DT044A

Tingting Zhang
072 5188878

Examination of Data Mining, AV 2015

Time: 2015-08-24

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) List 4 kind of applications that data mining can be useful.

2. (5 p) Briefly describe data mining process.

3. (10 p) How to estimate the error in
   a. Cross validation
   b. Bootstrap
   c. Holdout
   d. Leave-One-out

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

5. (5 p) List two methods of attributes selection.

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	75%	75%	90%	85%	80%	65%	80%	90%	60%	80%
Scheme 2	95%	95%	80%	95%	80%	85%	90%	80%	70%	90%

Tingting Zhang
072 5188878

7. (30 p) Classification and association rules

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	Reduced	B
18	No	Reduced	B
20	No	Normal	B
20	No	Normal	B
25	yes	Normal	A
25	yes	Reduced	A
25	yes	Reduced	A
30	yes	Normal	A
30	yes	Normal	A
30	yes	Normal	A
30	yes	Normal	A
35	yes	Normal	A
35	yes	Reduced	C
35	No	Reduced	B
40	No	Normal	B
40	No	Normal	B
40	No	Normal	A
45	No	Normal	A
45	No	Reduced	C
50	yes	Reduced	A
50	yes	Normal	A

b. What is the pruning a tree/rule? What is the purpose of pruning a tree/rule?

c. What is different between association rules and classification rules

d. Use Apriori algorithm find all frequent item sets with *minimum* 25% (5 instances).

e. Use the frequent item sets from c. to generate all the association rules that satisfy *min-cover*= 25% and *min-accurate* = 60%

Tingting Zhang
072 5188878

8. (10 p) Linear model

   a) 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**
0	-1	yes
0	1	no
1	1	no
1	1	Yes

9. (10 p) Instance based learning method

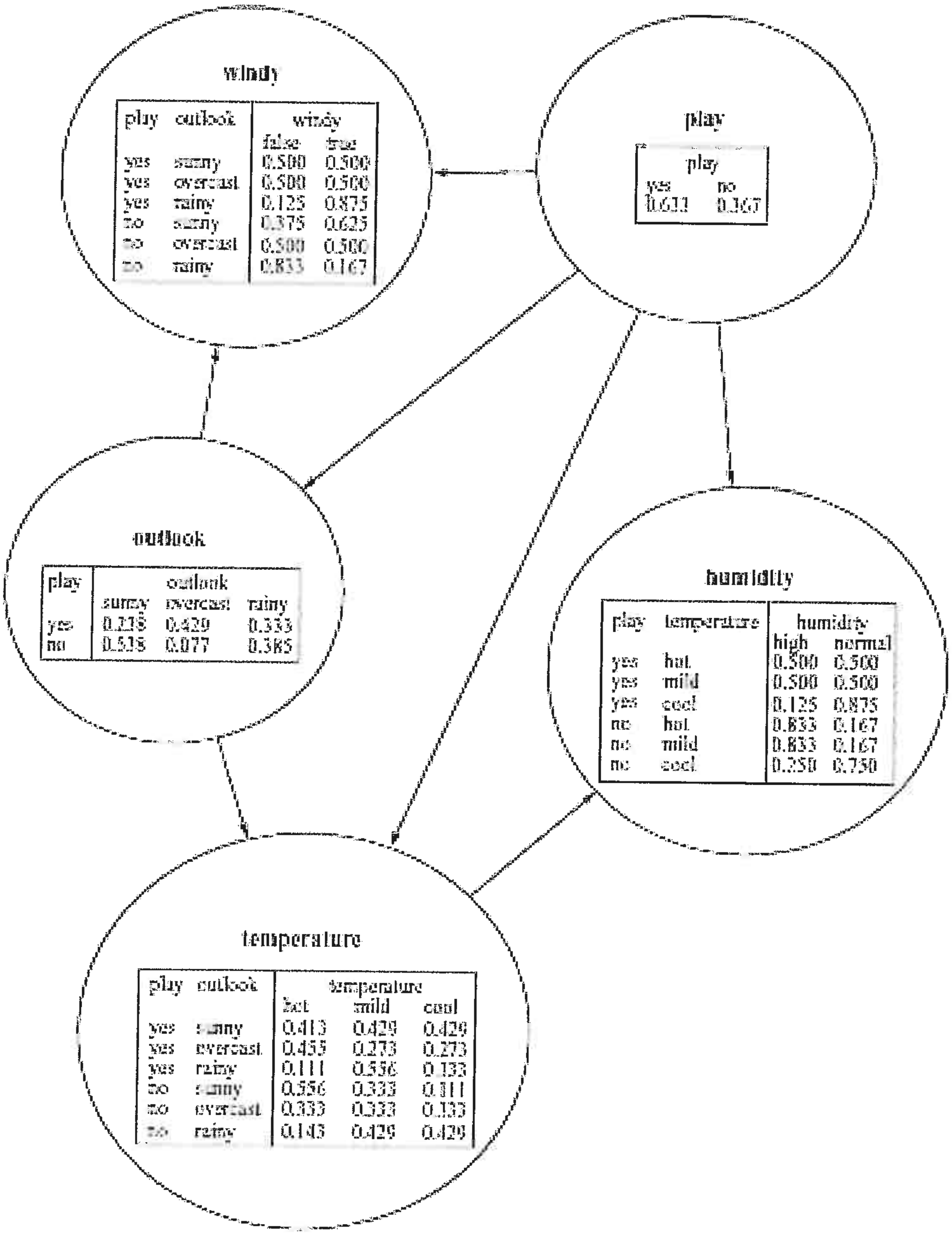
   a) Briefly describe instance based learning.

   b) What are main challenges of instance based learning? How to deal the challenge(s)?

Tingting Zhang
072 5188878

10. (10 p) Bayes classification

a. What is pr(play= yes, outlook = sunny)?
b. Given the following Bayes network. Find out:
Given outlook = sunny, temperature = hot, windy = true and humidity = normal, find out the play = yes or not ?



play	outlook	windy	
		false	true
yes	sunny	0.500	0.500
yes	overcast	0.500	0.500
yes	rainy	0.125	0.875
no	sunny	0.375	0.625
no	overcast	0.500	0.500
no	rainy	0.833	0.167

play	
yes	no
0.633	0.367

play	outlook		
	sunny	overcast	rainy
yes	0.238	0.429	0.333
no	0.538	0.077	0.385

play	temperature	humidity	
		high	normal
yes	hot	0.500	0.500
yes	mild	0.500	0.500
yes	cool	0.125	0.875
no	hot	0.833	0.167
no	mild	0.833	0.167
no	cool	0.250	0.750

play	outlook	temperature		
		hot	mild	cool
yes	sunny	0.413	0.429	0.429
yes	overcast	0.455	0.273	0.273
yes	rainy	0.111	0.556	0.333
no	sunny	0.556	0.333	0.111
no	overcast	0.333	0.333	0.333
no	rainy	0.143	0.429	0.429

Tingting Zhang
072 5188878

Index: distributions and formulas

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

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

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

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

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

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

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

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

Tingting Zhang
072 5188878

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

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

,

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

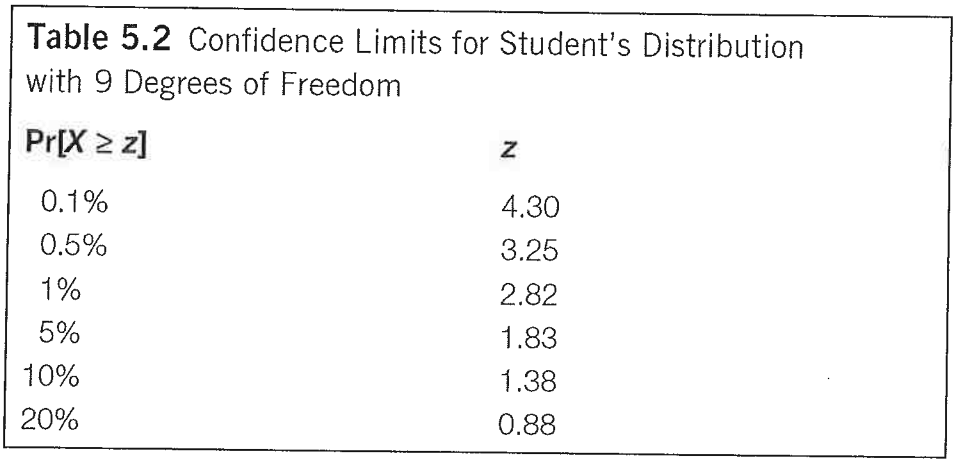


Table 5.2 Confidence Limits for Student's Distribution with 9 Degrees of Freedom

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

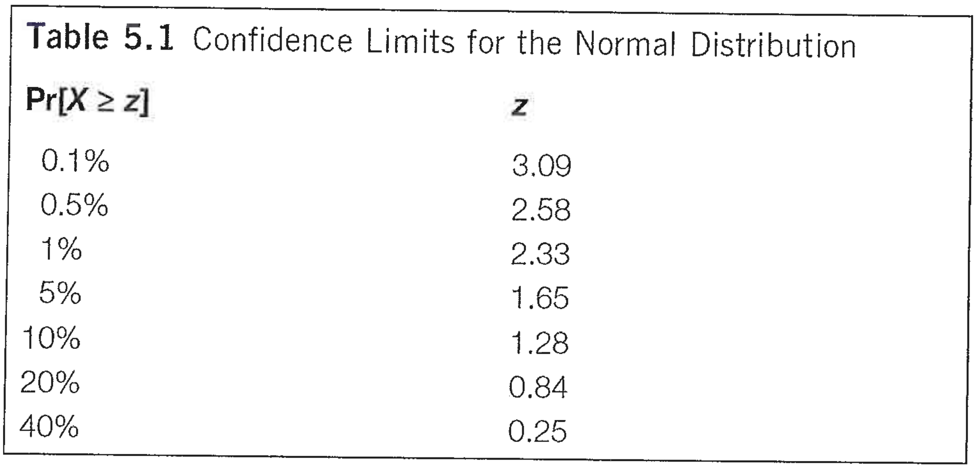


Table 5.1 Confidence Limits for the Normal Distribution

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