DTOYYA

Tingting Zhang
Tel: 0101428878

Examination of Data Mining, AV 2017

Time: 2017-05-29

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. (10 p) Briefly describe the data mining process.
2. (10 p) Why normalization should take place in data preprocessing? What are the value ranges of the following *normalization methods*?
   (a) min-max normalization
   (b) z-score normalization
   (c) Normalization by decimal scaling
3. (5 p) How to divided test and training data in holdout method, so that the error rate can be correctly estimated.
4. (7 p) What is bootstrap? Given the success rate of testing data set is 70% and the success rate of training data with 10000 instance is 90%, what is the estimated success rate? What is the standard deviation of the success rate? What is the lower band of the success rate given confidence limit 20%.
5. (8 p) 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 = 80, cost = 0	FN= 30, cost =10	110
	no	FP= 20, cost = 5	TN = 70, cost =0	90

Model 2		Predicted class		total
		yes	no	
Actual class	yes	TP = 70, cost = 0	FN= 40, cost =10	110
	no	FP= 10, cost = 5	TN = 80, cost =0	90

   a) Which model will give better cost sensitive prediction?
   b) What are success rates for model 1 and 2?

6. (20 p) Association rules
   (i) What is the difference of classification rules and association rules?
   (ii) What is the main challenges of find association rules?

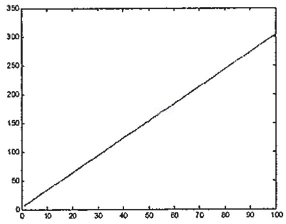
   (iii) Consider 16 data records in the testing data set of in following table. Use Apriori algorithm find all 2 item set, 3 item set and 4 item set with *minimum cover* = 35% (5 instances)

(iv) Use the frequent item sets from these item sets to generate all the association rules that satisfy *min-cover*= 35% and *min-accurate* = 50%

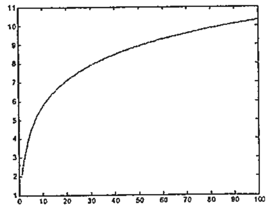
	Attribute variables			Target variable
	color	**size**	**Act**	**inflated**
1	yellow	small	stretch	True
2	yellow	small	stretch	True
3	yellow	small	Dip	True
4	yellow	large	Dip	True
5	yellow	large	Dip	False
6	yellow	large	Dip	false
7	yellow	small	Dip	false
8	yellow	small	Dip	True
9	purple	small	stretch	False
10	purple	small	stretch	false
11	purple	small	stretch	false
12	purple	large	Dip	false
13	purple	large	Dip	true
14	purple	large	Dip	true
15	purple	large	Dip	false
16	purple	large	Dip	false

7. (20 p) Linear regression
   (i) Briefly describe linear regression. What is logistic function that is used in linear regression? Why we need to use the logistic function?
   (ii) Given the following datasets, can single layer neural network be used to find the model?

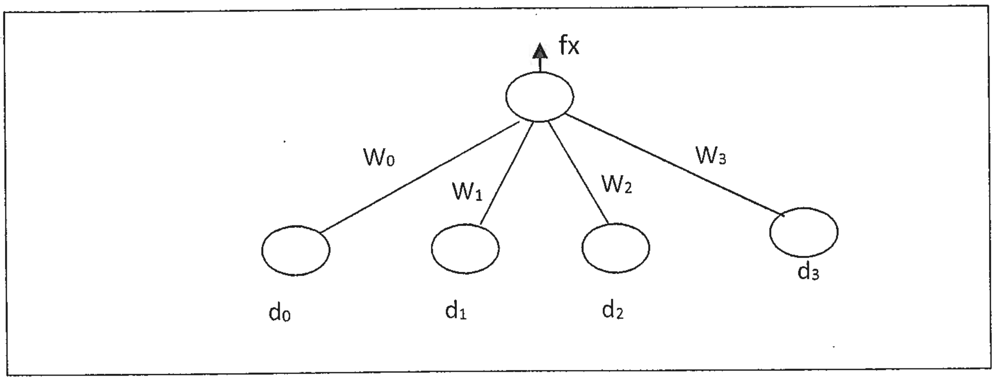
Dataset 1



Dataset 2



(iii) Given the following single receptron network. Suppose we know the output of the network is 0.9, real value is 0.4, and learning rate is 2, how to change the weight w_1?



8. (20 p) clustering
   - i) In which case, clustering can be used? Give an example to explain.
   - ii) Briefly describe the method of clustering a data set into k clustering.
   - iii) Briefly describe hierarchical clustering. How to measure distance between two clusters? Briefly describe two methods.

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))

$$\frac{mean_x - \mu}{\sqrt{\sigma_x^{\,2} / k}},$$

$$\frac{mean_d}{\sqrt{\sigma_d^{\,2} / k}}$$

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$U(A,B) = \frac{\sum_i \sum_j (a_i - a)(b_j - b)}{\sqrt{\left(\sum_i (a_i - a)^2\right)\left(\sum_i (b_i - b)^2\right)}}$$

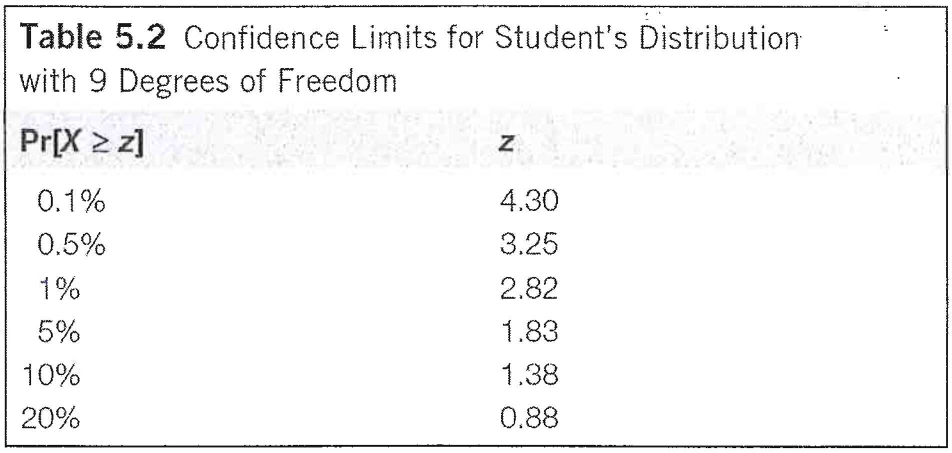


Table 5.2 Confidence Limits for Student’s Distribution with 9 Degrees of Freedom

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

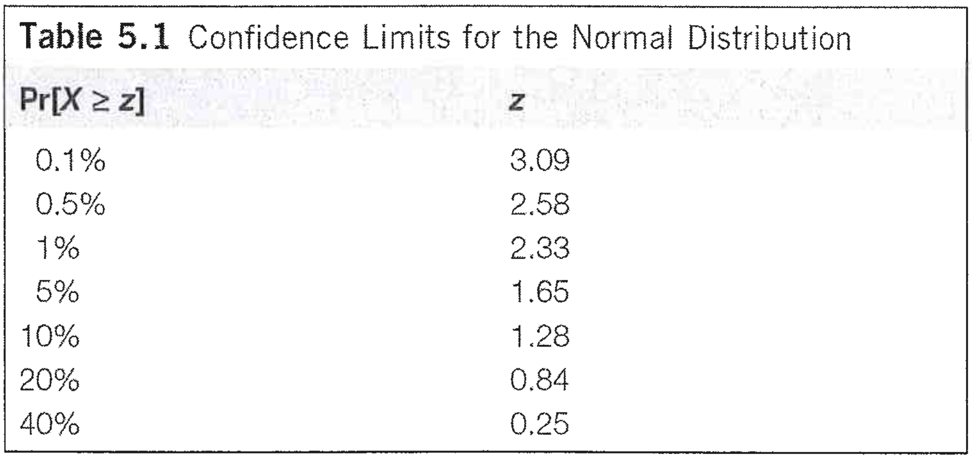


Table 5.1 Confidence Limits for the Normal Distribution

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