Mittuniversitetet
MID SWEDEN UNIVERSITY

Försättsblad Prov Original

Kurskod	Provkod	Tentamensdatum
DT044A	T101	2018-08-29
Kursnamn	Datateknik AV, Datamining	
Provnamn	Tentamen	
Ort	Sundsvall	
Termin	H18	
Ämne	Datateknik	

8982538474

0026841

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2018

Time: 2018-08-29

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. (8 p) In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe 3 methods for handling this problem.
3. (8 p) In which situation, we need to transform Numeric data to normal data. List out the methods of transforming numeric data to normal data.
4. (8 p) Explain the problem of Overfitting, and how to prevent it in association rule learning and decision tree learning.
5. (8 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 sets for the two different learning schemes are same and from same domain Find out if one scheme is better than other one in confidence level of 80% (the probability of the one scheme is better than other one is bigger than 80%).

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

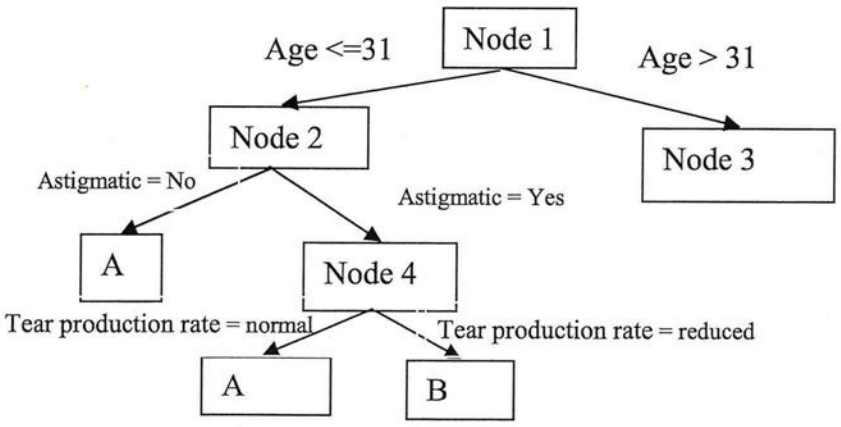
Tingting Zhang
Tel: 0101428878

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) Given the following instances and unfinished decision tree.

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



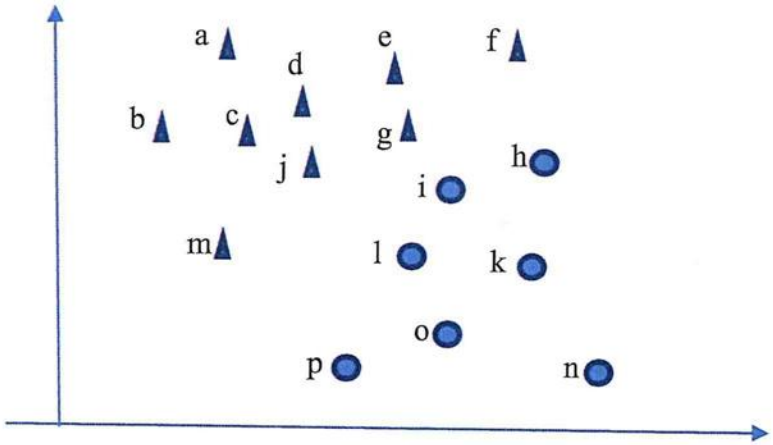
(i) Finish build the tree

(ii) Post pruning the tree based on train dataset with confidence = 80% (z = 0.25).

Tingting Zhang
Tel: 0101428878

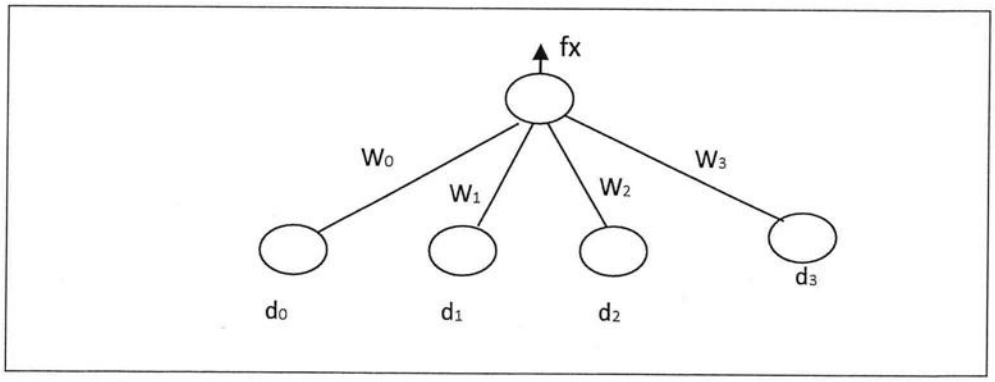
7. (20 p) Naïve Bayes Classify
   a. Briefly describe the principle of Naïve Bayes classify. What is the basic limitation of Naïve Bayes classify?
   b. Build a naïve Bayes classifier to classify the target variable from the above table in question 6.
   c. Use your Bayes classify to predict the target variable of age = 41, Astigmatic = yes and Tear production rate = normal.

8. (20 p) Linear regression
   a. Given the following graph. What are support vectors?



   b. What is logistic function that is used in linear regression? Why we need to use this logistic function?

   c. Given the following single layer neural network. Suppose we know the output of the network is 5, real value is 4, and learning rate is 0.5, how to change the weight w_1?



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