CSE342: Statistical Machine Learning

Assignment-1 (Question-5)

Problem

Given the following dataset, perform k-NN classification to predict the target variables for a given test point.

S. No.	Age	Loan (in Million $)	HPI	BHK
1.	25	40	135	2
2.	35	60	256	3
3.	45	80	231	3
4.	20	20	267	4
5.	35	120	139	4
6.	52	18	150	2
7.	23	95	127	2
8.	40	62	216	4
9.	60	100	139	2
10.	48	220	250	3
11.	33	150	264	4

For a test instance having the features **Age** $= 37$ and **Loan** $= 142$, the problem is to predict the continuous target variable **HPI** and discrete target variable **BHK** for the values of $k \in \{1, 2, 3\}$.

Notations Used

1. Let the Euclidean Distance d between two points x and y be denoted by $d(x, y)$.

$$d(x, y) = \sqrt{(x_0 - y_0)^2 + (y_0 - y_1)^2} \tag{1}$$

2. Let the set of k-nearest neighbors of a point x be denoted by $N_k(x)$.

3. Let x_i refer to the point whose serial number is i.
4. Let $x_Q(i)$ denote the value of the feature variable x_Q and $y_Q(i)$ denote the value of the target variable y_Q of the point x_i.
5. Let x_A and x_L denote the feature variables **Age** and **Loan** respectively, and y_H and y_B denote the target variables **HPI** and **BHK** respectively.
6. Then, each point x_i in the dataset is the feature vector $(x_A(i),\ x_L(i))$, and has a target vector $y_i = (y_H(i),\ y_B(i))$.

Solution

Let the given test point be $\hat{x} = (37, 142)$. The following table contains the Euclidean distances of the test sample $\hat{x}$ from all the points in the dataset sorted in ascending order of distance.

i	$x_A(i)$	$x_L(i)$	$d(\hat{x}, x_i)$
11.	33	150	8.944
5.	35	120	20.091
9.	60	100	47.885
7.	23	95	49.041
3.	45	80	62.514
10.	48	220	78.772
8.	40	62	80.056
2.	35	60	82.024
1.	25	40	102.703
4.	20	20	123.179
6.	52	18	124.904

Predicting the Continuous Target Variable HPI

The continuous target variable **HPI** is predicted by taking the mean of the **HPI** values of the k-nearest neighbors.

$$y_H = \frac{1}{k} \sum_{x_i \in N_k(\hat{x})} y_H(i) \tag{2}$$

Predicting the Discrete Target Variable BHK

The discrete target variable **BHK** is predicted by taking the mode of the **BHK** values of the k-nearest neighbors.

$$y_B = \text{ mode } \{y_B(i) \mid x_i \in N_k(\hat{x})\} \tag{3}$$

Hyperparameter: $k = 1$

For $k = 1$, we only consider the sample closest to the test point, which gives $N_1(\hat{x}) = \{x_{11}\}$.

S. No.	**Age**	**Loan (in Million $)**	**HPI**	**BHK**
11.	33	150	264	4

Target Variable: y_H **= HPI**

Using (2), we get:

$$y_H = \frac{264}{1} = 264 \tag{4}$$

Target Variable: y_B **= BHK**

Using (3), we get:

$$y_B = \text{ mode } \{4\} = 4 \tag{5}$$

Final Solution for $k = 1$

Hence, the final solution for $k = 1$ is $y = (264, 4)$, which means that the predicted value of the target variables **HPI** and **BHK** are 264 and 4 respectively.

Hyperparameter: $k = 2$

For $k = 2$, we consider two samples closest to the test point, which gives $N_2(\hat{x}) = \{x_{11}, x_5\}$.

S. No.	**Age**	**Loan (in Million $)**	**HPI**	**BHK**
11.	33	150	264	4
5.	35	120	139	4

Target Variable: y_H **= HPI**

Using (2), we get:

$$y_H = \frac{264 + 139}{2} = 201.5 \tag{6}$$

Target Variable: y_B **= BHK**

Using (3), we get:

$$y_B = \text{ mode } \{4, 4\} = 4 \tag{7}$$

Final Solution for $k = 2$

Hence, the final solution for $k = 2$ is $y = (201.5, 4)$, which means that the predicted value of the target variables **HPI** and **BHK** are 201.5 and 4 respectively.

Hyperparameter: $k = 3$

For $k = 3$, we consider three samples which are closest to the test point, which gives $N_3(\hat{x}) = \{x_{11}, x_5, x_9\}$.

S. No.	Age	Loan (in Million $)	HPI	BHK
11.	33	150	264	4
5.	35	120	139	4
9.	60	100	139	2

Target Variable: y_H = HPI

Using (2), we get:

$$y_H = \frac{264 + 139 + 139}{3} = 180.667 \tag{8}$$

Target Variable: y_B = BHK

Using (3), we get:

$$y_B = \text{ mode } \{4, 4, 2\} = 4 \tag{9}$$

Final Solution for $k = 3$

Hence, the final solution for $k = 3$ is $y = (180.667, 4)$, which means that the predicted value of the target variables **HPI** and **BHK** are 180.667 and 4 respectively.