Mark Chafin

IST – 5535

Homework 1

Chapter 2 section 2.4 (pages 52-53)

1. Question 2
   1. Regression – because CEO salary is a continuous variable. Inference – because we want to know what factors affect CEO salary. n = 500 – 500 firms provide our data points. p = 3 (profit, # of employees, and industry)
   2. Classification – because we want to know whether the product will be a success or failure. Prediction – we want to predict if it will fail or be successful. n = 20. p = 13 (price charged for the product, marketing budget, competition price, and ten other variables.)
   3. Regression – we are interested in % change in the $/€ which is a quantitative dependent variable. Prediction – we are interested in predicting the % change. n = 52 (weekly data for 2012). p = 3 (% change in the US market, the % change in the British market, and the % change in the German market.)
2. Question 4
   1. You want to use image recognition to see if someone is male or female. The response would be “Male” or “Female”. The goal of the application is to make an inference because we want to know whether it is male or female.

You want to know whether a stock is going to go up or down. The response would be true or false = up or down. The goal of the application is to make a prediction, so we know whether a specific stock it going to go up or down.

You want to know whether gas prices will rise or fall the next day. The response would be “up” or “down” if the gas prices are going to go up or down. The goal of the application is to make a prediction so we know whether the gas prices will rise or fall.

* 1. I want to predict how much a specific stock will move. This is a regression model and the goal of the application is to make a prediction of how much the stock will move.

I want to know how busy a store is going to be at a certain time, this is a prediction model because I am prediction how many people will be at the store.

I want to know if gas prices will rise or fall based on current political events. The goal of the application is inference because I want to know what factors affect gas prices.

* 1. Some real-life examples of cluster analysis: spam filter, identifying fake news, and identifying fraud or criminal activity.

1. Question 6
   1. The advantages of parametric ML is that the algorithms are simple, fast, and require less data to be trained. Some disadvantages are that they are constrained, they stick to their parameters. There is a limited complexity and when data does not fit a specific distribution, they can be a poor fit.

Some advantages of non-parametric ML they are a lot of features, and since it does not make a lot of presumptions it is very powerful. It learns from the training data set and has a better performance. Some disadvantages are that it requires more data and is slower than parametric algorithms. There is a higher chance of overfitting, it is harder to explain why a specific prediction is made on non-perimetric algorithm.