Chapter 2 Conceptual Exercises

1. Indicate whether we would generally expect the performance of a flexible statistical method to be better or worse than an inflexible method. Justify your answer.

a. n is extremely large, and p is small.

A flexible method would perform better because it can handle larger amounts of observations and can find a non-linear relationship if we do not assume a linear functional form.

b. p is extremely large, and n is small.

An inflexible method would perform better. With a small number of observations, the flexible method could follow a limited number of data points too closely. This can cause a higher variance and a small bias reduction, which can result in a high test MSE/error rate.

c. The relationship between the predictors and response is highly non-linear.

Because we cannot assume a linear functional form (which is inflexible), a flexible method would be ideal to identify the relationship between the response and predictors.

d. The variance of the error terms is extremely high.

With that much error, it is best to use an inflexible method so that the model does not pick up too much noise. Using a more flexible method would also increase variance in the model.

2. Explain whether each scenario is a classification or regression problem. Indicate whether we are most interested in prediction or inference. Provide n and p.

a. We collect a data on the top 500 firms in 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.

This is a regression problem because our response is quantitative (CEO salary).
We want to understand the relationship between predictors and response, so we are interested in inference.
n=500 (number of firms)
p=3 (profit, number of employees, industry)

b. We are considering launching a new product and wish to know whether it will be a success or 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.

This is a classification problem because our response is qualitative (success/failure).
We are only interested in making a prediction.
n=20 (similar products)
p=14 (success or failure, price charges, marketing budget, competition price, +10 other variables)

c. We are interested in predicting the percent 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.

This is a regression problem because our response is quantitative (percent change in USD/Euro exchange rate).
We are only interested in making a prediction.
n=52 (weeks in 2012)
p=3 (% changes in US market, British market, and German market)

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

OBSERVATION	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using KNN.

a. Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

We use the formula D = $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ to calculate Euclidean distances, where $y_i = 0$.

OBSERVATION	1	2	3	4	5	6
DISTANCE	3	2	sqrt(10), 3.16	sqrt(5), 2.24	sqrt(2), 1.41	sqrt(3), 1.73

NOTE: For parts b and c, we employ: P(Y =j | X = x_o) = $\frac{1}{K}\sum_{i\in\aleph_0} I(y_i = j)$. Our test point is (0,0,0).

b. What is our prediction with K=1? Why?

Our prediction is Green.
The closest observation to (0,0,0) is Observation 5 with distance 1.41. Therefore, X = x_5.

P(Y = Red | X = x_5) = $\frac{1}{1}I(x_5 = Red) = 0$.

P(Y = Green | X = x_5) = $\frac{1}{1}I(x_5 = Green) = 1$.

c. What is our prediction with K=3? Why?

Our prediction is Red.
The 3 closest observations to (0,0,0) are 5, 6, and 2.

P(Y = Red | X = x_o) = $\frac{1}{3}I(x_5 = Red) + I(x_6 = Red) + I(x_2 = Red) = \frac{2}{3}$.

P(Y = Green | X = x_o) = $\frac{1}{3}I(x_5 = Green) + I(x_6 = Green) + I(x_2 = Green) = \frac{1}{3}$.

d. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

A small value of K would be the best. With a small K, the decision boundary would be more flexible and can find patterns in the data that the Bayes decision boundary does not show.