COSC 311 Homework 3 (10 points)

Please finish the following tasks and submit your homework report via MyClasses. Your submission must contain your source code file (".py" or ".ipynb" file) and a PDF document. The PDF contains the results (may use screenshot) of your program.

1. Regression on the Computer Hardware Dataset

Please download the "Computer Hardware" dataset (https://archive.ics.uci.edu/dataset/29/computer+hardware). Unzip it and use the "machine.data" to conduct the following regression tasks. Assume the 3-8 columns of this dataset are attributes, and the last column ("ERP") is the ground truth.

- Task 1: Measure the correlation between each attribute and the "ERP" and select the most important 4 attributes for the following regression analysis;
- Task 2: Using the above 4 attributes and the "ERP" (ground truth) as the dataset, split it into two parts, 60% for training and 40% for testing.
- Task 3: Using training data to build a multiple linear regression model, and evaluate it using the testing data. Show the MAE, MSE, and RMSE.

2. Clustering on Hand-Written Digits

Please use the UCI ML hand-written digits dataset in our lecture note "COSC311_Module5_3_Kmeans clustering", which is included in the scikit-learn datasets.

- Task 1: Conduct PCA analysis on the dataset and find out how many principal components are needed to keep at least 85% variance (i.e. the ratio of variance loss, η, is less than 15%).
- Task 2: Assume m principal components are needed to keep at least 85% variance, transform the dataset from 64 dimensions to m dimensions.
- Task 3: Based on the above dimension-reduced dataset, conduct k-means clustering (k = 10, each cluster is a digit), output the center of each cluster.
- Task 4: Match each learned cluster label with the true label (i.e. ground truth) using mode function in scipy.stats package (i.e. based on most common value), calculate and output the clustering accuracy, and show the corresponding confusion matrix as a figure.

Policy

1. Each student MUST finish this homework independently. NO TEAM WORK and DISCUSSION are allowed. If you need any help, please feel free to contact the instructor.
2. You need to submit your source code (“.py” or “.ipynb” file) to MyClasses together with your PDF report.