# Homework # 3

August 11, 2019

## 1 S&P500 index

Your friend Joe, who has recently started a consluting firm. He is working on a project to help a client understand the S&P500 index. Since both Joe and his client are new to the financiel industry and had little knowledge about the market, they decide to starting with public available data and do some simple analysis. Joe learned that you are working on a master's program on data analytics and called you for your opinion. Joe provide you with two data set in ".csv" format.

**Data:**
S&P500 historcal index value: "**sp500indexdaily.csv**"
Stock price of S&P500 listed companies (2013-2018): "**sp500_cmpny_all_stocks_5yr.csv**"

### 1.1 Data Exploration

After getting the data, you decide to explore the data by visualizing it first. You started by looking at the records in "**sp500indexdaily.csv**" and made the following plots

(1) Use the "close" price of SP500 and plot it against trading dates (between 2009-01-01 and 2018-12-31). Instead of using actual date as x-axis, you can assign an integer to each trading date and set "2009-1-1" to "0", "2009-1-2" to "1", "2009-1-3" to "2" ... etc.).

(2) In addition to the temporal property of SP500, you look into the statistical distribution of the index. Specifically, you create a histogram of the SP500 index for each year between 2009 and 2018.

(3) Can you guess the underlying probability distribution of SP500?

### 1.2 Linear Regression - Single Predictor

After looking at the SP500 index change over time i.e. plots 1.1(1), you decide a linear regression can capture the trend fairly well. Therefore, you build a linear regression model with SP500 index as your target variable and time index as your predictors.

(1) How do you interpret the numbers?

(2) Make an in sample prediction for the SP500 index between 2009 and 2018. Compare it with the actual SP500 index. What does the model capture/not capture?

(3) Based on the model, where do you think the SP500 index will be by the end of year 2019.

(4) Can you design a validation test to ensure the linear model is not overfitting? Describe your approach (you don't have to implement it).

### 1.3 Linear Regression - Multiple Predictors

After searching the website, you learned that SP500 is an index built on the stock price 500+ companies. The largest 5 components are "Microsoft (MSFT)", "Apple Inc. (AAPL)", "Amazon.com Inc (AMZN)", "Berkshire Hathaway Inc (BRK.B)" and "Johnson & Johnson (JNJ)". You postulate that the SP500 index might be well replicated by the 5 top components already. So you did the following

(1) Build a multi-predictor linear regression model using SP500 index as the target variable and the stock price of MSFT, AAPL, AMZN, BRK.B and JNJ as the predictors (only use the stock price and SP500 index between 2013-02-08 and 2018-02-07). Hint: you need to reformat the company data so that each column represents the stock price and the row represents the date.

(2) Are there any missing values in the data?

(3) Check the significance of the variables by looking at the p-values of each variable. How do you interpret the coefficients?

(4) Can the model be used to predict future SP500 price?

## 2 Mammal Classification Tree

You were given a data set "**zoo.csv**" that includes 101 animals and a list of characteristics of the animals e.g. do they have feather, do they lay eggs or not etc. Build a CART model to classify if an animal is mammal or not.

(1) Calculate the overall entropy of of the target variable "ismammal", using the definition $H = p_1 log(p_1) + (1 - p_1) log(1 - p_1)$

(2) To build a classification tree, you need to decide the splitter for each nodes of a binary tree. Using the criterion that $hair > 0.5$ and split the dataset in to two branches. Calculate the entropy at each branch and the average entropy change.

(3) Check the entropy changes for all the following features i.e. 'feathers', 'eggs', 'airborne', 'aquatic' and 'backbone'. Which one would you use to make the first split?

(4) Build a CART model using the sklearn package and compare the model with your calculation. Is the first split the same as yours? You may use the python code provided in "CARTmammals.py"