DT044A

Tingting Zhang

Examination of Data Mining, AV 2015

Time: 2015-03-16

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 data mining? What knowledge areas are included in the data mining?

2. (5 p) Briefly describe data mining process.

3. (5 p) Briefly describe the wrapper attribute selection method.

4. (5 p) How to select instance training set and test set in holdout?

5. (5 p) Is 10 times repeated holdout same as tenfold cross validation? (why?)

6. (5) Suppose the following table present two set of mean success rate obtained by five 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 limit of 10%.

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

7. (5 p) What is bootstrap? Given the success rate of testing data set is 70% and the success rate of training data is 80%, what is the estimated success rate?

Tingting Zhang

8. (20 p) Classification

a. Build a decision tree for the following balloon 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	1
2	low	50	1
3	low	50	0
4	low	50	0
5	low	50	0
6	high	50	0
7	high	100	0
8	high	100	0
9	high	100	0
10	low	150	1
11	low	150	1
12	low	150	0
13	low	150	0
14	low	150	0
15	low	200	2
16	high	200	0
17	high	200	0
18	high	200	0
19	high	200	0
20	high	200	0
21	high	200	0

b. Is it good to build a decision tree that is perfect fit the training data set? Why?

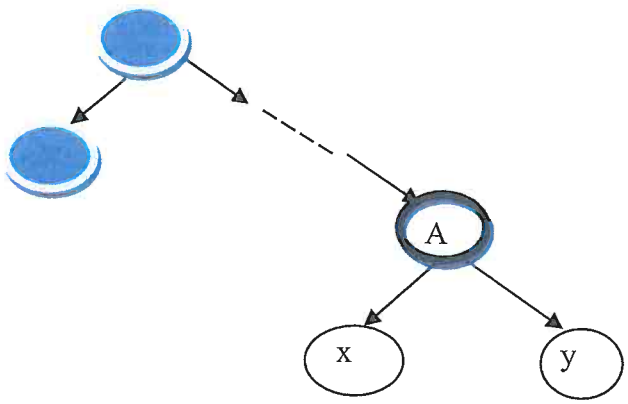
c. Suppose we have following information after running test data on the following decision tree

i. 100 test instances in node x, and 10% predict error

ii. 20 test instances in node y, with 5% predict error

iii. 120 test instances in nod A with 10% error

Should node x and y be pruned away?



Tingting Zhang

9. (20 p) Statistic model

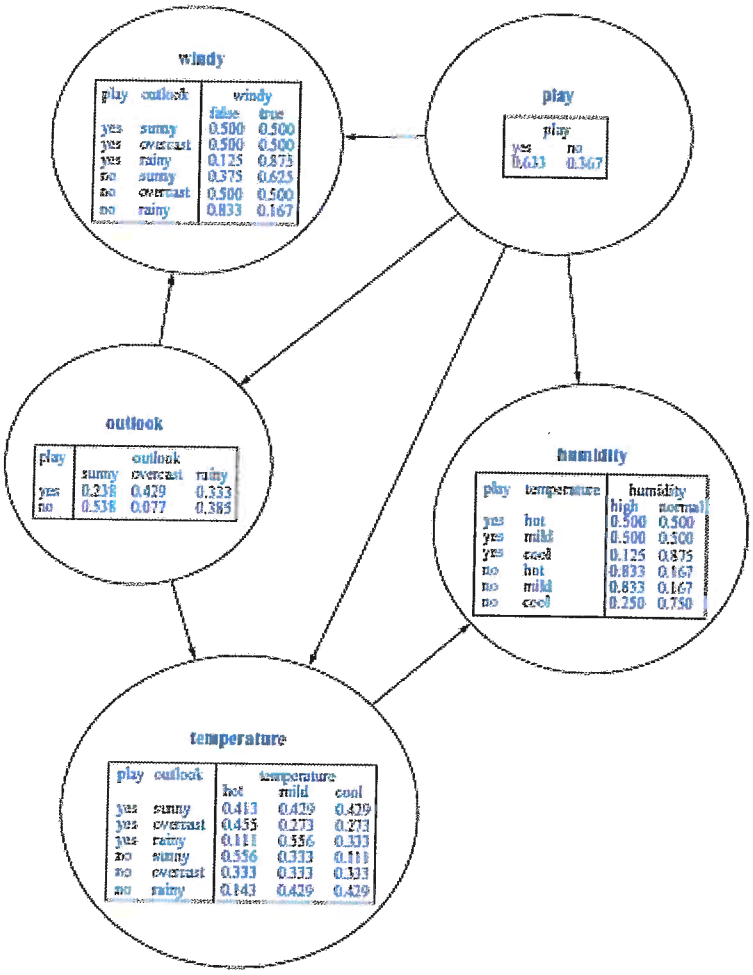
a) What is naïve bayes model? What is the main difference between bayes network and naïve bayes model?

b) Given the following bayes network. Find out:

a. Pr(outlook = sunny)

b. Pr(temperature = hot)

c. Given outlook = rainy, temperature = hot, windy = false and humidity = normal, find out if play = yes or not?



Tingting Zhang

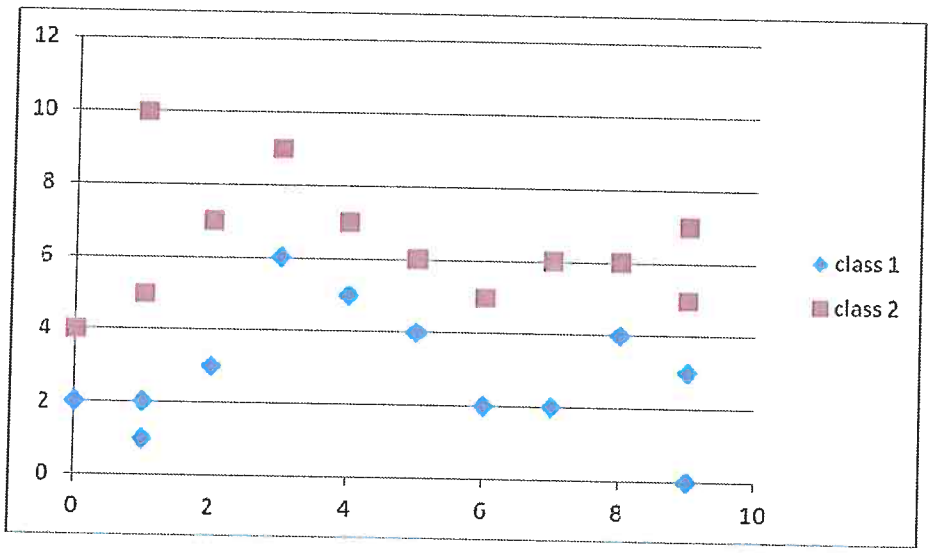
1O. (20 p) Linear model

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.

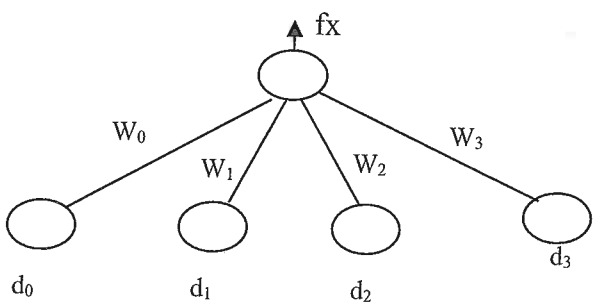
x	y	Target
0	0	yes
0	1	no
1	1	yes
1	-1	Yes

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

c) Given the following dataset, can single layer receptron be used to find the model? List two methods to solve this.



d) Given the following single receptron network. Suppose we know the output of the network is fx, difference between f(x) and real value is δ, and learning rate is α, how to change the weight w_1?



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