Homework # 3

October 16, 2020

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 analytic 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 trend of SP500 against time, you look into the statistical distribution of the index. Specifically, create a histogram of the SP500 index between 2009 and 2018.

(3) Does the histogram look like any distribution that you have learned so far? Provide your thoughts on how to describe a distribution like this (hints: where are the distribution peaks; could it be superposition of multiple normal distributions)?

1.2 Regression - Single Predictor

After looking at the SP500 index change over time i.e. plots 1.1(1), you decide to build a regression model to capture the trend of SP500 index.

(1) In the GLM framework that you learned, what kind of distribution will you choose from the exponential family?

(2) Build the GLM model

(3) How can you interpret the betas in your model, phrase a few words that you can explain to Joe?

(4) Make an in sample prediction for the SP500 index between 2009-2018. Compare it with the actual SP500 index (visualize it in a graph). What does the model capture/not capture?

(5) Calculate the summed-square-error (SSE) of your in sample prediction. Do one for all data points and do one for the data points between 2017-2018

(6) Based on the model, where do you think the SP500 index will be by the end of year 2020.

(7) In order to validate your model, you also decide to use the data between 2009-2016 as training set and use 2017-2018 as test data set. What's the summed-square-error (SSE) after applying the model on the test data set. Compare your validation SSE with the insample SSE from question (5)

1.3 Regression - Multiple Predictors

After searching the website, you learned that SP500 is an index built based 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 speculate that the SP500 index might be a weighted average of the stock prices i.e. $SP500 = \beta_1 MSFT + \beta_2 AAPL + \beta_3 AMZN + \beta_4 BRK.B + ... + \beta 0$.

(1) Build a multi-predictor 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. A sample code is provided to you.

(2) What is the in-sample prediction summed-square-error compare (SSE) now, is it better than what you get in 1.2(5) on average?

(3) Visualize your prediction and compare it with the true value of SP500?

(4) Check the significance of the variables by looking at the p-values of each variable, are the coefficients significant? How can you phrase a few words to explain the coefficients to Joe?

(5) Can you use the model to predict future SP500 price, what is preventing you from doing that?

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"