**Problem 2:** Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

**Ans:** This is a regression problem where we are interested in “inference”. Here, n=500 and p=4.

1. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

**Ans**: This is a classification problem where “prediction” is the main interest. Here, n=20 and p=14.

1. We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

**Ans:** This is a regression problem with interest in “prediction”. Here, n=52 and p=4.

**Problem 4:** You will now think of some real-life applications for statistical learning.

1. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

**Ans**:

1. Spam emails: This is a well-known real-life classification problem. Essentially, it’s a binary classification (response: Spam/Not Spam) and the goal is prediction. The possible predictors are Spam Emails, Non-Spam Emails.
2. Detecting Tumors: Classification techniques are extensively used in biomedical imaging for detecting tumors which in essence is a binary classification problem (response: Tumor / Non-tumor) and the goal is prediction. The predictors could be medical images (MRI scans) with and without tumors, DNA profiles, etc.
3. Understanding landscape changes: In the field of Earth Observation, large-scale satellite data are used for understanding landscape changes over time which involves classification of land use and land cover, such as identifying human settlements, water bodies, agriculture areas, etc. The primary goal is prediction, however, inference, also acts as a secondary goal. The predictors are image texture, heterogeneity, entropy, cloud cover percentage, vegetation indices, water body indices, etc., and the response variable is land-cover type.
4. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

**Ans:**

1. Weather forecasting: Machine learning is extensively used to “predict” weather events ahead of time which is essentially a regression problem. The predictors are air temperature, pressure, humidity, solar radiation, precipitation, wind speed, and wind direction. Here, the response could be Rainfall for forecasting rainfall events.
2. Stock Market Prediction: This is another well-known regression problem where the main goal is “prediction”. The predictors are currency exchange rates, share prices, GDP, etc., and the response is the “Stock Price”.
3. Sport Event Prediction: Regression is useful to predict the outcome of a sporting event such as soccer. The main goal is “prediction” wherein the predictors could be selected from various team or player statistics such as Past Wins, Player Strength, Team Formation, etc., with the response being “Outcome” (Win/Lose).
4. Describe three real-life applications in which cluster analysis might be useful.

**Ans:**

1. Market Analytics: Cluster analysis can be used to group customers based on some attributes such as spending power.
2. Recommender Systems: The generation of target advertisements can be done through cluster analysis.
3. Fraud Detection: Cluster analysis is useful for finding unknown patterns in online banking systems and is extensively used for fraud detection.

**Problem 6:** Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

**Ans:** The parametric statistical learning approach requires the assumption of underlying statistical distributions in the population data from which the samples are drawn (e.g., normal distribution) which is not necessary in case of the non-parametric approach.

## *Advantages of parametric approach:*

## The parametric approach is statistically more powerful when the assumptions are met.

## Easier interpretation than the non-parametric approach.

## *Disadvantages of parametric approach:*

1. This approach is less robust because of the stricter criteria enforced.
2. Leads to inconclusive results when the sample size is very low or very large.
3. Only works with numerical data, whereas, non-parametric approach can be applied to all the scales of measurements- nominal, ordinal, interval, and ratio.