DT044A

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2017

Time: 2017-08-31

Total: 100

A: 90
B: 80
C: 70
D: 60
E: 50
Fail < 50

The use of dictionaries and calculators are permitted.

Good Luck

Tingting Zhang
Tel: 0101428878

1. (8 p) List out the type of dataset that can be used in data mining. List out the kind of knowledge that will be produced from data mining.
2. (12 p) Suppose that a hospital tested the age and body fat data for 12 randomly selected adults with the following results:

age	23	27	29	31	49	50	52	54	56	57	58	60
%fat	9	8	15	26	27	31	35	30	33	30	34	41

Answer the following questions:

(a) Normalize the two attributes based on z-score normalization.

(b) Calculate a correlation coefficient. Are these two attributes positively or negatively correlated?

3. (5 p) In which situation, we need to transform Nominal data to Numeric data. List out the methods of transforming nominal to numeric data.
4. (8 p) Explain the problem of Overfitting, and how to prevent it. Give an example of overfitting (e.g. a plot of a function).
5. (7 p) Suppose the following table present two set of mean success rate obtained by ten folds Cross validation using two different learning schemes. All data set for the two different learning schemes are same and from same domain. Find out if one scheme is better than other one in confidence limit of 20%.

Scheme 1	90%	80%	90%	90%	100%	75%	90%	90%	80%	90%
Scheme 2	80%	90%	80%	90%	80%	85%	90%	80%	70%	95%

6. (20 p) decision tree

a) Suppose that a decision tree is build based on a training data set so that every leaf of the tree is 100% correct for the training tree. Is this a good decision tree? Why? If it is not how to improve the decision tree?

b) What kind of problem can be occur if we use highest information gain to decide the split of a node in decision tree?

Tingting Zhang
Tel: 0101428878

c) Given the following instances. What is information gain for each attributes? Which attribute is the best one to split the root?

Attribute variables			Target variable
age	Astigmatic	Tear production rate	**Lens type**
18	No	Normal	A
18	No	Reduced	A
20	No	Normal	A
20	No	Normal	A
21	yes	Normal	A
21	yes	Reduced	A
25	yes	Reduced	A
?	yes	Normal	A
29	yes	Normal	A
30	yes	Normal	A
31	yes	Normal	A
33	yes	?	A
33	yes	Reduced	C
35	No	Reduced	B
38	No	Normal	B
40	No	Normal	B
42	No	Normal	A
42	No	Normal	A
43	No	Reduced	C
43	yes	Reduced	C
48	yes	Normal	C

7. (20 p) Instance based learning method

a. When can the instance based learning be used? Give an example that the instance based learning is suitable.

b. Produce the 2-means clustering of the data points in above table using the Euclidean distance as the measure of dissimilarity and using the first and third data points to set up the initial centroids of the two clusters.

Instances	X1	X2	X3	X4
I1	1	0	0	0
I2	0	1	0	1
I3	0	0	1	1
I3	0	0	0	1
I4	0	0	0	0
I5	0	0	0	0
I6	0	1	0	0
I7	1	0	0	1
I8	0	0	1	1

c) Give one more distance function that can be used in clustering learning.

Tingting Zhang
Tel: 0101428878

8. (20 p) Linear regression
   a) Briefly describe linear regression.
   b) How to use linear method to learn un-linear model?
   c) Consider perceptron learning rule in the training data set for the following table. Assign 1 to initial weights and bias. Use the learning method to update weight w_0 (for bias) , w_1 for x and w_2 for y in one rounds.

x	**y**	**Target**
1	0	yes
1	2	no
-1	1	no
0	1	Yes

Tingting Zhang
Tel: 0101428878

Index: distributions and formulas

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Pr(E \mid H) = \Pi_{i=1}^{i=k}\left[\binom{N-\sum_{1}^{i-1} n_j}{n_i} p_i^{n_i}\right] = N!\Pi_{i=1}^{i=k}\frac{p_i^{n_i}}{n_i!}$$

$$p\left(\log\left(\frac{p}{t}\right) - \log\left(\frac{P}{T}\right)\right)$$

$$entropy(a) = \sum_i p_i \log(\frac{1}{p_i}) = -\sum_i p_i \log(p_i)$$

$$\inf(node) - \sum_i \frac{|subnode_i|}{|node|}\inf(subnode_i)$$

$$d([x_1,\ldots,x_n],[y_1,\ldots,y_n]) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

$$p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) \Big/ \left(1 + \frac{z^2}{N}\right)$$

$$\left(1 - \frac{1}{n}\right)^n = e^{-1} = 0.368$$

Tingting Zhang
Tel: 0101428878

Let f(x) is the logistic function, then f(x)' = f(x) (1-f(x))

$$\frac{mean_x - \mu}{\sqrt{\sigma_x^{\,2} / k}},$$

$$\frac{mean_d}{\sqrt{\sigma_d^{\,2} / k}}$$

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$U(A,B) = \frac{\sum_i \sum_j (a_i - a)(b_j - b)}{\sqrt{\left(\sum_i (a_i - a)^2\right)\left(\sum_i (b_i - b)^2\right)}}$$

Table 5.2 Confidence Limits for Student's Distribution with 9 Degrees of Freedom

Pr[X ≥ z]	z
0.1%	4.30
0.5%	3.25
1%	2.82
5%	1.83
10%	1.38
20%	0.88

Tingting Zhang
Tel: 0101428878

Table 5.1 Confidence Limits for the Normal Distribution

$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25