图形用户界面, 文本, 应用程序, 信件, 电子邮件

描述已自动生成1. We can learn that the rank of A is 4096-4086 = 10

Because of

We can get

So , we now know that the non-zero singular values of A have a square relationship with

Non-zero singular values of A is [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]

2. 4096 x 10

图表, 散点图

描述已自动生成

1. a. I will calculate the Euclidean distance
2. between the test data(x1=4, x2=6) and the sample data. For example, the first point (2,4)from class A, the distance of (2, 4) and (4, 6) is sqrt((x1-x2)^2 + (y1-y2)^2) = sqrt(8), and the point from

算法流程

总体来说，KNN分类算法包括以下4个步骤：[4] 

①准备数据，对数据进行预处理[4]  。

②计算测试样本点（也就是待分类点）到其他每个样本点的距离[4]  。

③对每个距离进行排序，然后选择出距离最小的K个点[4]  。

④对K个点所属的类别进行比较，根据少数服从多数的原则，将测试样本点归入在K个点中占比最高的那一类[4]  。

图表

描述已自动生成

从上面给出的四个点可以发现，ROC曲线图中，越靠近(0,1)的点对应的模型分类性能越好，所以，可以确定的是**ROC曲线图中的点对应的模型，它们的不同之处仅仅是在分类时选用的阈值(Threshold)不同，每个点所选用的阈值都对应某个样本被预测为正类的概率值**

From the above figure, it can be found that in the ROC curve graph, the closer the point (0, 1), the better the classification performed of the model. Therefore, the models corresponding to the points in the ROC curve are different in the threshold value replaced during classification. The threshold value selected by each point corresponds to the probability value of a certain sample being predicted as a positive class. So the KNN is better than the Naïve Bayes.

文本, 信件

描述已自动生成

Let D = defective items

Total production = 2000 + 3000 + 5000 = 10000

P(A) = 2000/10000 = 0.2

P(B) = 3000/10000 = 0.3

P(C) = 5000/10000 = 0.5

P(D|A) = 0.01

P(D|B) = 0.01

P(D|C) = 0.03

Now

P(A|D) = P(D|A) P(A) / P(D)

P(D) = (0.01 x 2000 + 0.01 x 3000 + 0.03 x 5000) / 10000 = 0.02

P(A|D) = (0.01 x 0.2) / 0.02 = 0.1

表格

描述已自动生成

Let the final result is attribute, the probability of attribute is P(A)

P(A|Y) = 1/3 x 2/3 x 2/3 = 4/27

P(A|N) = 2/6 x 1/6 x 1/6 = 1/108 We should assume that there is one ‘Marital status = Divorced’, because when the ‘Default borrower = NO’, the marital status has no dicorved.

P(A|Y)P(Y) = 4/27 x 1/3 = 4/81

P(A|N))(N) = 0 x 6/9 x 1/108= 0

P(A|Y)P(Y) > P(A|N))(N)

So the last borrower will default.

文本, 信件

描述已自动生成

Let the function is f(x)

So

= 0.5 x [2, 4, 2] – [2, 4, 2] + 2[1, -1, 1] = [1, -4, 1]

W1 = W0 – = [1, -1, 1]- 0.1[1, -4, 1] = [0.9, -0.6, 0.9]

图表

描述已自动生成

After the M=8, it is overfitting. The error on the training dataset is small while the error on the test dataset is large.

1. Reducing the complexity of the model to prevent the model from overfitting the training set.
2. Increase the size of training data.

文本, 信件

描述已自动生成

1. Hk
2. 2-dimensional cannot divided dataset C, so we use three-dimensional, so we use the 3D and make the kernel as small as possible.
3. A: y = -x + 2

B: x-1.5 = 0

C:

1. A: (0, 0), (1, -1), (-1, 1), (1, 1)

B: (2, 3), (2, 2), (2, -2), (1, -1)

C: (-1, 1), (2, 2), (1, -1), (2, -2), (-2, -2)

1. A: 2 < 2.5, so the x\* belongs to the positive class.

B: 2 > 1.5, so the x\* belongs to the negative class.

C: 2.5 x 2.5 = 6.26 > 5, so the x\* belongs to the positive class.

图表, 折线图

描述已自动生成

1. **Dropout. Randomly delete some hidden neurons to prevent the model from being overly dependent on some neurons.**
2. **Decrease the number of layers. Reduce the complexity of the model and prevent the model from overfitting the training set**

**Question6**

Consider the random variables X, Y, Z which have the following joint distribution:

p(X, Y, Z) = p(X)p(Y |X)p(Z|Y ).

Show that X and Z are conditionally independent given Y

Question 7 
[10 marks] 
Consider a linear regression problem of estimating a non-linear function f with 30 training data points 
{ (Xi, "i . As shown in Fig Q8, three linear regressions were independently performed with polyno- 
mial features of polynomial orders I, 4 and 15. 
1. Identify which polynomial degrees (out of 1, 4 and 15) could correspond to models A, B and C. 
2. The models A, B, and C independently reported sum-of-squared-error (SSE) values between all 
training data points and corresponding estimates as 0.35, 6.78 and 0.15, respectively. Explain why 
model C has a lower SSE, although seems to have a spurious fit. 
30 
SSE = - 

g. 
Describe a procedure for this particular dataset to determine n suitable model complexity, i.e. ibe 
polynomial order. 

1. The ‘1’is model B, the ‘4’is model A, the ‘15’is model C. It depends on the highest power (n) of the polynomial. The higher the highest power, the more times the line will bend, up to (n-1) times.
2. The model C has overfitting. About the whole training data points, the line of model C passes through almost all points, so its predicted value is close to the true value. Therefore, the model C has the smaller error.
3. Draw a simulated line to see the trend of the model. It depends on the highest power (n) of the polynomial. The higher the highest power, the more times the line will bend, up to (n-1) times.

Question 9 
Given that the SVD of a matrix M EVT. 
[9 marks) 
If it is not, how 
I. 
2. 
3. 
Is it correct to say: "The matrix MT M can be decomposed as MT AI = V" EVT„. 
to make it become correct. 
Chopse a correct matrix to fill in the question mark: MT NI V 
Based on the above, what are eigenvectors and eigenvalues of MT M ? 

Question 10 
14 marks) 
Give a vector a = (1, 3, 4) and other vectors = (4, 3, = (0.4, 10, 50), and = (1, 4, 10). 
1. Report the minimum distance of a to x2, and and the nearest neighbor of a. 
2. Find which of makes the smallest angle with a and report that angle. 

Question 12 
In Linear Regression given the following cost function 
2 
[9 marks] 
(1) 
where ho (xo)) = 9Tx(i), feature vector xo) € Rd of the i-th sample, and there are n data samples. We 
usually use the gradient descent to learn the minimum value of the cost function: O — OVJ(O). 
I. 
2. 
3. 
\Vhat is the name of the cost function above? 
Show step-by-step the gradient descent update for this cost function. 
How will the gradient update change if we add the regularization term? 

1. Mean square error
2. If will change to

Question 14 
[10 marks] 
Consider the problem of binary classification using the Naive Bayes classifier. You are given two di- 
mensional features (Xl, X2) and the categorical class conditional distributions in the tables below. The 
entries in the tables correspond to P(XI = ICi) and P(X2 = X21Ci) respectively. The two classes are 
equally likely. 
Class 
0.3 
0.2 
0.5 
Class 
-1 
1 
0.1 
0.6 
0.3 
-1 
o 
1 
Given a data point (I, I), calculate the following posterior probabilities: 
P(C21X1 = 1) 
0.5 
0.3 
0.2 
0.6 
0.3 
0.1 
P(CIIXI 1,X2 = 
I) and 

P(C1|X1 = 1, X2 = 1) =[ P(X1=1, X2=1|C1)P(C1)] / [P(X1 = 1, X2 = 1)]

=[P(X1=1|C1)P(X2=1|C1)P(C1)] / [P(X1=1|C1)P(X2=1|C1)P(C) + P(X1=1|C2)P(X2=1|C2)P(C2)]

=5/6

The following questions are multiple choices questions. Please check ALL CORRECT CHOICES 
and circle your answers. Note that every question should have at least one right answer. 
Question 15 
Which of the following are true about generative models? 
(5 marks) 
A. They model the joint distribution 
P(class = C AND sample = x). 
C. The Perceptron is a generative model. 
B. They can be used for classification. 
D. Linear discriminant analysis is a generative 
model. 

ABD

Question 16 
[5 marks) 
Suppose we train a hard-margin linear SVM on n > 100 data points in R2, yielding a hyperplane 
with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the 
maximum possible number of support vectors for the new hyperplane (assuming the n + 1 points are 
linearly separable)? 
B. 
3 

D

Question 10 
14 marks) 
Give a vector a = (1, 3, 4) and other vectors = (4, 3, = (0.4, 10, 50), and = (1, 4, 10). 
1. Report the minimum distance of a to x2, and and the nearest neighbor of a. 
2. Find which of makes the smallest angle with a and report that angle. 

Question 9 
Given that the SVD of a matrix M EVT. 
[9 marks) 
If it is not, how 
I. 
2. 
3. 
Is it correct to say: "The matrix MT M can be decomposed as MT AI = V" EVT„. 
to make it become correct. 
Chopse a correct matrix to fill in the question mark: MT NI V 
Based on the above, what are eigenvectors and eigenvalues of MT M ? 

Question 17 
[5 marks) 
Suppose we are given data comprising points of several different classes. Each class has a different 
probability distribution from which the sample points are drawn. We do not have the class labels. We 
use k-means clustering to try to guess the classes. Which of the following circumstances would undermine 
its effectiveness? 
A. Each class has the same mean. 
B. Choose k = n,the number of sample points. 
C. Some of the classes aren't normally distributedD. The variance of each distribution is small in all 
directions. 

ABC