1. Suggest the appropriate normalization (if applicable) for the values in the above table. Classify the label of a new sample with values [6, 7] using a k-nearest neighbours classifier using Euclidean distance for k=3 and k=5, respectively. Explain the results.

min max normalization

Minimum value of variable A is 0 and maximum is 10,
Minimum value of variable B is 1 and maximum is 10,

Then the variable A and variable B can be transformed as follows,

Samples	Variable A	Normalized A	Variable B	Normalized B
S1	1	0.1	1	0
S2	2	0.2	2	0.11111
S3	3	0.3	3	0.22222
S4	4	0.4	4	0.33333
S5	5	0.5	4	0.33333
S6	8	0.8	10	1
S7	10	1	9	0.88889
S8	0	0	4	0.33333
S9	4	0.4	1	0
New sample	6	0.6	7	0.66667

Then the point (6,7) will be transformed to (0.6,0.666667),

Euclidean Distance to new value:

Sample	Euclidean Distance
S1	0.833333
S2	0.684574
S3	0.536219
S4	0.38873
S5	0.34801
S6	0.38873
S7	0.457584
S8	0.686375
S9	0.69602

1.1.1 k=3:

when the k is equals to three, then we will choose top three smallest Euclidean Distance samples, which is S5,S4 and S6.

The Label of S4,S5 and S6 is “Malignant”, “Malignant”, “Benign”, respectively. There are

2 votes for “Malignant” and 1 for “Benign” .

Then for k=3, the label of new sample is “Malignant”.

1.1.2 k=5

when the k is equals to 5, then we will choose top three smallest Euclidean Distance samples, which is S5,S4, S6, S7 and S3.

The Label of S3,S4,S5,S6 and S7 is “Benign”, “Malignant”, “Malignant”, “Benign”, “Benign”, respectively. There are 2 votes for “Malignant” and 3 for “Benign” .

Then for k=5, the label of new sample is “Benign”.

As a matter of fact, since the number difference is not so high. Maybe the normalization is not necessary.

2. Given the following training data set, build a Naïve Bayes classification model and classify the following test sample.

$P(Solen = Yes) = 1/2$

$P(Solen = No) = 1/2$

1. **Color:**

   1.1. Red

   $P(Color = Red|Yes) = \frac{3}{5} = 0.6$

   $P(Color = Red|No) = \frac{2}{5} = 0.4$

   1.2. Yellow

   $P(Color = Yellow|Yes) = \frac{2}{5} = 0.4$

   $P(Color = Yellow|No) = \frac{3}{5} = 0.6$

2. **Type**

   2.1. SUV

   $P(Type = SUV|Yes) = \frac{1}{5} = 0.2$

   $P(Type = SUV|No) = \frac{3}{5} = 0.6$

   2.2. Sports

   $P(Type = Sports|Yes) = \frac{4}{5} = 0.8$

$$P(Type = Sports|No) = \frac{2}{5} = 0.4$$

3. Origin

3.1.Domestic

$$P(Origin = Domestic|Yes) = \frac{2}{5} = 0.4$$

$$P(Origin = Domestic|No) = \frac{3}{5} = 0.6$$

3.2.Imported

$$P(Origin = Imported|Yes) = \frac{3}{5} = 0.6$$

$$P(Origin = Imported|No) = \frac{2}{5} = 0.4$$

The Naïve Bayes classification is as follows:

$P(Color = Red|Yes) = 0.6$
$P(Color = Yellow|Yes) = 0.4$
$P(Color = Red|No) = 0.4$
$P(Color = Yellow|No) = 0.6$
$P(Type = SUV|Yes) = 0.2$
$P(Type = Sports|Yes) = 0.8$
$P(Type = SUV|No) = 0.6$
$P(Type = Sports|No) = 0.4$
$P(Origin = Domestic|Yes) = 0.4$
$P(Origin = Imported|Yes) = 0.6$
$P(Origin = Domestic|No) = 0.6$
$P(Origin = Imported|No) = 0.4$

In this case, the label of test sample can be predicted as follows:

$$\begin{aligned} &P(Color = Red, Type = SUV, Origin = Domestic|Stolen = Yes) \\ &\qquad = P(Color = Red|Yes) \times P(Type = SUV|Yes) \\ &\qquad \times P(Origin = Domestic|Yes) = 0.6 \times 0.2 \times 0.4 = 0.048 \end{aligned}$$

$$\begin{aligned} &P(Color = Red, Type = SUV, Origin = Domestic|Stolen = No) \\ &\qquad = P(Color = Red|No) \times P(Type = SUV|No) \times P(Origin = Domestic|No) \\ &\qquad = 0.4 \times 0.6 \times 0.6 = 0.144 \end{aligned}$$

$$\begin{aligned} &P(Stolen = Yes|Color = Red, Type = SUV, Origin = Domestic) \\ &\qquad = P(Color = Red, Type = SUV, Origin = Domestic|Stolen = Yes) \\ &\qquad \times P(Yes) = 0.024 \end{aligned}$$

$$\begin{aligned} P(Stolen = No|Color = Red, Type = SUV, Origin = Domestic) \\ = P(Color = Red, Type = SUV, Origin = Domestic|Stolen = No) \times P(No) \\ = 0.072 \end{aligned}$$

0.072>0.024, then the Stolen label of this test sample is No.

3. Discuss in what situations an ensemble classification scheme may fail to improve the classification performance.

3.1 Success rate of base classifiers is less than 0.5

The success rate of ensemble learning is as follows:

$$P_{emsemble} = \sum_{k=\left(\frac{T}{2}\right)+1}^{T} \binom{T}{k} p^k 1 - p^{T-k}$$

If base classifiers success rate is less than 0.5, when T increases, the rate of ensemble rate will decrease gradually to 0.

3.2 The base classifiers are highly correlated with each other

If all base classifiers are highly corrected, then their performance will also similar with each other, in this case, their decision about the label also will be similar, then wrong answer will always be wrong answer. a lower correlation among ensemble model members will increase the error-correcting capability of the model. So, it is preferred to use models with low correlations when creating ensembles.

3.3 The mechanism of voting will lead to wrong answer

The answer of majority is wrong, and minority is correct, then voting will leading to the wrong answers which are the answers of majority.

4. Consider the following 6 training data points in a 2D space for a binary classification task (circles are "+" squares are "–"). What is the hyperplane function learned by a linear SVM?

Target function: $min\colon \frac{1}{2} ||w||^2$

Set (2,2) as s1,(4,1) as s2, S3 as (3,3

Then change to k1(2,2,1), k2(4,1,1) and k3(3,3,1)

$$\left\{\begin{array}{c} a1\ k1k1 + a2k2k1 + a3k3k1 = -1, \\ a1\ k1k2 + a2k2k2 + a3k3k2 = -1, \\ a1\ k1k3 + a2k2k3 + a3k3k3 = 1, \end{array}\right\}$$

Then,

$$\left\{\begin{array}{c} 9a1 + 11a2 + 13a3 = -1, \\ 11a1 + 18a2 + 16a3 = -1, \\ 13a1 + 16a2 + 19a3 = 1, \end{array}\right\}$$

Then a1=-143/9, a2=-2/9, a3=100/9

Then

$$a1(2,2,1) + a2(4,1,1) + a3(3,3,1) = (\frac{2}{3}, \frac{4}{3}, -5)$$

So the hyperplane parameter can be calculated and show as follows,

$$\bar{w} = \left(\frac{2}{3}, \frac{4}{3}\right), b = -5$$

Then function of hyperplane is $2X1 + 4X2 - 15 = 0$

5. Assume that S1, S2, and S3 are initially assigned in one cluster and S4 and S5 in another cluster. Apply the k-means method to cluster the five samples into two clusters. Show the detailed workings.

Idea: Use the mean value of each cluster value as the centroid, and apply Euclidean distance as the metrics to update the centroids.

5.1 First Round:

Centroid 1 in cluster (S1,S2, and S3) is ((10+12+20)/3=14,(5+3+10)/3=6)=>(14,6)
Centroid 2 in cluster(S4,S5) is ((22+18)/2=20,(12+8)/2=10)=>(20,10)
Then the Euclidean distance between five samples and each two centroids is as follows:

Samples	Euclidean distance with Centroid 1	Euclidean distance with Centroid 2
S1	4.123106	11.18034
S2	3.605551	10.63015
S3	7.211103	0
S4	10	2.828427
S5	4.472136	2.828427

We could find that S3 in cluster 1 has a larger Euclidean distance with Centroid 1 but the distance with centroid 2 is lower.
Then the cluster can be updated by two clusters (S1,S2) and (S3,S4,S5).

5.2 Second Round:

Centroid 1 in cluster (S1,S2) is (10+12)/2=11,(5+3)/2=4=>(11,4)
Centroid 2 in cluster (S3,S4,S5) is (20+22+18)/3=20,(10+12+8)/3=10=>(20,10)

Then the Euclidean Distance is show as follows.

Samples	Euclidean distance with Centroid 1	Euclidean distance with Centroid 2
S1	1.414214	11.18034
S2	1.414214	10.63015
S3	10.81665	0
S4	13.60147	2.828427
S5	8.062258	2.828427

Then the cluster in second round can be update to two , (S1,S2) as one cluster and (S3,S4,S5) as another. The Centroid is the same with the beginning of the Second round. Then we could finish the algorithm.

After k-mean algorithm with k=2, the five samples can be clustered to two cluster with one is(S1,S2) and another is (S3,S4,S5).

6. Given the set of 1-dimensional data points {6, 12, 18, 24, 30, 42, 48} and the initial centroids are {18, 45} (K=2), cluster the points to the nearest centroid, calculate the total sum of squared error (SSE) for the two clusters. Show both the clusters & total SSE for each centroid. Also show cluster membership and centroid.

Choose Euclidean Distance as the metrics

6.1 First Round:

Set centroid 1 is 18 and centroid 2 is 45,

The Distance of seven samples with each centroid is show as follows:

Samples	value	Euclidean distance with Centroid 1	Euclidean distance with Centroid 2
S1	6	12	39
S2	12	6	33
S3	18	0	27
S4	24	6	21
S5	30	12	15
S6	42	24	3
S7	48	30	3

Compared with the distance, we could see that S1,S2,S3,S4,S5 has smaller distance with centroid 1 and S6, S7 has smaller distance with centroid 2.
Then the cluster can be updated as (6,12,18,24,30) (42,48).
Centroid 1 is updated to (6+12+18+24+30)/5=18,
Centroid 2 is updated to (42+48)/2=45.

We could see that centroids don’t change after the first round, then we finish the algorithms.

Cluster 1 has 18 as is centroid and its membership is {6,12,18,24,30}
Cluster 2 has 45 as is centroid and its membership is {42,48}

SSE of cluster 1 is:

$$\mathrm{SSE}_1 = (6-18)^2 + (12-18)^2 + (18-18)^2 + (24-18)^2 + (30-18)^2 = 360$$

SSE of cluster 2 is:

$$\mathrm{SSE}_2 = (42 - 45)^2 + (48 - 45)^2 = 18$$

Then total SSE for these clusters is:

$$SSE = 360 + 18 = 378$$