COSC 311 Project 2 (15 points)

Please finish the following tasks and submit your project report via MyClasses. Your submission must contain your source code file (one Jupyter Notebook file for all code) and a PDF document. For each task, please include your explanation and/or output/results of your program (may use screenshot).

1. Clustering for Wireless Indoor Localization (5 points)

1) Dataset

Please use the "Wireless Indoor Localization Data Set". The description of this dataset can be found at https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization. You may download the dataset from this website or from the project 2 assignment on MyClasses.

2) Tasks

- Take the first 7 columns of this dataset as attributes to conduct k-means clustering (k = 4, each cluster is a room), output the center of each cluster
- Use the last column (room number) as ground truth to match each cluster with its label (room number), calculate and output the clustering accuracy (i.e. out of all samples, how many samples are correctly identified), and show the corresponding confusion matrix as a figure
- Calculate and output the clustering accuracy of each room (i.e. out of all samples for EACH room, how many samples are correctly clustered)

2. PCA based hand-written digits classification (5 points)

1) Dataset

Please use the UCI ML hand-written digits dataset in our lecture note "COSC311_Module5_4_Kmeans clustering", which is included in the scikit-learn library.

2) Tasks

- Conduct PCA analysis on the dataset and find out how many principal components are needed to keep at least 90% variance (i.e. the ratio of variance loss, η, is less than 10%).
- Assume m principal components are needed to keep at least 90% variance, transform the dataset from 64 dimensions to m dimensions.

- Based on the above dimension-reduced dataset, build a classification model (any algorithm) with optimized parameters to do cross-validation test (CVT) (fold = 10), show the CVT accuracy and corresponding confusion matrix (in a figure).

3. Regression model for median house value prediction (5 points)

1) Dataset

Please use the "Housing Dataset" attached. You may download the dataset from the project 2 assignment on MyClasses.

2) Tasks

- Assume the "median_house_value" is related to the following attributes in this dataset: "housing_median_age", "total_rooms", "total_bedrooms", "population", and "median_income", use correlation coefficient analysis to select 3 attributes that have higher correlation with the target variable (i.e. "median_house_value").
- Randomly split all the samples (each sample include the 3 selected attributes and one target variable) into two parts: 60% for training and 40% for testing.
- Use training data to build a Multiple Linear Regression model and test it using the testing data. Show the performance of the regression model, including MAE, MSE, and RMSE.

Policy

1. Each student MUST finish this project independently. NO TEAM WORK and DISCUSSION are allowed. If you need any help, please feel free to contact the instructor.
2. You need to write your code in a jupyter notebook file and save your source code and outputs. This jupyter notebook file will be submitted to MyClasses together with your PDF report.