Data Preparation and Analysis

Homework 1

Question 1.1:

Problem 1:

a) Flexible statistical learning is **better** than inflexible method. Since the number of observations or records are more, the data points will lie close to the flexible method.

b) Flexible statistical method will be **Worse** than inflexible method. As the number of observations are in small number the flexible approach might overfit the data.

c) When the relationship between the Predictors and the Response is highly non-linear the flexible method will be a **better** fit.

d) When the variance of the error is high flexible model will try to fit noise or extreme values in the error term and will **Worsen** the variance.

Problem 2:

a) **Regression, Inference -** Here given the predictors we are trying to infer what are the values responsible for affecting the CEO salary.
N – 500 firms in the U.S.
P – profit, number of employees, industry.

b) **Classification, Prediction –** Here we are trying to predict whether the new product that we are trying to launch will be a success or a failure.
N – 20 similar products that were previously launched.
P – price charged, Marketing Budget, Competition price and ten other variables.

c) **Regression, Prediction –** Here we are trying to predict the % of change in the value.
N – Weekly data for the year 2012.
P - % change in the US market, % change in the British market, % change in the German market.

Problem 4:

a) **1-** Checking whether a movie that is going to be released tomorrow will be a success or a failure. (Classification, Prediction).
Predictors – Money spent, Release date.
2- Checking whether the Indian team will win or fail the next world cup. (Classification, Prediction).
Predictors – **Current Form, Playing Location, Injury details.**
3- Checking for health problems. (Classification, Inference).
Predictors – Blood sugar, Blood Pressure, Temperature level.

b) **1-** Predicting the sale profit of an advertising industry. (Regression, prediction).
Predictors – Amount of money spent on tv, radio, newspaper advertising.
2- Checking what are the predictors responsible for increase in the CEO salary. (Regression, Inference)
Predictors – Profit, Number of employees.
3- Predicting next week stock price of a company. (Regression, Prediction).
Predictors – Last quarter performance, new deals with the company.

c) **1- Image processing.**
2- Pattern Recognition.
3- Tumor Analysis in medical field.

Problem 6:

Parametric Statistical Learning – First we make assumptions about the functional form or shape of **f**. Once we assume **f** the problem of estimating **f** is greatly simplified. Finally, after the model has been selected we fit the data to the selected model.

Non-Parametric Statistical Learning – This method doesn't make explicit assumptions about the functional form of **f.** So, it is necessary that we have large number of records to estimate a goof **f.**

Advantage – Parametric method requires only few parameters to model the function **f** for regression or classification when compared to the non-parametric approach.

Disadvantage – Since we are assuming **f** in the parametric statistical learning, the assumed function could be inaccurate.