THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331:Introduction to Data Mining
Fall 2019 Assignment 1
Due time and date: 11:59pm, Oct 26 (Sat), 2019.

IMPORTANT NOTES

- **Your grade will be based on the correctness, efficiency and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**

Q1. Implement the following procedure on the Iris dataset (`https://github.com/jiaxinxie97/COMP4331/tree/master/assign1/Q1`). The dataset has 4 input features, and the output label is the type of flower.

1. Show the boxplot for each feature. If a sample has an outlying feature, remove that sample.
   (a) Save the processed dataset to Q1_1.csv, using the same format as the given cvs file.
   (b) Include the boxplots of the original and processed datasets into the report.
2. Implement z-score normalization, and use it to normalize the dataset obtained in step 1. Save the normalized dataset to Q1_2.csv.
3. We want to see if any two input features are highly related. Explain how this can be done by using a numerical measure we discussed in class, and show the corresponding plots.
4. Using sklearn, train a decision tree (with the gini ratio as splitting criterion) using all the dataset obtained in step 2. To avoid overfitting, limit the maximum tree depth to 2.
   (a) You can use the class sklearn.tree.DecisionTreeClassifier (`https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`).
   (b) Show the resultant tree, using the function sklearn.tree.export_graphviz (`https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_graphviz.html`).
5. Instead of fixing the maximum tree depth to 2, we now use cross-validation to find this value.
   (a) Let the candidate maximum depth values be $\{1, 2, 3, 4\}$. Report the 10-fold cross-validated accuracies for each candidate maximum depth value, and then find the best maximum depth. You can use the function sklearn.model_selection.cross_val_score(`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html`).

Q2. You are given the following dataset, with four inputs

1. "chills": taking values "Y"/"N";
2. "runny nose": taking values "Y"/"N";
3. "headache": taking values "No", "Mild", and "Strong";
4. "fever": taking values "Y"/"N",

and output label "flu". We want to predict if a patient with "(chills=Y, runny nose=N, headache=No, fever=Y)" have flu or not? Show clearly how to obtain the prediction using the naive Bayes classifier, both with and without using Laplacian correction.

chills	runny nose	headache	fever	flu
Y	N	Mild	Y	N
Y	Y	Mild	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Q3. You are given the MNIST dataset (`https://github.com/jiaxinxie97/COMP4331/tree/master/assign1/Q3`). The first column is the class label. The other columns are the intensity values for each individual pixel in each MNIST image. Note that the feature dimensionality is 784. Also, this dataset has been split into a training set and a test set.

In this question, you have to implement in python:

- the 1-nearest neighbor classifier, using the Euclidean distance as distance measure. Note that you need to implement it from scratch, directly calling a function is not allowed.
- PCA, using the class sklearn.decomposition.PCA (`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`).

Then, perform the following steps:

1. Run your 1-nearest neighbor classifier, and report the test set accuracy.
2. (Use PCA to reduce the feature dimensionality) For each PCA dimension d in $\{1, 50, 100, 200, 300\}$, perform PCA to extract d features, and then go back to step 1.
    - On the test data, you should use the same PCA transform as obtained on the training set, which can be retrieved from your_pca_object.components_, your_pca_object in the object created by sklearn.decomposition.PCA.
3. What trend do you observe?

Submission Guidelines Package your code (named Q1.py and Q3.py), your report and the output csv files to a zip file. Name the zip file "COMP4331_Assignment_1_yourstudentID". Email the whole package to jxieax@connect.ust.hk before the deadline.

Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. Plagiarism will lead to zero point on this assignment.