Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2018

Time: 2018-03-21

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) Briefly describe the process of data mining process.
2. (8 p) List out three normalization methods. In which cases, attributes should be normalized?
3. (8 p) In which situation, we need to transform Numeric data to nominal data. Briefly describe two transform methods.
4. (8 p) What is bootstrap? In which situation, bootstrap method is more preferred? Given 5000 instance, the success rate of testing data set is 80% and the success rate of training data is 90%, what is the estimated success rate?
5. (8p) Given the following 2 cost matrix and prediction accurate results model 1 and model 2.

Model 1		Predicted class		total
		yes	no	
Actual class	yes	TP = 100, cost = 0	FN= 50, cost =10	150
	no	FP= 20, cost = 5	TN = 50, cost =0	70

Model 2		Predicted class		total
		yes	no	
Actual class	yes	TP = 90, cost = 0	FN= 60, cost =10	150
	no	FP= 40, cost = 5	TN = 30, cost =0	70

What model is better? Why?

6. (20 p) Bayes

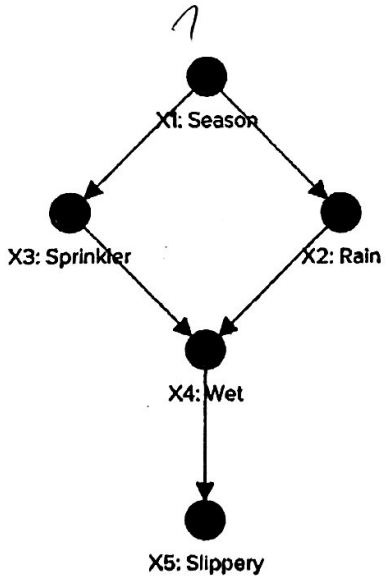
a) Based on the following training instances, Build a naïve Bayes classifier to classify the target variable.

b) From your classification model from above question, predict lens type given age is 34, astigmatic is no, tear reduction is normal,

Tingting Zhang
Tel: 0101428878

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	Normal	A
25	yes	Reduced	A
26	yes	Normal	A
29	yes	Normal	A
30	yes	Normal	A
31	No	Normal	A
33	No	Normal	A
40	No	Reduced	No
45	No	Reduced	B
45	No	Normal	B
46	No	Normal	B
47	No	Normal	A
48	No	Normal	A
51	yes	Reduced	No
59	yes	Reduced	A
70	yes	Normal	No

c) Given the following Bayes network. What is $P(x_2 = 1)$? What is $P(x_5 = 1 | x_1 = 1)$?



$P(x_3|x_1)$

$P(x_3 = 0|x_1 = 0) = 0.7$ $P(x_3 = 0|x_1 = 1) = 0.3$

Tingting Zhang
Tel: 0101428878

$P(x_3 = 1|x_1 = 0) = 0.3$ $P(x_3 = 1|x_1 = 1) = 0.7$

$P(x_2|x_1)$

$P(x_2 = 0|x_1 = 0) = 0.7$ $P(x_2 = 0|x_1 = 1) = 0.1$

$P(x_2 = 1|x_1 = 0) = 0.3$ $P(x_2 = 1|x_1 = 1) = 0.9$

$P(x_4|x_3, x_2)$

$P(x_4 = 0|x_2 = 0, x_3 = 0) = 0.8,$ $P(x_4 = 1|x_2 = 0, x_3 = 0) = 0.2$

$P(x_4 = 0|x_2 = 0, x_3 = 1) = 0.5,$ $P(x_4 = 1|x_2 = 0, x_3 = 1) = 0.5$

$P(x_4 = 0|x_2 = 1, x_3 = 0) = 0.4,$ $P(x_4 = 1|x_2 = 1, x_3 = 0) = 0.6$

$P(x_4 = 0|x_2 = 1, x_3 = 1) = 0.1,$ $P(x_4 = 1|x_2 = 1, x_3 = 1) = 0.9$

$P(x_5|x_4)$

$P(x_5 = 0|\, x_4 = 0) = 0.7,$

$P(x_5 = 1|x_4 = 0) = 0.3,$

$P(x_5 = 0|\, x_4 = 1) = 0.2,$

$P(x_5 = 1|x_4 = 1) = 0.8,$

$P(x1=0) = 0.2,\ P(x1=1) = 0.8$

7. (20 p) Association rules

   a) What is overfitting problem in the association rules finding? How to prevent overfitting?
   b) Briefly describe Prior method
   c) Given the following dataset, use prior method to find all association rules with coverage is 5 and accuracy is more than 80%.

Attribute variables			Target variable
age	Astigmatic	Tear production rate	**Lens type**
y	No	Normal	A
y	No	Reduced	A
y	No	Normal	A
y	No	Normal	A
y	yes	Normal	A
y	yes	Normal	A
y	yes	Reduced	A
y	yes	Normal	A
y	yes	Normal	A
y	yes	Normal	A
m	No	Normal	A
m	No	Normal	A
m	No	Reduced	No
m	No	Reduced	B
m	No	Normal	B
m	No	Normal	B
m	No	Normal	A
o	No	Normal	A
o	yes	Reduced	No
o	yes	Reduced	A
o	yes	Normal	No

Tingting Zhang
Tel: 0101428878

8. (20p) Instance based learning
   a) What is instanced based learning. When can the instance based learning be used? Give an example that the instance based learning is suitable. Given an example that instance based learning is not preferred.
   b) How to decide k in K-NN methods?
   c) Briefly describe two methods that can be used to reduce the sample dataset in instance based learning.

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