CS-345/M45 Lab Class 3

Release date: 29/10/2019
Due date: 13/11/2019

This lab is about utilizing Linear Regression for continuous value prediction, and Principal Component Analysis and Linear Discriminant Analysis for feature dimension reduction. We will be implementing these algorithms on some example data by adding our own code to a skeleton jupyter notebook.

☐ Task 3.1 – Linear Regression

The first task involves using Linear Regression to fit a predictive model to a collection of observed datapoints. You are provided with the data in the form of two numpy arrays, `x_points.npy` and `y_observations.npy`, on BlackBoard. Download the files to your local directory.

- Load the data and visualise the data with a scatter plot.
- Go through Task 3.1 in the notebook, completing the TODOs.
- You should be able to show a visualisation of the model updating in each iteration, and provide the final predicted values for the intercept (w0) and coefficient (w1) of the linear model.

☐ Task 3.2 – Principal Component Analysis

The second task is to use Principal Component Analysis to reduce the dimensionality of the Wine Dataset. A lot of the functionality is provided, your main tasks are to divide the dataset into a training and testing set, mean centre the data for Singular Value Decomposition, and project the data into the new principal component space. You are provided with the data in the form of two numpy arrays, `wineData.npy` and `wineLabels.npy`, on BlackBoard. Download the files to your local directory.

- Load the data and plot them to get a feel for the data.
- Go through Task 3.2 in the notebook, completing the TODOs.
- Divide the data into two sets: training and testing.
- Mean-centre the data and compute the SVD.
- Select a number of principal components to keep, based on the explained variance of the new component space.
- Project the two sets into the new space and visualize them.

☐ Task 3.3 – Linear Discriminant Analysis

The third task is to use Linear Discriminant Analysis to reduce the dimensionality of the Wine Dataset. Again, a lot of the functionality is already provided. Follow through the TODOs, your main tasks are to load in the Wine Dataset, calculate the class means, calculate the explained variance, project the data and visualize the datapoints on the new linear discriminant space.

- Reload the 'wineData.npy' and 'wineLabels.npy' files.
- Go through Task 3.3 in the notebook, completing the TODOs.
- Calculate the per-class mean vectors.
- Calculate the explained variance of the identified linear discriminants.
- Select a number of linear discriminants to keep, based on the explained variance of the new discriminant space.
- Project the two sets into the new space and visualize them.

☐ Challenge Task 3.4

Some questions to consider:

1. Why do we not compute a projection matrix and mean value for the testing sets?
2. Why does LDA give us nice distinct clusters for our Wine Dataset when PCA does not?
3. What benefit does dimensionality reduction provide? What are the drawbacks?