DTOYYA

Tingting Zhang
072 5188878

Examination of Data Mining, AV 2015

Time: 2015-06-02

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
072 5188878

1. (10 p) What is data mining? What is unsupervised data mining? List out data mining method(s) that can be used in unsupervised learning.

2. (5 p) What is the purpose of attribute selection?

3. (10 p) What is
   a. cross validation method
   b. holdout method
   c. Bootstrap
   d. Leave-One-Out

4. (5 p) What is the Minimum Description Length principle?

5. (5 p) How to select a learning scheme?

6. (5) 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 10%.

Scheme 1	80%	80%	90%	90%	80%	65%	80%	90%	60%	80%
Scheme 2	95%	95%	85%	95%	80%	85%	90%	80%	70%	95%

Tingting Zhang
072 5188878

7. (20 p) Classification

a. Build a decision tree for the following data set using gain to decide the split attribute. Stop condition: when number of instance less than 3.

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
29	yes	Normal	A
29	yes	Normal	A
30	yes	Normal	A
31	yes	Normal	A
33	yes	Normal	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

b. Generate classification rules from the decision tree.

c. What is different between association rules and classification rules

Tingting Zhang
072 5188878

8. (20 p) Instance based learning method

Instances	X1	X2	X3	X4	X5	X6
I1	1	0	0	0	1	0
I2	0	1	0	1	0	0
I3	0	0	1	1	0	1
I3	0	0	0	1	0	0
I4	0	0	0	0	1	0
I5	0	0	0	0	0	1
I6	0	1	0	0	1	0
I7	1	0	0	1	0	0
I8	0	0	1	1	1	0

a) 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.

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

c) What is hierarchical clustering?

9. (20 p) Naïve Bayes classification

a) Build a naïve Bayes classifier to classify the target variable from the above table in 7 a.

b) Use your classifier to decide the type of lenses of the following instances.

Attribute variables		
age	Astigmatic	Tear production rate
25	yes	Reduced
33	yes	Reduced

c) What is main limitation of Naïve Bayes classification?

Tingting Zhang
072 5188878

Index: distributions and formulas

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

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

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

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

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

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

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

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

Tingting Zhang
072 5188878

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

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

,

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

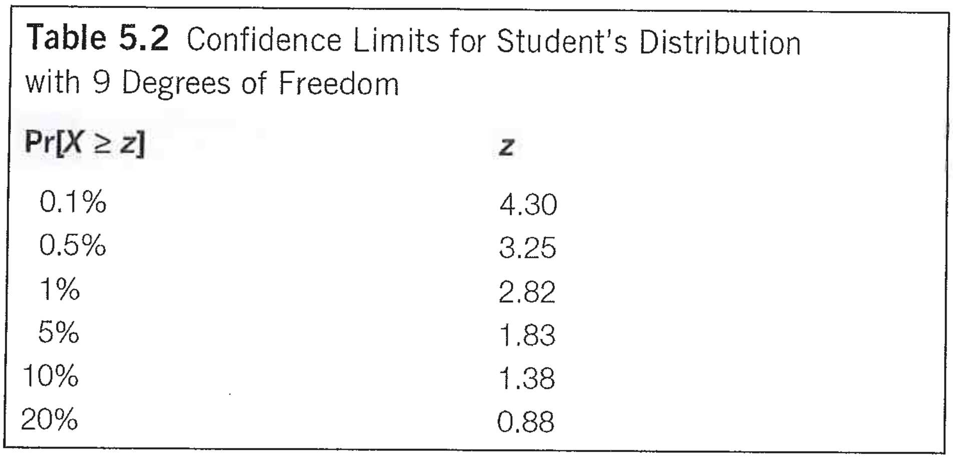

Table 5.2 Confidence Limits for Student’s Distribution with 9 Degrees of Freedom

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

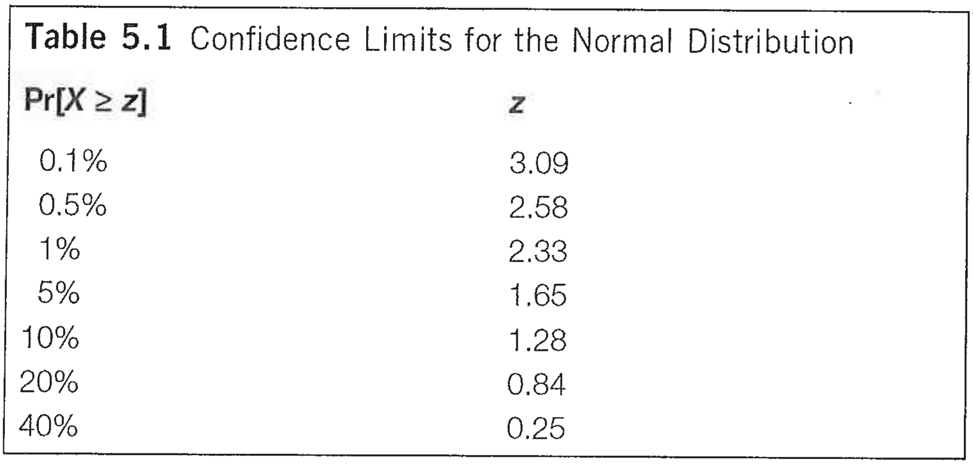

Table 5.1 Confidence Limits for the Normal Distribution

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