**3. 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.**

(a) You want to predict whether a particular customer is going to click on an online advertisement or not. You have information on whether or not they clicked on 200 other ads, in addition to whether the ad was in the same category, whether the ad was shown during regular working hours, whether the ad was shown on a weekend, and the percent of all customers who had previously clicked on the ad.

This is a classification problem as response variable is qualitative or categorical (Yes ,No)

* n or number of training examples/observations = 200
* p or number of predictor variables x0 to x4 = 5
* y = response variable = whether ad was clicked or not
* x0 = 1
* x1 = whether the ad was in the same category
* x2 = whether the ad was shown during regular working hours
* x3= whether the ad was shown on a weekend
* x4= the percent of all customers who had previously clicked on the ad

--------------------------------------------------------------------------------------------------

(b) Suppose it is the end of the quarter and you wish to predict your score on the final exam. You have data from 20 classes you have previously taken, consisting of your final exam scores, your average scores on the midterms (i.e., one average midterm score per class), your average homework scores (i.e., one average homework score per class), and whether the final exam was take-home or not.

This is a Regression problem as response variable is quantitative or numerical

* n or number of training examples/observations = 20
* p or number of predictor variables x0 to x3 = 4
* y = final exam scores
* x0 = 1
* x1 = average scores on the midterms
* x2 = your average homework scores
* x3= whether the final exam was take-home or not.

--------------------------------------------------------------------------------------------------

(c) You work for an ice cream shop and are in charge of determining what factors affect how much ice cream is sold each day. For 600 days you have information on how much ice cream the shop sold, in addition to whether the day was sunny or not, what the temperature was, whether school is in session or not, whether your most popular flavor was available that day, and whether you had recently run any advertisements

This is a Regression problem as response variable is quantitative or numerical

* n or number of training examples/observations = 600
* p or number of predictor variables x0 to x5 = 6
* y = how much ice cream is sold each day
* x0 = 1
* x1 = whether the day was sunny or not
* x2 = what the temperature was
* x3 = whether school is in session or not
* x4 = whether your most popular flavor was available that day
* x5 = whether you had recently run any advertisements

--------------------------------------------------------------------------------------------------

**4. In this problem you will brainstorm real-life applications for statistical learning. Your answers aren’t allowed to be the same as any of the examples in the other homework problems.**

(a) Describe three real-life applications in which classification might be useful, one from political science, one from sports, and one from an area of your choice. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

**Political Science:** Given variables from a survey result such as trust factor with scale of 1-10 where 10 is most trusted party, political ideology, past success rate, predict if a political party would win an upcoming election or not. Number of survey records=1000

This is a classification problem and the goal of this problem is prediction in nature.

* n or number of training examples/observations = 1000
* p or number of predictor variables x0 to x3 = 4
* y = political party win or not
* x0 = 1
* x1 = Trust factor
* x2 = political ideology
* x3 = past success rate

**Sports:** Given variables from polling such as performance, favorite team or not, past success rate, predict if a sports team would win the match or not. Number of poll records =1000

This is a classification problem and the goal of this problem is prediction in nature.

* n or number of training examples/observations = 1000
* p or number of predictor variables x0 to x3 = 4
* y = sports team win or not
* x0 = 1
* x1 = performance
* x2 = favorite team or not
* x3 = past success rate

**Health:** Given a patient with variables such as Tumor size, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, classify if the tumor is malignant or benign as we are not interested in the relation between response and predictors but rather interested in predicting if tumor is benign or malignant. Number of patient records = 50

This is a classification problem and the goal of this problem is prediction in nature.

* n or number of training examples/observations = 50
* p or number of predictor variables x0 to x4 = 5
* y = tumor is benign or malignant
* x0 = 1
* x1 = Tumor size
* x2 = Clump Thickness
* x3 = Uniformity of Cell Size
* x4 = Uniformity of Cell Shape

--------------------------------------------------------------------------------------------------

(b) Describe three real-life applications in which regression might be useful, one from agriculture, one from business, and one from an area of your choice. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

**Agriculture:**

In agriculture, soil variables are known to be important factors that determine the level of crop productivity. Let the 6 soil variables be soil reaction, organic matter, total nitrogen, total phosphorus, total potassium and soil texture. We are interested in evaluating how each of the soil variables affect the crop productivity and whether there’s any association between the soil variables and crop productivity. What is the relationship between the crop productivity and each soil variable? Number of observations = 100

This is a regression problem and the goal of this problem is inference in nature.

* n or number of training examples/observations = 100
* p or number of predictor variables x0 to x6 = 7
* y = crop productivity
* x0 = 1
* x1 = soil reaction
* x2 = organic matter
* x3 = total nitrogen
* x4 = total phosphorus
* x5 = total potassium
* x6 = soil texture

**Business:**

A florist wants to predict the sales of number of flower bouquets per day based on certain variables namely Season, whether it’s a weekday or weekend, whether it’s a sunny day or not and whether flowers are fragrant or not. The florist has recorded this data for 100 days.

This is a regression problem and the goal of this problem is prediction.

* n or number of training examples/observations = 100
* p or number of predictor variables x0 to x4 = 5
* y = number of flower bouquets
* x0 = 1
* x1 = season
* x2 = whether it’s a weekday or weekend
* x3 = whether it’s a sunny day or not
* x4 = whether flowers are fragrant or not.

**Education:**

Predict tuition costs for a 4-year public university course in USA based on variables inflation index, Debt\_to\_GDP, and interest rates. There are 50 entries from 1969-2019 for making this prediction.

This is a regression problem and the goal of this problem is prediction.

* n or number of training examples/observations = 50
* p or number of predictor variables x0 to x3 = 4
* y = tuition costs for a 4-year public university course in USA
* x0 = 1
* x1 = Inflation index
* x2 = Debt\_to\_GDP
* x3 = Interest rates

--------------------------------------------------------------------------------------------------

(c) Describe three real-life applications in which cluster analysis might be useful, one from education, one from meteorology, and one from an area of your choice. Be sure to describe why it would be useful.

**Education**: From a database of students, group the students into groups that are somehow similar or related by different variables, such as final scores, specialization and so on. This will help in the analysis of students’ academic performance and provide them the required help.

**Meteorology:** Classify the rainfall distribution in space and time from historical data of similar meteorological and climatological conditions based on clustering.

**Astronomical Data Analysis:** Clusters of galaxies are a useful proxy to trace the distribution of mass in the universe. By measuring the mass of clusters of galaxies on different scales, one can follow the evolution of the galaxies. Analyses of smaller sample of galaxies helps in discovering new classes of galaxies

**Organize large computer clusters:** In large data centers, there are large computer clusters and we want to figure out which machines tend to work together. If we put those machines together, the data center work can be done more efficiently.