# **Data Mining**

# **Homework #3**

**Problem 1 (50 points)**

In this problem, we apply Principal Component Analysis to stock market index data. We will use records of daily closing prices of S&P 500 stocks from January 1, 2011 to December 31, 2014 retrieved through Yahoo Finance. The data is stored in the attached file named “SP500\_close\_price\_no\_missing.csv”. There are actually only 471 stocks in this file, the rest of the stocks were not included either because Yahoo Finance returned an error or because the stock was not listed as of January 1, 2011. The file named “SP500\_ticker.csv” contains ticker information for each included stock, as well as the corresponding company name and its industry sector assignment.  *Note: I also included the raw data before missing values are imputed, for those who would like to challenge yourself, you may try to do missing value imputation yourself.*

1. Fit a PCA model to **log returns** (log return = log( Price [t+1]/Price [t]) derived from stock price data and complete the following tasks (25 points)
   1. Derive log returns from the raw stock price dataset
   2. \* Plot a scree plot which shows the distribution of variance contained in subsequent principal components sorted by their eigenvalues.
   3. \* Create a second plot showing cumulative variance retained if top N components are kept after dimensionality reduction (i.e. the horizontal axis will show the number of components kept, the vertical axis will show the cumulative percentage of variance retained).
   4. \* How many principal components must be retained in order to capture at least 80% of the total variance in data?
2. Analysis of principal components and weights (25 points)
   1. \*Compute and plot the time series of the 1st principal component and observe temporal patterns. Identify the date with the lowest value for this component and conduct a quick research on the Internet to see if you can identify event(s) that might explain the observed behavior.
   2. Extract the weights from the PCA model for 1st and 2nd principal components.
   3. \*Create a plot to show weights of the 1st principal component grouped by the industry sector (for example, you may draw a bar plot of mean weight per sector). Observe the distribution of weights (magnitudes, signs). Based on your observation, what kind of information do you think the 1st principal component might have captured?
   4. \*Make a similar plot for the 2nd principal component. What kind of information do you think does this component reveal? (Hint: look at the signs and magnitudes.)
   5. \*Suppose we wanted to construct a new stock index using one principal component to track the overall market tendencies. Which of the two components would you prefer to use for this purpose, the 1st or the 2nd? Why?

(c)Bonus points (20 points)

* Rerun the PCA analysis with the most updated data downloaded from the available stock price API.

**Problem 2 (50 points)**

Let us experiment with a few feature selection methods. We will use data stored in ‘BMI.csv’ file. This data contains measurements of Body Mass Index (BMI) obtained for a number of human subjects. The goal is to predict fat percentage (fatpctg) using all other features available in data.

a) **Wrapper method:** Search for the best set of features using backward and forward stepwise regression

b) **Filter method:** output a ranking of features using correlation statistics (i.e. between any of the input variables and output)

c) **Embedded method:** (1) Lasso regression; (2) random forest (feature importance ranking)

\*Write a paragraph to summarize your findings from the above experiments.

**GPT policy:**

You are allowed to use GPT to complete this assignment, if you did, please make sure you summarize your GPT usage (GPT statement), and share a link to the chat history.

**Deliverable**: (please do not upload any documents to blackboard, only hyperlink)

* An **editable link** to Google Doc with answers (i.e. anyone with the link can edit). Only those marked as \* are to be answered in Google doc, otherwise, it should be in your Python Notebook
* A link to the Python Notebook uploaded to GitHub
* If you use GPT, the GPT statement and link to the chat history