DT044A

Tingting Zhang

Examination of Data Mining, AV 2015

Time: 2015-04-24

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

1. (10 p) What is supervised data mining? List out four kinds data mining that can be used in supervised learning

2. (5 p) How to divided test and training data in holdout method, so that the error rate can be correctly estimated.

3. (5 p) Name two main attribute selection methods.

4. (5 p) Given the following table suppose after learning concludes that the probability of true is 5%.

    a. What is lift fact of top 10% instance with 20% true?
    b. What is lift fact 20% sample with 10% true?

5. (5 p) Name two methods of evaluate numeric prediction learning result.

6. (5 p) How to select a learning scheme?

7. (5 p) Describe one method of data discretization of one nominal attribute color with 3 values.

Tingting Zhang

8. (25 p) Classification

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
number of O-rings	**Launch temperature**	**Leak-Check Pressure**	**Number of O-rings with Stress**
1	low	50	0
2	low	50	0
3	low	50	0
4	low	50	0
5	low	50	0
6	high	50	1
7	high	100	0
8	high	100	0
9	high	100	0
10	low	150	1
11	low	150	1
12	low	150	1
13	low	150	1
14	low	150	0
15	low	200	1
16	high	200	2
17	high	200	2
18	high	200	2
19	high	200	2
20	high	200	1
21	high	200	2

b. What is the problem of using gain to decide the splitting attribute?
c. Generate classification rules from the decision tree.
d. Use the same dataset to evaluate confidence interval of the error rate of the classification, given confidence limit (z) is 10%

Tingting Zhang

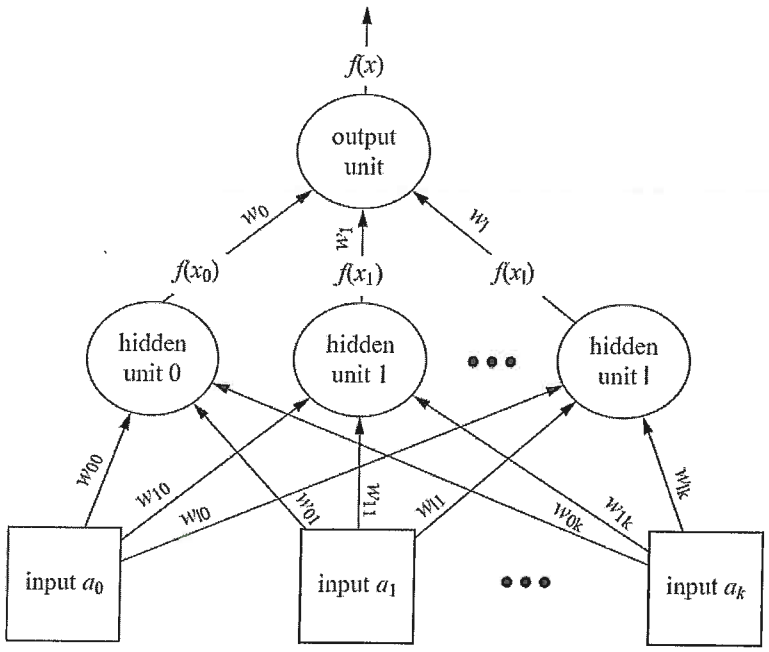
9. (20 p) Instance based learning method
   a) Briefly describe instance based learning.
   b) Give three distance functions that used in instance based learning.
   c) What are main challenges of instance based learning? How to deal the challenge(s)?

10. (15 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 round.

x	y	Target
0	0	no
0	1	yes
1	1	yes
1	0	Yes

   b) Consider the following multilayer receptron. Suppose we know the difference between f(x) and real value is δ (f(x) – ReavlValue) = δ) , and learning rate is α, how to change the weight w_0, and w_{00}



Tingting Zhang

Index: distributions and formulas

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

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

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

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

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

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

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

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

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 ≥ 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 ≥ 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