Mittuniversitetet
MID SWEDEN UNIVERSITY

Försättsblad Prov Original

Kurskod	Provkod	Tentamensdatum
DT044A	T101	2018-06-15
Kursnamn	Datateknik AV, Datamining	
Provnamn	Tentamen	
Ort	Sundsvall	
Termin	V18	
Ämne	Datateknik	

8982538474

0026772

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2018

Time: 2018-06-15

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) What is supervised learning? What is unsupervised learning? What is semi-supervised learning?
2. (8 p) Briefly describe two attribute selection approach; Filter and Wrapper. Compare these two methods.
3. (8 p) How to divided test and training data in holdout method, so that the error rate can be correctly estimated.
4. (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 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 level of 80% (the probability of the one scheme is better than other one is bigger than 80%).

Scheme 1	80%	80%	85%	80%	80%	75%	90%	90%	70%	90%
Scheme 2	90%	95%	90%	95%	80%	85%	90%	80%	80%	95%

5. (8p) Given the following sample table suppose after learning concludes that the probability of response is pr(person) = 100 – person.age. What is lift fact of 10% sample? What is lift fact 20% sample? Draw first 5 points of a ROC curve (x-axis: false positive rate, y-axis: true positive rate).

Customer Name	address	Age	Actual Response
Alan	C	39	N
Alex	C	18	Y
Amy	C	18	N
Bob	B	21	Y
Catherine	B	23	N
Chris	D	42	N
Elizabeth	A	30	Y
Erin	C	35	Y
Fred	A	45	N
Hilary	A	19	Y
Jessica	B	25	Y
John	C	70	N
Kim	D	33	N
Laura	A	29	Y
Margot	C	51	N
Nancy	B	24	Y
Philip	A	20	Y
Preston	A	25	N
Sean	D	65	Y
Trent	A	55	N

Tingting Zhang
Tel: 0101428878

6. (20 p) Decision tree
   a. Briefly outline the major steps of decision tree classification.
   b. What is the purpose of pruning a tree? What is pre pruning? What is post pruning?
   c. Given the below mushroom data set.
   What is information gain for each attributes? Which attribute is the best one to split the root? What is the problem of using information gain?

attributes				**Target class**
Colour	**Height**	**Stripes**	**Texture**	**Poisonous**
Red	10	Yes	Hairy	No
Blue	10	No	Smooth	No
Blue	12	Yes	Hairy	Yes
Blue	14	Yes	Smooth	Yes
Blue	14	Yes	Rough	Yes
Red	18	No	Smooth	No
Purple	18	?	Hairy	Yes
Purple	20	Yes	Rough	Yes
Purple	22	Yes	Smooth	Yes
Red	23	No	Hairy	No
Blue	24	Yes	Smooth	Yes
Blue	?	No	Hairy	No
Red	31	Yes	Hairy	No
Purple	33	Yes	Hairy	Yes
Purple	34	No	Rough	No
Purple	35	No	Smooth	No

7. (20 p) Regression

   a) Can linear regression learn un-linear mode? If cannot name two methods to learn un-linear model using linear regression.

   b) Briefly describe support vector method (SVM).

   c) Consider perceptron learning rule in the training data set for the following table. Assign 0 to initial weights and bias. Use the learning method to learn weight w_0 (for bias) , w_1 for x and w_2 for y.

Tingting Zhang
Tel: 0101428878

x	y	Target
1	0	no
1	2	yes
-1	5	yes
0	1	no

8. (20p) K-means
   a) When will the clustering method be used? What is the main challenges of clustering methods?
   b) Suppose that the data mining task is to cluster the following points (with .x, y/ representing location) into two clusters.

	x	y
A1	1	0
A2	-1	-1
A3	2	0
A4	2	1
A5	3	1
A6	-3	-1
A7	-2	0
A8	-1	-3
A9	1	3

The distance function is Euclidean distance. Suppose initially we assign A1, A5, as the center of each cluster, respectively. Use the k-means algorithm to show only

- The two cluster centers after the first round of execution.
- The final two clusters center and custers.

c) Why the *k*-means algorithm may not find the global optimum? How to increase the chance of finding global optimum?

Tingting Zhang
Tel: 0101428878

reference: 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