CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Spring 2019: Homework 2 (10 points)

Due date: Sunday Mar 10, 2019 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

This is the final version of the homework.

1. Exercises (4 points divided evenly among the questions). Please submit a PDF file containing answers to these questions. Any other file format will lead to a loss of 1 point. Non-PDF files that cannot be opened by the TAs will lead to a loss of 3 points.

1.1 Tan, Chapter 3

Exercise 2, 3, 5.

1.2 Tan, Chapter 4

Exercise 18 (show your work, don't just provide the answer without showing how you derived it).

1.3 Zaki, Chapter 8

Exercises 1,4, 6(a), 6(b)

1.4 Multiclass classification

Using the confusion matrix from multiclass.Rmd notebook (from Lecture 7), create a binary-class confusion matrix using the "one-vs-many" strategy for each class. Then, for each class, compute the sensitivity, specificity and precision to **two decimal places**. Show all work, including the binary class confusion matrices.

2. Practicum problems (3 points distributed as indicated) Please label your answers clearly, see Template.Rmd R notebook for an example (Template.Rmd R notebook is available in "Blackboard→Assignment and Projects → Homework 0".) Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points.

2.1 Decision tree classification (3 points divided evenly by all sub-questions)

You are provided two datasets from the 1994 US Census database: a training dataset (adult-train.csv) and a testing dataset (adult-test.csv). Each observation of the datasets has 15 attributes as described below. The class variable (response) is stored in the last attribute and indicates whether a person makes more than $50K per year.

The attributes are as follows:

age: Age of the person (numeric)
workclass: Factor, one of Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: Final sampling weight (used by Census Bureau to handle over and under-sampling of particular groups).
education: Factor, one of Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: Number of years of education (numeric).
marital-status: Factor, one of Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Factor, one of Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Factor, one of Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: Factor, one of White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Factor, one of Female, Male
capital-gain: Continuous
capital-loss: Continuous
hours-per-week: Continuous
native-country: Factor, one of United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
income: class variable (response), factor, one of ">50K", "<=50K".

You are to build a decision tree using rpart to predict whether a person makes more than $50K per year.

Both the training and testing dataset are not clean; some fields have '?' in them. You will remove those observations that contain '?'.

Do the questions below in the order that they appear since there will be dependency between some questions.

First, set the seed as follows before reading in the datasets.

```
> set.seed(1122)
```

(a) Remove all the observations that have '?' in them. Hints: To find out which attributes have a '?' in them, use sum(df$occupation == "?"). If this method returns a non-zero value, the value returned represents the number of times a '?' is seen in that column. Then, use which(df$occupation == "?") to determine the index of the rows containing the attribute that has the '?'. Recall that which() accepts as a parameter a logical vector (or array), and returns the indices where a TRUE occurs in the vector (or array). Consequently, the return value of which() will be a vector of indices. (See R-intro-1.r in Lecture 1 for an example that involves the use of which().) Collect all the indices of the columns where a '?' occurs into a vector, and use that vector to weed out the rows containing the columns with '?'

As a sanity check, when you are done with weeding out the '?', you should be left with 30,161 observations in the training set.

Do the same thing for the test dataset. Again, as a sanity check, you should be left with 15,060 observations in the test dataset after you have removed the rows containing '?' in a column.

The rest of the questions below assume that the training and testing datasets are clean.

(b) Build a decision tree model using rpart() to predict whether a person makes <=50K or >50K using all of the predictors. Answer the following questions through model introspection:
(i) Name the top three important predictors in the model?

(ii) The first split is done on which predictor? What is the predicted class of the first node (the first node here refers to the root node)? What is the distribution of observations between the "<=50K" and ">50K" classes at first node?

(c) Use the trained model from (b) to predict the test dataset. Answer the following questions based on the outcome of the prediction and examination of the confusion matrix: (for floating point answers, assume 3 decimal place accuracy):
(i) What is the balanced accuracy of the model? (Note that in our test dataset, we have more observations of class "<=50" than we do of class ">50". Thus, we are more interested in the *balanced accuracy*, instead of just accuracy. Balanced accuracy is calculated as the average of sensitivity and specificity.)
(ii) What is the balanced error rate of the model? (Again, because our test data is imbalanced, a balanced error rate makes more sense. Balanced error rate = 1.0 – balanced accuracy.)
(iii) What is the sensitivity? Specificity?
(iv) What is the AUC of the ROC curve. Plot the ROC curve.

(d) Print the complexity table of the model you trained. Examine the complexity table and state whether the tree would benefit from a pruning. If the tree would benefit from a pruning, at what complexity level would you prune it? If the tree would not benefit from a pruning, provide reason why you think this is the case.

(e) Besides the class imbalance problem we see in the test dataset, we also have a class imbalance problem in the training dataset. To solve this class imbalance problem in the training dataset, we will use undersampling, i.e., we will undersample the majority class such that both classes have the same number of observations in the training dataset.
(i) In the *training* dataset, how many observations are in the class "<=50K"? How many are in the class ">50K"?
(ii) Create a new training dataset that has equal representation of both classes; i.e., number of observations of class "<=50K" must be the same as number of observations of class ">50K". Call this new training dataset. (Use the sample() method on the majority class to sample as many observations as there are in the minority class. **Do not use any other method for undersampling as your results will not match expectation if you do so.)**
(iii) Train a new model on the new training dataset, and then fit this model to the testing dataset. Answer the following questions based on the outcome of the prediction and examination of the confusion matrix: (for floating point answers, assume 3 decimal place accuracy):
i) What is the balanced accuracy of this model?
(ii) What is the balanced error rate of this model?
(iii) What is the sensitivity? Specificity?
(iv) What is the AUC of the ROC curve. Plot the ROC curve.

(f) Comment on the differences in the balanced accuracy, sensitivity, specificity, positive predictive value and AUC of the models used in (c) and (e).

2.2 Association Analysis (3 points divided as indicated below)

In this assignment you will be using the *Extended Bakery* dataset, which describes transactions from a chain of bakery shops that sell a variety of drinks and baked goods.

The dataset is presented as a series of transactions, 1,000, 5,000, 20,000 and 75,000 transactions, stored in files named tr-1k.csv, tr-5k.csv, tr-20k.csv and tr-75k.csv, respectively. Each file contains the data in a sparse vector format, i.e., each line of the file has the following format:

1, 7, 15, 44, 49
2, 1, 19
...

The first column is the transaction ID and the subsequent columns contain a list of purchased goods from the bakery represented by their product ID code. In the example above, the first line implies that transaction ID one contained four items: 7, 15, 44, and 49. The mapping of the product ID to product name is provided in the products.csv file.

(a) **[0.25 points]** For each series of transaction files (i.e., tr-5k.csv, tr-20k.csv, …) create a canonical representation of the transaction file. A canonical representation for each dataset will be a file that contains a list

of product names (not IDs) on a line, each product separated by a comma and a newline ends the line. So, as an example, the vector shown above would correspond to the following canonical representation:

Coffee Eclair, Blackberry Tart, Bottled Water, Single Espresso
Lemon Cake, Lemon Tart
…

Save the canonical representation in files with the canonical suffix, i.e., tr-5k-canonical.csv, and so on. Use these files for the rest of the work. Include these files in the archive that you upload to Blackboard.

You can use a programming language of your choice to do part (a).

(b) **[1.25 points]** Given the database of transactions, where each transaction is a list of items, find rules that associate the presence of one set of items with that of another set of items. Ideally, we only want to find rules that are substantiated by the data; we want to avoid spurious associations.

Find association rules that exceed specific values of *minimum support* and *minimum confidence*. You are free to experiment with different values until you find something that produces meaningful results. Recall that this process requires two steps: finding the *frequent itemsets* and discovering strong *association rules* within them. You will use the R **arules** package as shown in class.

Your output should contain the following:

- For each frequent itemset:
    1. All items in it, described by the product names.
    2. The support of the itemset.
- For each rule:
    1. The antecedent.
    2. The consequent.
    3. The support of the rule.
    4. The confidence of the rule.

(c) **[1.25 points]** Given the above output, respond to the following question: Compare the rules you obtained for each different subset (1,000 – 75,000 transactions). How does the number of transactions affect the results you observed? (Write the answer in your R markup file, easily identified.)

(d) **[0.25 points]** Answer the following questions for the 75,000 transactions dataset using the same support level as determined in (b):

(i) What is the most frequently purchased item?
(ii) What is the least frequently purchased item?