Data Mining: Homework #3

Due on November 30, 2014 at 5:00pm

Professor Y L 712009H

Libaier

1234567

Problem 1

Cluster the following 8 points into three clusters:
A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9)
The distance function is Euclidean distance. Suppose initially we assign A1, A4, A7 as the center of each cluster, respectively. Use the K-Means algorithm to show only
(1)The three cluster center after the first round execution

Solution

Step 1.

$$\begin{aligned} K &= 3 \\ Z_1(1) = A1 &= (2\ 10)^t \\ Z_2(1) = A4 &= (5\ 8)^t \\ Z_3(1) = A7 &= (1\ 2)^t \end{aligned}$$

Step 2.

$$\begin{aligned} ||Ai - Z_1(1)|| &< min(||Ai - Z_2(2)||, ||Ai - Z_2(2)||), i = 1 \\ ||Ai - Z_2(1)|| &< min(||Ai - Z_1(2)||, ||Ai - Z_3(2)||), i = 3,4,5,6,8 \\ ||Ai - Z_3(1)|| &< min(||Ai - Z_1(2)||, ||Ai - Z_2(2)||), i = 2,7 \end{aligned}$$

$$\begin{aligned} S_1(1) &= \{A1\} \\ S_2(1) &= \{A3, A4, A5, A6, A8\} \\ S_3(1) &= \{A2, A7\} \end{aligned}$$

Step 3.Calculate the new cluster center

$$\begin{aligned} Z_1(2) &= \frac{1}{N} \sum_{x \in S_1(1)} x = (2\ 10)^t \\ Z_2(2) &= \frac{1}{N} \sum_{x \in S_2(1)} x = (6\ 6)^t \\ Z_3(2) &= \frac{1}{N} \sum_{x \in S_3(1)} x = (1.5\ 3.5)^t \end{aligned}$$

(2)The final three clusters

Solution

$$\begin{aligned} S_1(1) &= \{A1, A8, A4\} \\ S_2(1) &= \{A3, A5, A6\} \\ S_3(1) &= \{A2, A7\} \end{aligned}$$

Problem 2

Perform AGNES clustering on the data set in Question 1. Show your results by drawing a dendrogram (each step merges two clusters with the minimum distance). The dendrogram should clearly show the order in which the points are merged.

Solution

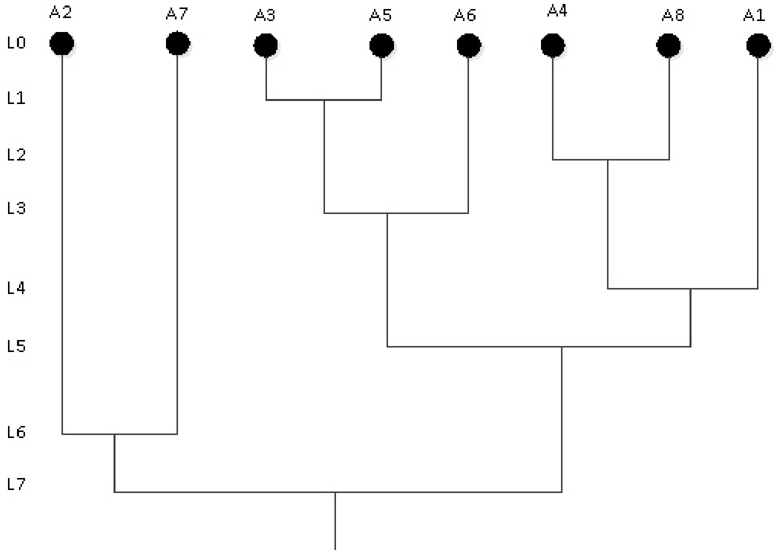


Figure 1: The dendrogram.

Problem 3

1. Production Recommendation The data contains the following fields. Row 1 contains the headers.

Table 1: Production Recommendation

id	a unique identification number
age	age of customer in years
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer
married	is the customer married (YES/NO)
children	number of children
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP after the last mailing (YES/NO)

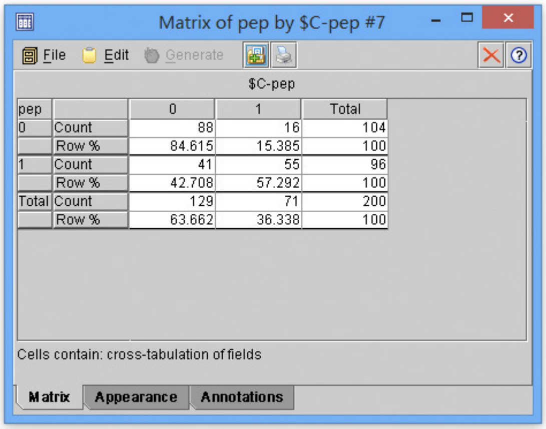
Each record is a customer description where the ”pep” field indicates whether or not that customer bought a PEP. For other existing customers in the database, we would like to see if PEP should be RECOMMENDED to the customers in the roll-out data.

The firm decides to use decision tree to build the models for PEP recommendation. Develop a decision tree model using the estimation data. For building this model, you are expected to use the following steps.

Using the bank-estimation-data, estimate the decision tree that predicts pep as a function of the other variables. Select Expert and set pruning severity at 75. Set the Type of pep as Flag and the Direction as out. Build decision trees using three options Minimum records per child branch values being (a) 56, (b) 15 and (c) 10, not selecting use global pruning.

1) Hand in: the confusion matrix for (a), (b) and (c) on the validation data.

Solution



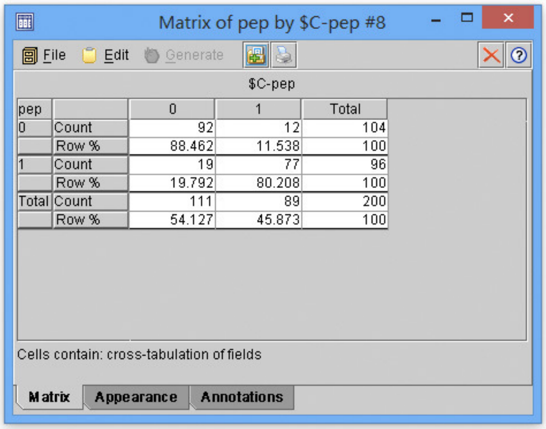
$C-pep

pep		0	1	Total
0	Count	88	16	104
	Row %	84.615	15.385	100
1	Count	41	55	96
	Row %	42.708	57.292	100
Total	Count	129	71	200
	Row %	63.662	36.338	100

Cells contain: cross-tabulation of fields

Figure 2: confusion matrix for (a).

Problem 3 continued on next page. . .

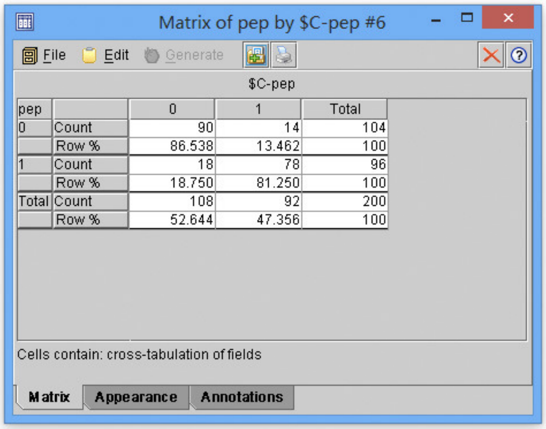


$C-pep

pep		0	1	Total
0	Count	92	12	104
	Row %	88.462	11.538	100
1	Count	19	77	96
	Row %	19.792	80.208	100
Total	Count	111	89	200
	Row %	54.127	45.873	100

Cells contain: cross-tabulation of fields

Figure 3: confusion matrix for (b).



$C-pep

pep		0	1	Total
0	Count	90	14	104
	Row %	86.538	13.462	100
1	Count	18	78	96
	Row %	18.750	81.250	100
Total	Count	108	92	200
	Row %	52.644	47.356	100

Cells contain: cross-tabulation of fields

Figure 4: confusion matrix for (c).

2) Hand in: Which of the three trees will you use to score the data in a holdout data list and why? 2-3 lines

Solution

I will use the tree with minimum records per child branch of 15 for it's high recognition rate of the 1-1 state(The correct prediction for the one who will buy PEP).

3) Hand in: For the following data (Appendix 1), using the rules from the best decision tree, fill in the recommendation.

Solution

Table 2: Appendix 1

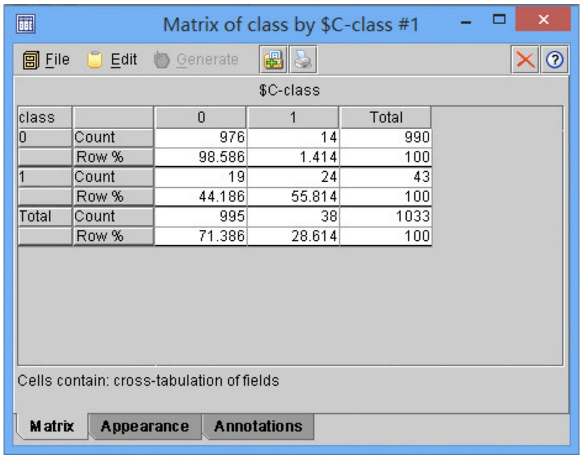
region	income	married	children	car	save_act	current_act	mortgage	RECOMEND PEP
1	14000	0	3	0	1	1	0	F
0	33000	0	0	1	1	0	0	T
0	16700	1	1	0	1	1	0	F
1	43400	1	1	1	1	1	0	T
2	60000	1	1	0	1	1	0	T
0	27700	0	1	1	0	0	0	T
0	38784	1	0	0	1	1	0	F
0	10200	1	0	0	1	1	1	F
0	22000	1	1	1	1	0	1	T
1	37400	1	2	0	1	1	0	T

Problem 4

The goal of this assignment is to introduce churn management using decision trees, logistic regression and neural network. You will try different combinations of the parameters to see their impacts on the accuracy of your models for this specific data set. This data set contains summarized data records for each customer for a phone company. Our goal is to build a model so that this company can predict potential churners. Two data sets are available, churn_training.txt and churn_validation.txt. Each data set has 21 variables. They are:
State: Account_length: how long this person has been in this plan
Area_code:
Phone_number:
International_plan: this person has international plan=1, otherwise=0
Voice_mail_plan: this person has voice mail plan=1, otherwise=0
Number_vmail_messages: number of voice mails
Total_day_minutes:
Total_day_calls:
Total_day_charge:
Total_eve_minutes:
Total_eve_calls:
Total_eve_charge:
Total_night_minutes:
Total_night_calls:
Total_night_charge:
Total_intl_minutes:
Total_intl_calls:
Total_intl_charge:
Number_customer_service_calls:
Class: churn=1, did not churn=0
Each row in churn_training represents the customer record. The training data contains 2000 rows and the validation data contains 1033 records.

1)Perform decision tree classification on training data set. Select all the input variables except state, area_code, and phone_number (since they are only informative for this analysis). Set the Direction of class as out, type as Flag. Then, specify the minimum records per child branch as 30, pruning severity as 60, click use global pruning. Hand-in the confusion matrices for validation data.

Solution



$C-class

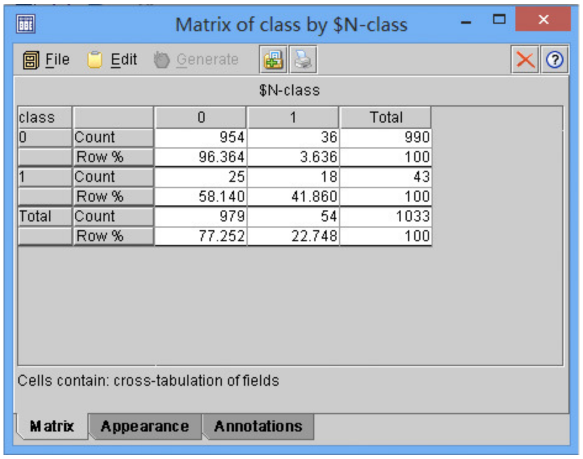
class		0	1	Total
0	Count	976	14	990
	Row %	98.586	1.414	100
1	Count	19	24	43
	Row %	44.186	55.814	100
Total	Count	995	38	1033
	Row %	71.386	28.614	100

Cells contain: cross-tabulation of fields

Figure 5: confusion matrix for decision tree.

2)Perform neural network on training data set using default settings. Again, select all the input variables except state, area_code, and phone_number. Hand-in the confusion matrix for validation data.

Solution



$N-class

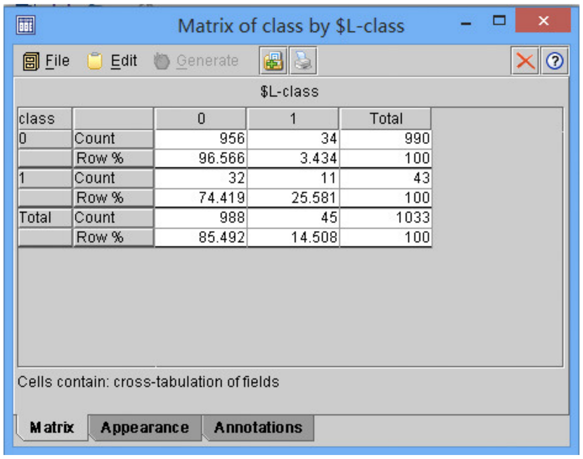
class		0	1	Total
0	Count	954	36	990
	Row %	96.364	3.636	100
1	Count	25	18	43
	Row %	58.140	41.860	100
Total	Count	979	54	1033
	Row %	77.252	22.748	100

Cells contain: cross-tabulation of fields

Figure 6: confusion matrix for neural network.

3)Perform logistic regression on training data set using default settings. Again, select all the input variables except state, area_code, and phone_number. Hand-in the confusion matrix for validation data.

Solution



Matrix of class by $L-class

$L-class

class		0	1	Total
0	Count	956	34	990
	Row %	96.566	3.434	100
1	Count	32	11	43
	Row %	74.419	25.581	100
Total	Count	988	45	1033
	Row %	85.492	14.508	100

Cells contain: cross-tabulation of fields

Figure 7: confusion matrix for logistic regression

4)Hand-in your observations on the model quality for decision tree, neural network and logistic regression using the confusion matrices.

Solution

The confusion matrices for decision tree,neural network,logistic regression is shown in the Figure 5,6,7.Through my observations I found that decision tree get the best result.And the result of neural network is changed for different training and testing process.