment consists only of your own work; use of copied material (including material obtained from Generative AI tools) will be treated as plagiarism.

Discussion is only permitted on general issues, and it is absolutely forbidden to discuss specific details with anyone and/or share results.

Please sign the plagiarism form that can be found here:
https://www.um.edu.mt/ict/students/formsguidelines/

and submit the jupyter notebook (with all cells executed) zipped together with the signed plagiarism form.

Aim of assignment

The main aim of this assignment is to study the time complexity of the brute search and kdtree methods as used in the k-Nearest-Neighbour (kNN) classification algorithm. The experiments will be carried out in Python within a Jupyter Notebook. The sci-kit learn package should be used for implementing the kNN models;

https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

Task 1:

Generate the following synthetic binary classification datasets using the sklearn.datasets. For both datasets generate scatter plots, the distribution of the output labels and the mean and standard deviation of all input variables (features).

a) Generate a dataset from make_circles with the following options; n_samples=1000, random_state=0, noise=0.2, factor =0.6, and inputs=2 **[5 marks]**

b) Generate a dataset from make_blobs with the following options (n_samples=1000, centers=2, cluster_std=1.0, n_features=2, random_state=0) **[5 marks]**

Task 2:

Develop the following functions in Python from scratch:

a) A function that re-shuffles and splits the dataset into two portions **[5 marks]**

b) A function that computes and returns the classification accuracy, recall, precision and F1-score for a binary classification task. **[15 marks]**

You are expected to develop unit test functions that adequately test the functions developed.

Task 3:

Develop a k-NN model for each dataset in task 1:

a) Using the function developed in task 2, split the dataset into a train set and test set in the ratio 80%/20% respectively. **[3 marks]**

b) Find the best k that maximises the F1-score (use the function developed in task 2 to compute F1 and you may want to split the train set further into train/validation) **[6 marks]**

c) Find the best k that maximises accuracy (use function developed in task 2 to compute accuracy and you may want to split the train set further into train/validation) **[6 marks]**

Task 4:

In this task you will empirically carry out a time complexity analysis of both the standard brute search and the kdTree search methods. Both algorithms are parameter values in the sci-kit learn kNN class. For this study you will use the synthetically generated datasets from task 1 with the option of varying the dataset size.

a) Design and describe a computational experiment to compare the time complexity of the two algorithms during inference (prediction). In other words, you have to measure the computational time taken for the models to infer the category for a fixed number of test samples given various training set sizes, for example, size = 100, 1000, 10000, etc. **[20 marks]**

b) Implement the experiment in python and extract the results. **[20 marks]**

c) Discuss the outcome of the experiment. Do the results corroborate theoretical analysis? **[15 marks]**