DTOYYA

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2017

Time: 2017-03-15

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. (10 p) What is supervised data mining? What is unsupervised data mining? What is semi-supervised data mining?

2. (10 p) Is it good to collect all the attributes in data mining? Why? Briefly describe two attribute selection approach; Filter and Wrapper. Compare these two methods.

3. (5 p) In what situations use Bootstrap better than Cross-validation. In which situation use cross-validation is better than bootstrap?

4. (10 p) Suppose the following table present two sets of observed success rates obtained by ten holdout validation using two different learning schemes. All data set for the two different learning schemes are same and from same domain.

   i) Find out the estimated success rates for these two schemes in confidence limit of 20%
   ii) Find out if one scheme is better than other one in confidence limit of 20%.

Scheme 1	90%	80%	90%	90%	80%	95%	90%	90%	70%	90%
Scheme 2	60%	90%	80%	90%	80%	80%	70%	80%	60%	80%

5. (5 p) How to evaluate a numeric prediction model? Name one method.

6. (20 p) Naïve Bayes Classify
   (i) Briefly describe Naïve Bayes classify. What is the basic limitation of Naïve Bayes classify? How to improve it?
   (ii) Build a naïve Bayes classifier to classify the target variable from the following table.
   (iii) Use your Bayes classify to predict the target variable of color = yellow, size = 65 and act = stretch.

	Attribute variables			Target variable
	color	**size**	**Act**	**inflated**
1	yellow	40	stretch	True
2	yellow	30	stretch	True
3	yellow	50	Dip	True
4	yellow	55	Dip	True
5	yellow	85	stretch	False
6	yellow	60	stretch	false
7	yellow	70	Dip	false
8	yellow	60	Dip	True
9	purple	35	stretch	False
10	purple	20	stretch	false
11	purple	10	Dip	false
12	purple	15	Dip	false
13	purple	80	stretch	true
14	blue	95	stretch	false
15	blue	75	Dip	false
16	blue	90	Dip	false

Tingting Zhang
Tel: 0101428878

7. (20 p) Nearest Neighbor Method

1. Briefly describe 1 nearest neighbor method. What is the main challenge of 1NN method? What kind of methods can be used to face the challenge?
2. Briefly describe two methods that can be used to reduce the samples in instance based learning database
3. Given the sample datasets (a, b, c, d, e, f, g) using Condensed NN Algorithm (IB2) method to produce a sample dataset for further learning.

   Class 1 has instance a, c, e. Class 2 has instance b, d, f.

   Distance (a, b) = 1, Distance (c, b) = 1
   Distance (a, d) = 2, distance (b, d) = 1
   Distance (b, e) =2, distance (c, e) = 1
   Distance (a, e) =2.5, distance (c,d) = 1.9
   Distance (a, c) = 2
   Distance (d, f) = 1, distance (e, f) = 2
   Distance (f, a) = 2.1, distance(f, b) = 1.9
   Distance (f, c) = 2.1

8. (20 p) decision tree
   (i) Briefly describe the decision tree construction method.
   (ii) Why should the decision tree be post pruned? What is basic principle of post pruning a decision tree?
   (iii) 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
	color	**size**	**Act**	**inflated**
1	yellow	40	stretch	True
2	yellow	50	stretch	True
3	purple	60	Dip	True
4	purple	70	Dip	True
5	yellow	88	stretch	True
6	yellow	90	stretch	false
7	yellow	?	stretch	false
8	yellow	100	Dip	false
9	purple	110	Dip	false
10	?	120	stretch	false
11	purple	130	stretch	false
12	purple	140	Dip	false

Tingting Zhang
Tel: 0101428878

Index: distributions and formulas

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

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

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

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

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

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

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

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

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