Project 2 of FTE 4560 Basic Machine Learning
Financial data mining

Jicong Fan, CUHK-Shenzhen, April. 2021

In this project, there are two datasets. The first one is for unsupervised learning. The second one is for regression. You need to write a report containing tables or/and figures, implementation details of your algorithms (e.g. NN structure and hyper-parameter settings), and discussion about the machine learning methods and real problems. You can use any toolboxes of machine learning such as sk-learn and Tensorflow.

You can finish the project in your original groups for Project 1 or re-form new groups of size at most 5. The deadline is May 03 (23:59). There will be a presentation session on May 04 (specific time will be determined later). In the presentation session (6 minutes for presentation and 2 minutes for Q&A), every group should let two members present the work on the two datasets respectively.

1 Unsupervised learning on the bankruptcy data

1.1 Dataset description

The Polish companies bankruptcy data[1] (a subset of the original data) is about bankruptcy prediction of Polish companies. We have already used the dataset for classification in our first project. In this project, you need to **put the training data and test data together** to form a larger dataset for the unsupervised learning problems. In addition, for unsupervised learning, remember to remove the label values (last column) from the data.

1.2 Task Description

1.2.1 Clustering

Perform k-means and spectral clustering.

In spectral clustering, use $\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)\right)$ to construct the similarity matrix and use the normalized symmetric Laplacian matrix for clustering. Draw a curve to show the influence of σ (with different values) on the clustering accuracy[2] (average of a number of repeated trials). Run k-means and spectral clustering for 10 times and compute the mean values and standard deviations of the clustering accuracy. Note that you need to re-match your cluster results with the true labels according to the sizes of the clusters. Otherwise, your clustering accuracy may be very low.

1.2.2 Data visualization

Use PCA, KPCA, LLE, and t-SNE to reduce the dimension of the data to 2 and visualize the data. In the 2-D visualization, mark the data points in different classes by different colors or symbols. Analyze the performance of the four methods.

1.2.3 Missing data imputation

Use matrix factorization method to impute the missing values of the dataset. Then perform k-means clustering or spectral clustering again to show whether the missing data imputation based on matrix factorization can improve the clustering accuracy.

[1] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[2] $acc = \dfrac{\text{number of correctly classified data}}{\text{total number of the data}}$

Randomly remove 30% $\sim$ 50% of the observed values to form a very incomplete dataset (matrix), fill the missing values by the mean values, and then perform k-means or spectral clustering. Use matrix factorization to recover the missing values (30% $\sim$ 50% newly removed+the original missing values) and then perform k-means or spectral clustering. See whether the missing data imputation based on matrix factorization can improve the clustering accuracy.

2 Data analysis for China stock market

We consider the two indices "000001.SS" and "399001.SZ" of China stock market from 2019.04.08 to 2021.04.02. There are five attributes "Open price", "High price", "Low-price", "Close price", and "Volume". The datasets are in the excel files. You can stack the two indices together to form a dataset with 10 variables. You may need to normalize the data because the value of "Volume" is very different from the price value.

2.1 Task description

Perform ICA (the Fast ICA algorithm) on the data (10 variables) and plot the sources. Try to reconstruct the data using a few number (e.g. 3 or 5) of sources.

Make prediction for 5-day/15-day "Close price" of "399001.SZ". Use the first 434 samples as testing data. The remained data are the testing data. You need to choose two from MLP (fully connected feedforward NN), vanilla RNN, LSTM, and CNN. Compare the performance of the two methods and analyze the influence of NN structures and hyper-parameter settings on the prediction accuracy. Try to improve your prediction accuracy by hyper-parameter tuning, data preprocessing, or modifying the model structures if possible. This task is independent from the ICA task.