**CS480**

**Computational Statistics II**

**Homework #3 1.**

Problem 1 on page 52

For each of parts (a) through (d), indicate whether we would expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer. (a) The sample size n is extremely large, and the number of predictors p is small.

Flexible statistical learning method would perform better than inflexible method. As the sample size n is exceptionally large, a flexible method can handle non-parametric and does not need to follow one specific model to extract additional Information, without adhering to a particular model, which lowers the risk of overfitting.

(b) The number of predictors p is extremely large, and the number of observations n is small.

Since there are few observations, a rigid statistical learning method would be preferable to a flexible one. This is because when n is small, there are fewer opportunities to describe non-linear or non-additive connections between the input features and the output variable, which increases the danger of overfitting in flexible methods.

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

Flexible statistical learning can model non-linear or non-additive relationships between the input features and output variable. As it justifies that Flexible statistical learning will be best fit for relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e., σ 2 = Var(ɛ), is extremely high.

Since the variance of error terms is extremely high, the sample will have a lot of noise. In this case, an inflexible method is better as it reduces the chance of overfitting.

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

2. Problem 2 on page 52

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.

Prediction is about estimating the value of the response variable based on predictor or input variables.

Inference is about understanding the relationship between the response and the predictor variables.

(a) 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.

Given,

N=500, P=3... The issue is regression and interests between the predictors and response, and the interface aids in our knowledge of the relationship between the predictors and response.

(b) 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 comparable 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.

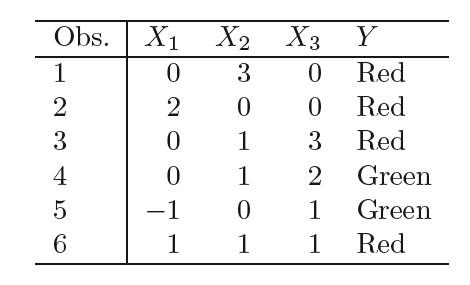
To determine whether the product is a failure or success is the problem, which points to a classification issue. As it states, based on the inputs supplied to the model, we must forecast the response. It's a prediction issue.

(c) 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.

This is a Regression problem, as the % change in the USD/Euro is quantitative Dependent variable. This is a prediction problem as we are predicting the % change in the USD/Euro. Here, n=52 (52 weeks in a year) and p=3

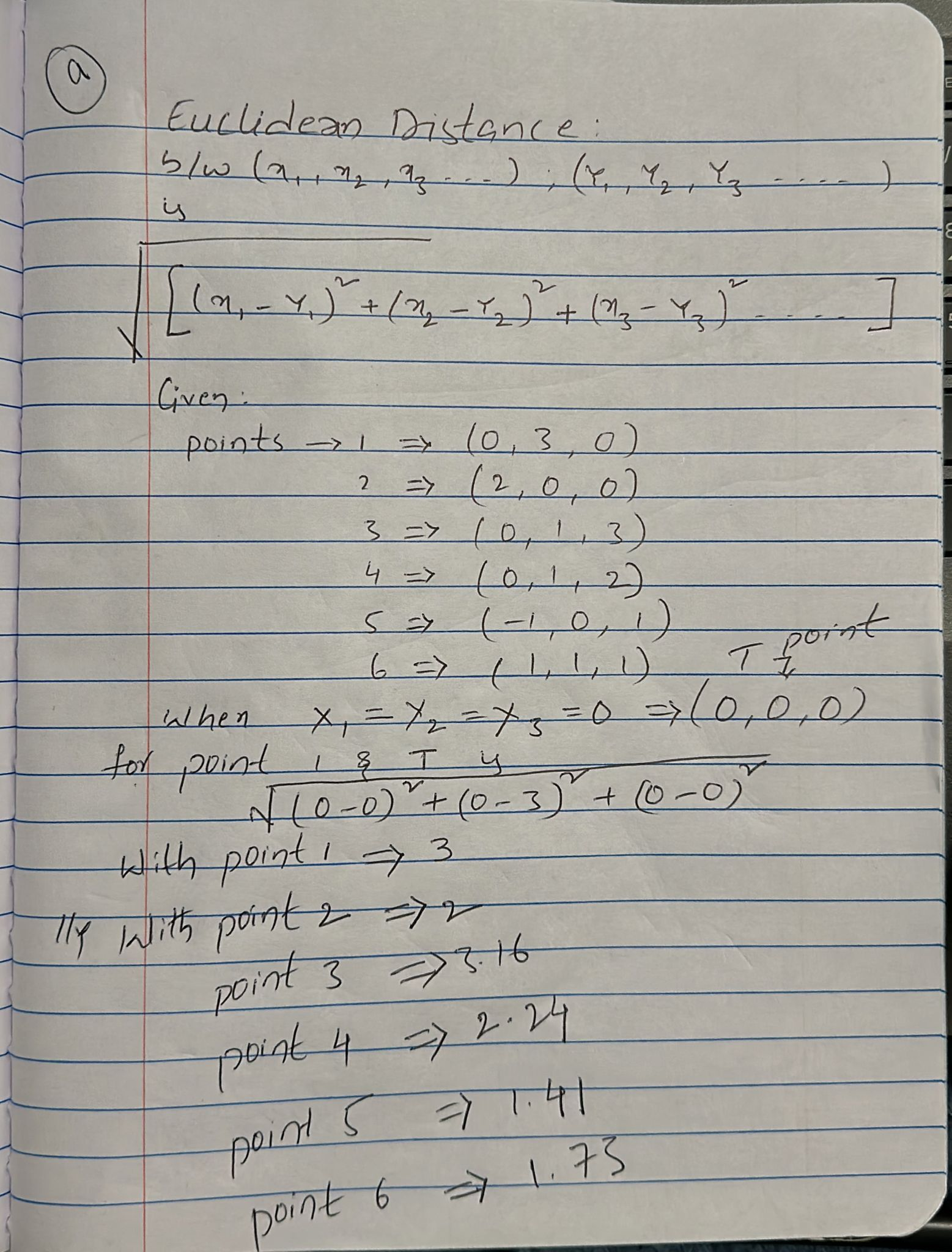
3. Problem 7 on page 53

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



Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.



(b) What is our prediction with K = 1? Why?

As, K = 1. The nearest one Neighbour to the given point (0, 0, 0) is point 5 with the lowest Euclidean distance as 1.41

(c) What is our prediction with K = 3? Why?

As, K = 3. The nearest three Neighbours to the given point (0, 0, 0) are:

* Point 5 with lowest Euclidean distance as 1.41
* Point 6 with second highest Euclidean distance as 1.73
* Point 2 as it is the next lowest Euclidean distance with 2

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

The best value for k would be small as the Bayes decision boundary in this problem is highly nonlinear. Small k value will limit the variability caused by the non-linear boundary.