**Machine Learning & Data Mining, Spring 2020**

**Homework 1**

Due March 27

1. Provide n (the number of samples) and p (the number of features) of the data collected in the following scenario (Note. The outcome variable to predict is not considered as a feature).

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, marking budget, competition price, and ten other variables.

#answer

n = 20 , p = 13

1. Explain in which circumstances the median can be more useful than the mean to represent a certain population data.

#answer

data에 극단적인 값(outlier)가 많아서 평균값에 영향을 많이 주는 경우에 는 모집된 데이터의 평균값(mean)보다 중앙값(median)을 쓰는 경우가 더 유용하다.

1. Fill the missing values in the following data table using 1) mean-based imputation, 2) KNN-based mean imputation with K=1, and 3. Assume that we use the typical Euclidean distance to find the nearest neighbors (e.g. dist( (1,2), (2,3)) =( (2-1)2 + (3-2)2 )1/2

|  |  |  |  |
| --- | --- | --- | --- |
| ID | x1 | x2 | x3 |
| 1 | 10 | NaN | 5 |
| 2 | 8 | 6 | 5 |
| 3 | 3 | 6 | 2 |
| 4 | 15 | 7 | 4 |
| 5 | 4 | 4 | 7 |
| 6 | 2 | 3 | 2 |
| 7 | 12 | 10 | 8 |

#answer

1) mean-based imputation

NaN = = 6

2) KNN-based mean imputation with K=1, and 3

d(y1,y2)^2 =(10-8)^2 +(5-5)^2 = 4

(1)k=1

NaN = 6

(2)k=3

NaN = (6+7+10)/3 = 23/3

d(y1,y3)^2 =(10-3)^2 +(5-2)^2 = 58

d(y1,y4)^2 =(10-15)^2 +(5-4)^2 = 26

d(y1,y5)^2 =(10-4)^2 +(5-7)^2 = 40

d(y1,y6)^2 =(10-2)^2 +(5-2)^2 = 68

d(y1,y7)^2 =(10-12)^2 +(5-8)^2 = 13