KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

Data Mining
HW5

Anil Varma B

November 23, 2019

1. Download the dataset adult and use it to build a binary decision tree for target attribute income Note: you need to group those catagorical values into 2, 3 or 4 groups. For example, for matrital.status, you can group those values into two groups or levels: married, unmarried. Use "level" function in R. Sample codes have been provided in last class. After building the tree, you need provide error rate for this model for your training dataset.

1 READING AND UNDERSTANDING THE DATA SET

```
adult<-read.csv(file.choose(),head=TRUE)
dim(adult)
> dim(adult)
[1] 25000    15
head(adult)
```

```
> adult<-read.csv(file.choose(),head=TRUE)
> head(adult)
  age         workclass demogweight education education.num     marital.status        occupation  relationship  race    sex
apital.gain
1  39         State-gov       77516 Bachelors            13      Never-married      Adm-clerical Not-in-family White   Male
       2174
2  50 Self-emp-not-inc       83311 Bachelors            13 Married-civ-spouse   Exec-managerial       Husband White   Male
          0
3  38           Private      215646   HS-grad             9           Divorced Handlers-cleaners Not-in-family White   Male
          0
4  53           Private      234721      11th             7 Married-civ-spouse Handlers-cleaners       Husband Black   Male
          0
5  28           Private      338409 Bachelors            13 Married-civ-spouse    Prof-specialty          Wife Black Female
          0
6  37           Private      284582   Masters            14 Married-civ-spouse   Exec-managerial          Wife White Female
          0
  capital.loss hours.per.week native.country income
1            0             40  United-States <=50K.
2            0             13  United-States <=50K.
3            0             40  United-States <=50K.
4            0             40  United-States <=50K.
5            0             40           Cuba <=50K.
6            0             40  United-States <=50K.
```

Figure 1.1: Dimensions of the data set

Checking the levels in categorical columns by using the levels command

```
levels(adult$marital.status)
levels(adult$workclass)
levels(adult$education)
levels(adult$occupation)
levels(adult$relationship)
levels(adult$race)
levels(adult$sex)
levels(adult$native.country)
```

```
> levels(adult$marital.status)
[1] "Divorced"              "Married-AF-spouse"     "Married-civ-spouse"    "Married-spouse-absent"
[5] "Never-married"         "Separated"             "Widowed"
> levels(adult$workclass)
[1] "?"                 "Federal-gov"       "Local-gov"         "Never-worked"      "Private"           "Self-emp-inc"
[7] "Self-emp-not-inc" "State-gov"         "Without-pay"
> levels(adult$education)
 [1] "10th"         "11th"         "12th"         "1st-4th"      "5th-6th"      "7th-8th"      "9th"
 [8] "Assoc-acdm"   "Assoc-voc"    "Bachelors"    "Doctorate"    "HS-grad"      "Masters"      "Preschool"
[15] "Prof-school"  "Some-college"
> levels(adult$occupation)
 [1] "?"                 "Adm-clerical"      "Armed-Forces"      "Craft-repair"      "Exec-managerial"
 [6] "Farming-fishing"   "Handlers-cleaners" "Machine-op-inspct" "Other-service"     "Priv-house-serv"
[11] "Prof-specialty"    "Protective-serv"   "Sales"             "Tech-support"      "Transport-moving"
> levels(adult$relationship)
[1] "Husband"        "Not-in-family"  "Other-relative" "Own-child"      "Unmarried"      "Wife"
> levels(adult$race)
[1] "Amer-Indian-Eskimo" "Asian-Pac-Islander" "Black"              "Other"              "White"
> levels(adult$sex)
[1] "Female" "Male"
> levels(adult$native.country)
 [1] "?"                         "Cambodia"                  "Canada"                    "China"
 [5] "Columbia"                  "Cuba"                      "Dominican-Republic"        "Ecuador"
 [9] "El-Salvador"               "England"                   "France"                    "Germany"
[13] "Greece"                    "Guatemala"                 "Haiti"                     "Holand-Netherlands"
[17] "Honduras"                  "Hong"                      "Hungary"                   "India"
[21] "Iran"                      "Ireland"                   "Italy"                     "Jamaica"
[25] "Japan"                     "Laos"                      "Mexico"                    "Nicaragua"
[29] "Outlying-US(Guam-USVI-etc)" "Peru"                     "Philippines"               "Poland"
[33] "Portugal"                  "Puerto-Rico"               "Scotland"                  "South"
[37] "Taiwan"                    "Thailand"                  "Trinadad&Tobago"           "United-States"
[41] "Vietnam"                   "Yugoslavia"
```

Figure 1.2: Levels for categorical columns

Grouping the categorical values into different levels below.

```
levels(adult$marital.status)[2:4]<- "Married"
levels(adult$workclass)[c(2,3,8)]<- "Gov"
levels(adult$workclass)[c(5,6)]<- "Self"
levels(adult$education)[c(1,2,3,4,5,6,7,12,14,16)] <- "Pre-Grad"
levels(adult$relationship)[c(1,4,6)] <-"Family"
```

Checking the new grouped levels

```
levels(adult$marital.status)
levels(adult$workclass)
levels(adult$education)
levels(adult$relationship)
```

```
> levels(adult$marital.status)
[1] "Divorced"      "Married"       "Never-married" "Separated"     "Widowed"
> levels(adult$workclass)
[1] "?"           "Gov"         "Never-worked" "Private"      "Self"         "Without-pay"
> levels(adult$education)
[1] "Pre-Grad"    "Assoc-acdm"  "Assoc-voc"   "Bachelors"   "Doctorate"   "Masters"     "Prof-school"
> levels(adult$relationship)
[1] "Family"         "Not-in-family"  "Other-relative" "Unmarried"
```

Figure 1.3: New categorical levels

```
adult$workclass = as.factor(adult$workclass)
adult$marital.status = as.factor(adult$marital.status)
adult$education = as.factor(adult$education)
adult$relationship = as.factor(adult$relationship)
```

2 BUILD A BINARY DECISION TREE FOR TARGET ATTRIBUTE INCOME AND PLOTTING A DECISION TREE .

```
library("rpart")
library("rpart.plot")
newrfit <- rpart(income ~ .,data= adult,method="class")
rpart.plot(newrfit)
```

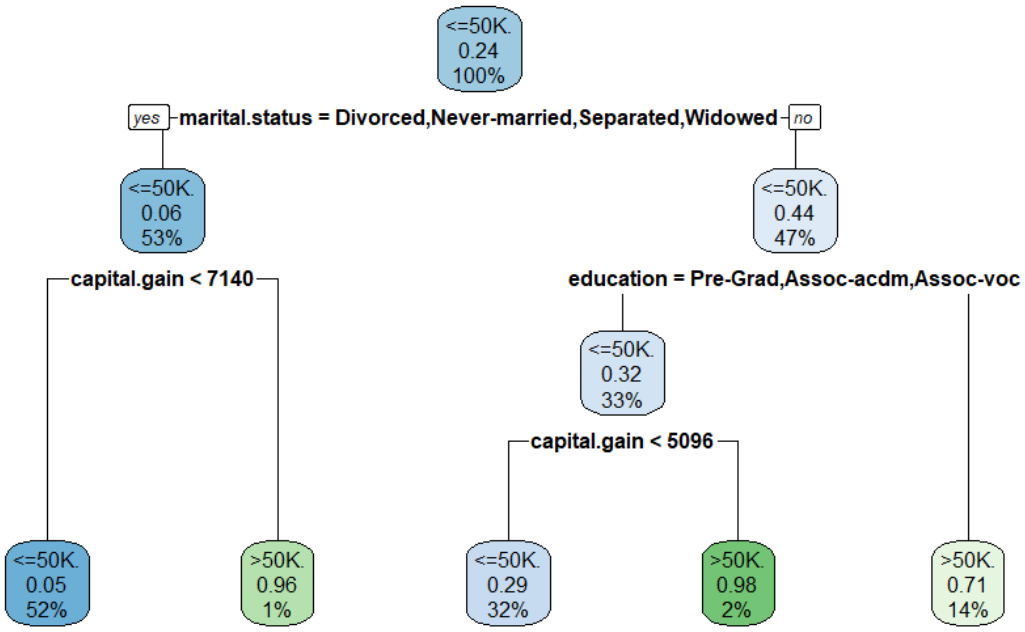

Figure 2.1: Binary tree plot

3 MAKING PREDICTIONS AND CALCULATING THE ERROR RATE

```
outcomes <- predict(newrfit, adult[1:14])
predictions <- ifelse(outcomes[,1] >= .5, " <=50K", " >50K")
predictions_table<- table(adult$income,predictions)
predictions_table
error<- sum(predictions_table[row(predictions_table) != col(predictions_table)])
    / sum(predictions_table)
error
```

```
> predictions_table
        predictions
          <=50K  >50K
  <=50K.  17975  1041
  >50K.    2905  3079
> error<- sum(predictions_table[row(predictions_table) != col(predictions_table)]) / sum(predictions_table)
> error
[1] 0.15784
```

Figure 3.1: prediction and error rate