Take Test: Assignment 5

Test Information

Description

Instructions

Multiple Attempts This test allows multiple attempts.

Force Completion This test can be saved and resumed later.

Saving All Answers | Close Window | Save and Submit

QUESTION 1

36 points Saved

Suppose we want to predict whether a restaurant is popular based on its cuisine, price, and whether it delivers, and we collected training data as in table 1. Answer following questions.

Table 1: Training dataset (P - popular, NP - not popular)

ID	Cuisine	Price	Delivery	Popularity
1	Thai	$	Yes	P
2	Korean	$$$	No	P
3	Thai	$$	No	NP
4	American	$	No	P
5	American	$$	Yes	NP
6	Korean	$$	Yes	P
7	Thai	$$	Yes	P
8	Korean	$	No	NP
9	American	$$$	No	P
10	American	$	Yes	NP

1a.

i. What is the expected information (entropy) needed to classify a tuple in D, i.e., *Info*(D)? 0.971

ii. What is the information gain for the "Price" attribute? 0.371

1b.

i. What is the Gini index for the attribute "Delivery"? 0.48

ii. What's the reduction in impurity, in terms of Gini Index, with respect to the "Delivery" attribute? 0

1c. Based on the training data, we want to construct a Naive Bayes classifier. (No smoothing is required.) Please estimate the following terms, rounding to three decimal places.

i. Pr(Popularity = `P') 0.6

ii. Pr(Popularity = `NP') 0.4

iii. Pr(Price = '$', Delivery = `Yes' , Cuisine = `Korean' | Popularity = `P') 0.056

iv. Pr(Price = '$', Delivery = `Yes' , Cuisine = `Korean' | Popularity = `NP') 0.063

1d. Suppose a restaurant has the values: Price = '$', Delivery = 'Yes' , Cuisine = 'Korean'. Based on the calculation in part (c.), is this restaurant classifed as popular? Answer 'yes' or 'no'. yes

QUESTION 2

27 points Saved

We have ten training points, which are listed and plotted in figure 1; in addition, four test points with their true labels are shown in table 2. Please answer the following questions.

id	x1	x2	y
1	1.6	1.1	+1
2	3.7	2.2	-1
3	1.1	3.1	+1
4	2.5	3.7	+1
5	0.4	0.6	+1
6	4.0	0.5	-1
7	3.4	2.9	+1
8	2.6	0.5	-1
9	2	1	-1
10	3.7	1.4	-1

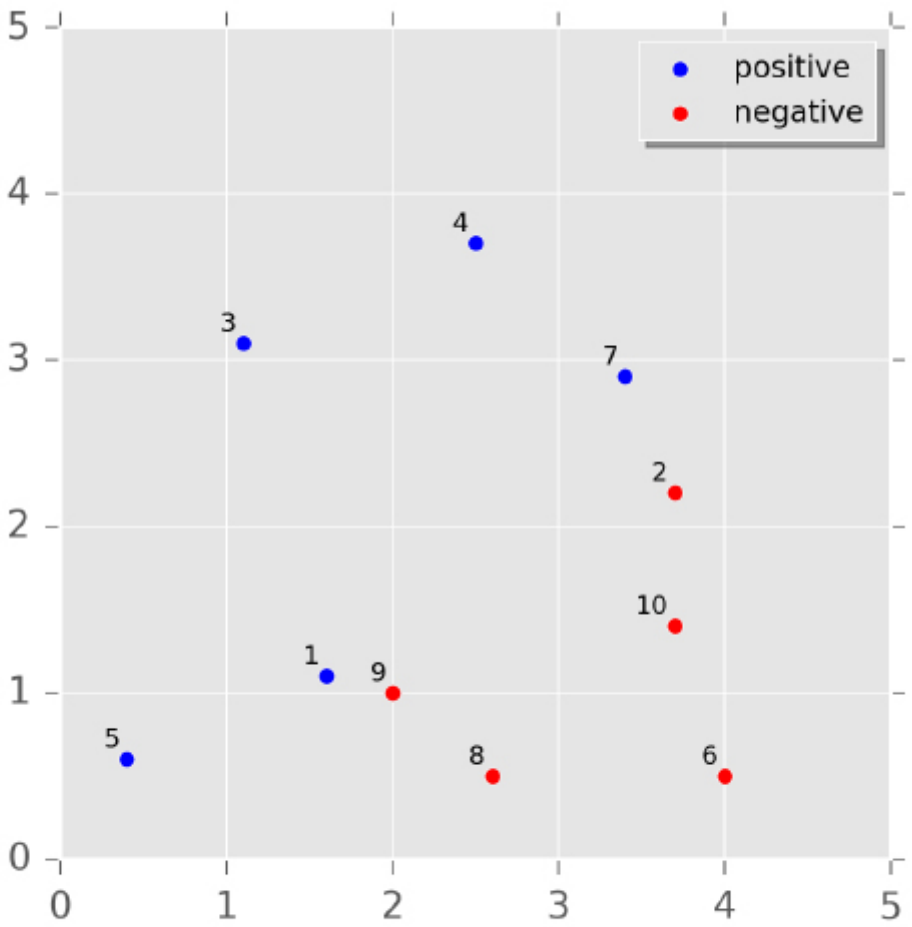

Figure 1: training data

Table 2: Test data with true labels

id	x1	x2	y
11	1.9	1.0	-1
12	2.7	1.8	+1
13	0.6	2.1	-1
14	3.4	2.4	-1

For this question you will perform k-nearest neighbor classification with k=3. Please use Euclidean distance and ties are broken at random. If a question asks you to list points, please format your answer like "6, 2, 10"; the order of the points is not important as long as you have the correct points. If a question asks you what label will be assigned to a point, please answer "+1" or "-1".

2a.

i. Which three data points are nearest to the test point with id = 11? 1,8,9

ii. What label will be assigned to this point? -1

2b.

i. Which three data points are nearest to the test point with id = 12? 2,9,10

ii. What label will be assigned to this point? -1

2c.

i. Which three data points are nearest to the test point with id = 13? 1,3,5

ii. What label will be assigned to this point? 1

2d.

i. (3') Which three data points are nearest to the test point with id = 14? 2,7,10

ii. (1') What label will be assigned to this point? -1

2e. What is the testing error, i.e., what percent of test points were wrongly classified? (Please format your answer like "10%") 50%

QUESTION 3

20 points Saved

For this problem we have nine data points, which are listed and plotted in figure 2.

id	x	y
1	8.2	6.4
2	0.1	6.7
3	1.5	7.8
4	2.2	3.4
5	1.6	3.5
6	4.3	9.7
7	2.5	2.2
8	6.2	3.1
9	5.5	0.3

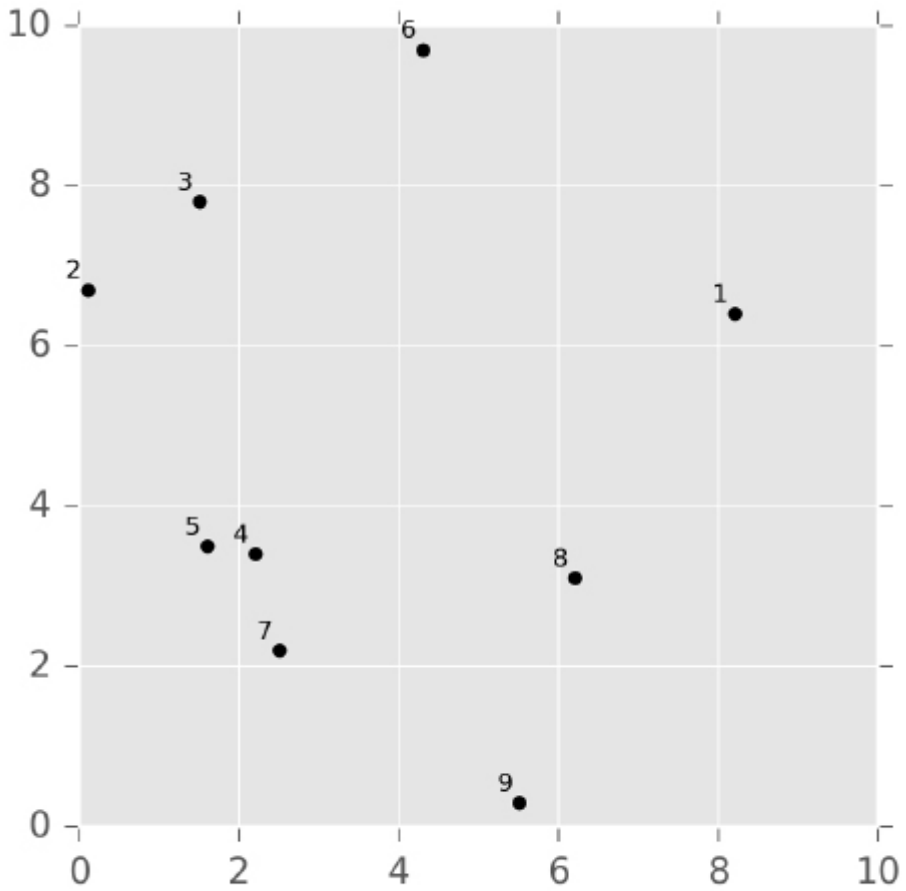

Figure 2: Data

You will perform k-means clustering with k=3 using Euclidean distance. Remember that each step of k-means consists of two parts: first, updating the cluster centroids based on the points in each cluster, and second, reassigning each point to the cluster represented by its closest centroid.

Begin with the following initial cluster centroids:

- Cluster 1: (2.0, 5.0)
- Cluster 2: (6.0, 0.5)
- Cluster 3: (6.0, 5.0)

For questions asking which points are assigned to a cluster, please use the point ids and format your answer like "2, 6, 7, 9"; order does not matter, as long as the points are correct. For questions asking for a cluster centroid, please format your answer like "(1.265, 0.300)" remembering to round your answer to 3 decimal places.

Question Completion Status:

4a. Which point(s) are initially assigned to

i. Cluster 1? 2,3,4,5,7

ii. Cluster 2? 9

iii. Cluster 3? 1,6,8

4b. After the first step of k-means, what is the new cluster centroid for

i. Cluster 1? (1.580,4.720)

ii. Cluster 2? (5.500,0.300)

iii. Cluster 3? (6.233,6.400)

4c. How many additional steps are required before k-means terminates? 2

4d. After k-means has finished, which point(s) are assigned to

i. Cluster 1? 2,3,4,5,7

ii. Cluster 2? 8,9

iii. Cluster 3? 1,6

QUESTION 4

20 points Saved

Using the same data as in question 3, repeat k-means clustering with k=3, but now use the following initial cluster centroids:

- Cluster 1: (1.0, 4.0)
- Cluster 2: (0.2, 6.0)
- Cluster 3: (4.3, 2.0)

Again, for questions asking which points are assigned to a cluster, please use the point ids and format your answer like "2, 6, 7, 9"; order does not matter, as long as the points are correct. For questions asking for a cluster centroid, please format your answer like "(1.265, 0.300)" remembering to round your answer to 3 decimal places.

4a. Which point(s) are initially assigned to

i. Cluster 1? 4,5

ii. Cluster 2? 2,3,6

iii. Cluster 3? 1,7,8,9

4b. After the first step of k-means, what is the new cluster centroid for

i. Cluster 1? (1.900,3.450)

ii. Cluster 2? (1.967,8.067)

iii. Cluster 3? (5.600,3.000)

4c. How many additional steps are required before k-means terminates? 2

4d. After k-means has finished, which point(s) are assigned to

i. Cluster 1? 4,5,7

ii. Cluster 2? 2,3,6

iii. Cluster 3? 1,8,9

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Saving All Answers | Close Window | Save and Submit

Question Completion Status: