جـــامـــعـــة مـحـمـد بـــن زايـــد
للـــذكـــاء الاصـــطـــنـــاعـــي
MOHAMED BIN ZAYED UNIVERSITY
OF ARTIFICIAL INTELLIGENCE

AI701: Foundations of Artificial Intelligence
Fall 2023

Assignment-02

Instructions:

- Group Assignment. Maximum number of students per group: 3.
- This assignment has three sections and carries 30 marks.
- The deadline to submit the assignment is by the end of November 10, 2023 (23:59 UAE time).
- Assignment deliverables: three completed jupyter notebooks and a report. The report is expected to be within 4 pages excluding references. All the required material should be zipped in one folder (per group). The zipped folder should be named according to the group numbers (e.g Group_01.zip). Please inform the TA if you wish to change your group.

---

1 Classification (10 points)

The target of this task is to predict the risk of credit default. The dataset is provided in 'default.xls'. It contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan. Information on all the included variables is available here.
Task 1.1: Create a feature matrix X and a target vector y from data. The target to be predicted is given in the last column.
In the following, you are required to evaluate your models with a cross-validation score. Make sure your splits are fixed across the experiments.
Task 1.2: Fit a Decision Tree classifier and evaluate.
Task 1.3: Transform categorical columns into one-hot encoded features. Fit and test a decision tree classifier with the updated features.
Task 1.4: Scale and apply PCA to the suitable payment features. Fit and test a decision tree classifier with the updated features.
Task 1.5: Use grid search to find the PCA ratio that gives the best accuracy.
Task 1.6: Fit a Random Forest classifier to your data with the updated features and compare to the decision tree classifier.
Task 1.7: Use grid search to find best random forest classifier parameters min_samples_split and max_depth.
Can you suggest other feature pre-processing methods that would further improve the result?

2 Clustering (10 points)

A list of paper titles is provided in 'paper_labels.tsv', The target of this task is to cluster papers using k-means.
Task 2.1: Read the text file and parse the paper titles. The titles appear in the second column before the comma delimiter.
Task 2.2: Use TfidfVectorizer to extract text features.
Task 2.3: Apply K-means clustering to the text features.
Task 2.4: Use cluster centers and vectorizer to interpret the output clusters. Visualize the output to propose an adequate number of clusters.

3 Neural Network for Regression (10 points)

In this part, you are required to apply a Neural Network on a non-linear regression problem.
Data: The training data can be found in Reg_Train.txt and the testing data in Reg_Test.txt. Each example is described by 18 features in one line. The last column (the last value of each line) is the target value.
Task: Learn a Neural Network from the training data and use it to predict the target values in testing data. What is the best prediction result? Give the average and standard deviation of prediction errors on the test data.