Mittuniversitetet
MID SWEDEN UNIVERSITY

Försättsblad Prov Original

Kurskod	Provkod	Tentamensdatum
DT047A	T101	2019-03-22
Kursnamn	Datateknik AV, Datamining	
Provnamn	Tentamen - Sundsvall	
Ort	Sundsvall	
Termin	VT2019	
Ämne	Datateknik	

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2019

Time: 2019-03-22

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 data mining process

2. (8 p) For supervised learning what kind of attributes should be included in the learning process?
3. (8 p) Is 10 times repeated holdout same as tenfold cross validation? (why?)
4. (8p) Given the following cost matrix and prediction accurate results model 1 and model 2.
   a) What is success rate of model 1? What is success rate of model 2?
   b) Which one is better? Why?

Model 1		Predicted class		total
		yes	no	
Actual class	yes	TP = 50, cost = 0	FN= 50, cost =10	100
	no	FP= 20, cost = 3	TN = 80, cost =0	100

Model 2		Predicted class		total
		yes	no	
Actual class	yes	TP = 70, cost = 0	FN= 30, cost =10	100
	no	FP= 60, cost = 3	TN = 40, cost =0	100

5. (8 p) What is bootstrap? Given that there are 2500 instance in dataset for bootstrap.
   - Model 1: The success rate of testing data set is 70% and the success rate of training data is 80%
   - Model 2: The success rate of testing data set is 75% and the success rate of training data is 85%

   Given confidence limit c = 60%. Which model is better? Why?

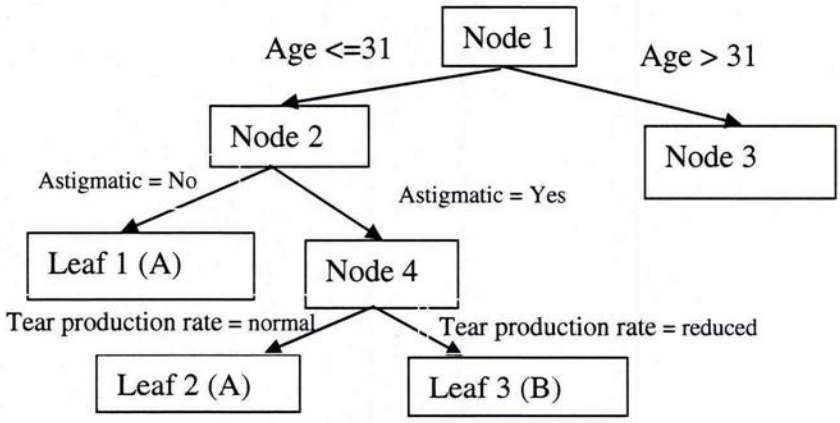
Tingting Zhang
Tel: 0101428878

6. (20 p) decision tree

a) Suppose that a decision tree is build based on a training dataset so that every leaf of the tree is 100% correct for the training dataset. Is this always 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
48	yes	Reduced	B
60	yes	?	A



(i) Which attribute will be used to split node 3?

(ii) Given confidence = 25% (z = 0.32). Can leaf 2 and Leaf 3 be pruned away?

Tingting Zhang
Tel: 0101428878

7. (20 p) clustering
   a. When can the clustering method be used?

   b. Suppose that the data mining task is to cluster points (with .x, y/ representing location) into 2 clusters, where the points are A1 (0,0), A2 (2,5), A3 (0,4), B1 (2,0), B2 (7,5), B3 (6,4), C1 (1,2), C2 (4,9).

      The distance function is Euclidean distance. Suppose initially we assign A1, B1 as the center of each cluster, respectively. Use the k-means algorithm to show:
      i) The two cluster centers after the first round of execution.
      ii) The final two clusters.

   c. Does the cluster method always lead to optimal solution? If it is not how to find a better solution?

8. (20 p) Linear regression

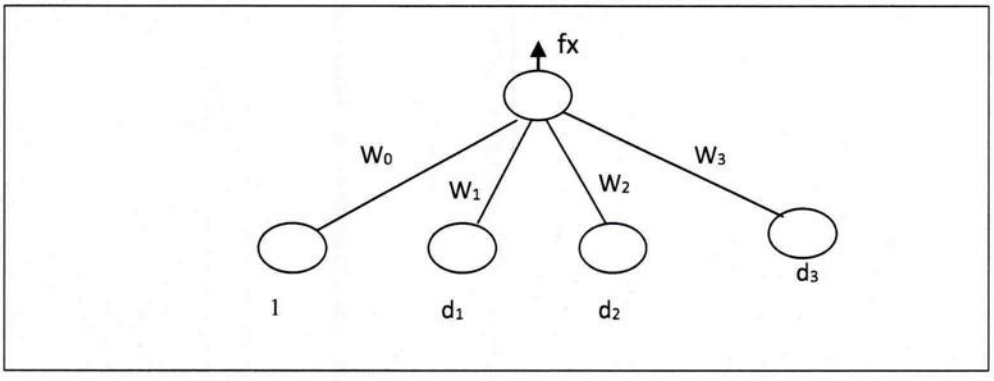
   a. Name two methods of extend linear regression to solve nonlinear problem.

   b. Consider perceptron learning rule in the training data set for the Boolean function y = NOT. Assign 1 to initial weights and bias. Use the perceptron learning rule method to learn weight w_0 (for bias) and w_1, w_2 in one round.

X1	X2	Y (target)
0	0	-1
2	0	-1
0	1	-1
2	1	1
1	2	1

(hint: $f(x) = w_0 + w_1x_1 + w_2x_2$. if $f(x) > 0$, then belong to class Y = 1, else belong to Y = -1 class)

   c. Given the following single layer neural network. The output node use sigmoid function as activation function Suppose we know the output of the network is 0.8, real value is 1, $d_1=2$, $d_2=3$, $d_3= 0$ and learning rate is 0.5. How to change the weight w_1?



Tingting Zhang
Tel: 0101428878

Distributions and formulas

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

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

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

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

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

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

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

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

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