**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, (because of collecting data on 20 similar products)

p=13 (because of price charged for the product, marking budget, competition price, and ten other variables.)

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

answer : circumstances where sorted data is given as input

sorting time complexity is O(nlogn)

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 : K=1, ID=1 x2= 6

K=3, ID=1 x2 = 8

Current data is considered integer data.  
I rounded it up. (The calculated value is 7.666666666....)

dist(1,2) = sqrt(4) k = 1, k = 3

dist(1,3) = sqrt(58)

dist(1,4)= sqrt(26) k =3

dist(1,5) = sqrt(40)

dist(1,6) = sqrt(73)

dist(1,7) = sqrt(13) k =3