CIS8695

Homework III

What to Turn In:

Please write your answers and paste required results/reports in **a MS Word file**. **Note: Please do NOT submit any other file format, because it will cause grading inconveniency. Failing to submit in correct file format will cause the loss of homework grades! You can use screenshots if it is convenient for you.**

Problem 1 (CART): Competitive auctions on eBay.com:

The file **eBayAuctions.csv** contains information on 1972 auctions that transacted on eBay.com during May-June in 2004. The goal is to use these data in order to build a model that will classify competitive auctions from non-competitive ones. A *competitive auction* is defined as an auction with at least 2 bids placed on the auctioned item. The data include variables that describe the auctioned item (auction category), the seller (his/her eBay rating) and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price that the auction closed at. The goal is to predict whether the auction will be competitive or not.

1. Note that in the dataset, the original variables of **Category** (11 categories), **Currency** (USD, nonUS), and **EndDay** (Weekend, Week) are categorical variables. Therefore, the dataset also contains their corresponding dummy variables.
2. Import the dataset and split the data into training and validation datasets using a 60%-40% ratio.
3. Fit a classification tree. Use ***Competitive*** as the output variable and the rest of variables as predictors. In the model, make sure that you exclude one dummy variable from each group of dummy variables (e.g. exclude Category_SportingGoods, Currency_nonUS and EndDay_Weekend). To avoid overfitting, set the **maxdepth=6**.
    a. Report the tree (copy and paste the tree diagram).
    b. Report the prediction Confusion Matrix of Validation Data.
    c. What predictors are used by the tree?
    d. List the decision rules. For example, if variable1<0 AND variable2<2, class=0.
4. Are the rules practical for predicting the outcome of a new auction? Explain why (Hint: are you able to use the rules to classify a new auction before the auction ends? Do you know the values of all predictors in the rules before the auction ends? Some of them may not be known before the end of auction. What are them?). What variables should **NOT** be included in the predictor set? Explain why.
5. Fit another classification tree using the same setting in question 3. This time only use the predictors that can be used for predicting the outcome of a new auction.

a. Report the tree (copy and paste the tree diagram).
b. Report the prediction Confusion Matrix of Validation Data.
c. What predictors are used by the tree?
d. List the decision rules.

6. Examine and compare the summary reports in questions 3 and 5. Compare the overall performance (e.g., accuracy or error rates) between these two decision trees. Which model has better predictive performance? Explain why.

7. Build a random forest model for this prediction problem. Report:

a. Variable importance.
b. The prediction Confusion Matrix of Validation Data.

Problem 2. (Naïve Bayes Classifier).

Look at Problem 8.2 in Chapter 8 of the Textbook, page 202-203. (A picture of the problem is attached below). Complete (a), and (c.i)-(c.iv) of Problem 8.2.

etc.), and the custo... Among these 5000 customers, only ... offered to them in the earlier campaign. In this exe... predictors: Online (whether or not the customer is an active user of online banking services) and Credit Card (abbreviated CC below) (does the customer hold a credit card issued by the bank), and the outcome Personal Loan (abbreviated Loan below).

Partition the data into training (60%) and validation (40%) sets.

a. Create a pivot table for the training data with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the table should convey the count. In R use functions *melt()* and *cast()*, or function *table()*.

b. Consider the task of classifying a customer who owns a bank credit card and is actively using online banking services. Looking at the pivot table, what is the probability that this customer will accept the loan offer? [This is the probability of loan acceptance (Loan = 1) conditional on having a bank credit card (CC = 1) and being an active user of online banking services (Online = 1)].

c. Create two separate pivot tables for the training data. One will have Loan (rows) as a function of Online (columns) and the other will have Loan (rows) as a function of CC.

d. Compute the following quantities [$P(A \mid B)$ means "the probability of A given B"]:

i. $P(\text{CC} = 1 \mid \text{Loan} = 1)$ (the proportion of credit card holders among the loan acceptors)

ii. $P(\text{Online} = 1 \mid \text{Loan} = 1)$

iii. $P(\text{Loan} = 1)$ (the proportion of loan acceptors)

iv. $P(\text{CC} = 1 \mid \text{Loan} = 0)$

v. $P(\text{Online} = 1 \mid \text{Loan} = 0)$

vi. $P(\text{Loan} = 0)$

e. Use the quantities computed above to compute the naive Bayes probability $P(\text{Loan} = 1 \mid \text{CC} = 1, \text{Online} = 1)$.

f. Compare this value with the one obtained from the pivot table in (b). Which is a more accurate estimate?

g. Which of the entries in this table are needed for computing $P(\text{Loan} = 1 \mid \text{CC} = 1, \text{Online} = 1)$? In R, run naive Bayes on the data. Examine the model output on training data, and find the entry that corresponds to $P(\text{Loan} = 1 \mid \text{CC} = 1, \text{Online} = 1)$. Compare this to the number you obtained in (e).

8.2 **Automobile Accidents.** The file *Accidents.csv* contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting).

Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value "yes" if MAX_SEV_IR = 1 or 2, and otherwise "no."

a. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

b. Select the first 12 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R.

   i. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns.

   ii. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.

   iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5.

   iv. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.

   v. Run a naive Bayes classifier on the 12 records and two predictors using R. Check the model output to obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

c. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%).

   i. Assuming that no information or initial reports about the accident itself are available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (Use the Data_Codes sheet.)

   ii. Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

   iii. What is the overall error for the validation set?

   iv. What is the percent improvement relative to the naive rule (using the validation set)?

   v. Examine the conditional probabilities output. Why do we get a probability of zero for P(INJURY = No | SPD_LIM = 5)?