Project Requirements for Decision Trees

Introduction

You are provided with 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 ‘?

- 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.
- 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:
- Name the top three important predictors in the model.
- 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?

- Use the trained model from above 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)
    1. 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.)
    2. 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.)
    3. What is the sensitivity and AUC of the ROC curve Plot the ROC curve.
- 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.
- 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.
    1. In the training dataset, how many observations are in the class “<=50K”? How many are in the class “>50K”?
    2. 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 under sampling as your results will not match expectation if you do so.)
    3. 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:
    4. Print, the balanced accuracy of this model, balanced error rate of this model, sensitivity, Specificity, AUC of the ROC curve, Plot the ROC curve

5. Comment on the differences in the balanced accuracy, sensitivity, specificity, positive predictive value and AUC of the models used.