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

**Homework 1**

Due March 27

**Name: dowoon Kim**

**Student ID: 201520893**

**Department: Cyber Security**

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.

* Y(outcome variable): success / failure
* Sample(n): 20
* Feature(p): price, marking budget, competition price, other 10 variables…
* The number of samples(n) is 20 and the number of features(p) is 13.

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

* If there are a lot of outliers in the data and we need to consider even outliers or if the distribution of the data is not normal distribution, median can be more useful than the mean because mean is heavily influenced by outliers. As a result, if we want to describe the center of the data it’s better to use median, but if we want common center, it’s better to use mean.

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 |

1. Mean-based imputation

* NaN = (6+6+7+4+3+10)/6 = 6

1. KNN-based imputation with k=1, and 3.

D(1,2) = =

D(1,3) = =

D(1,4) = =

D(1,5) = =

D(1,6) = =

D(1,7) = =

* K= 1) NaN= 6/1 = 6
* K= 3) NaN= (6+7+10)/3 = 7.67