Homework 1 R markdown

Abra Brisbin

2021-07-29

Possible Solutions to Selected Questions

Important: This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework.* We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet.* By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1:

Question:

Set R's seed to 1 (for Homework 1) with: **set.seed(1)**

Create a `groups` vector of 256 copies of the number 1 (to represent observations that will be in the training set) and 136 copies of the number 2 (to represent observations that will be in the validation set. Then use **sample()** to randomize the order of the vector.

Make a vector that contains TRUE for each data point of `Auto` that will be in the training set, and FALSE for each data point that will be in the test (or validation) set.

Possible Answer:

```r
set.seed(1)
groups = c(rep(1, 256), rep(2, 136)) # 1 represents the training set
random_groups = sample(groups, 392)

in_train = (random_groups == 1)
```

Question:

Standardize the `weight` and `year` columns of the training set. Then standardize the `weight` and `year` columns of the test set, *using the original mean and standard deviation of the training set.*

Possible Answer:

```r
train_std = scale(Auto[in_train, c("weight", "year")])
test_std = scale(Auto[!in_train, c("weight", "year")],
    center = attr(train_std, "scaled:center"),
    scale = attr(train_std, "scaled:scale"))
```

Question:

Use a for() loop to apply K-nearest neighbors regression to the same training and validation sets, for values of k from 1 to 50. Make a plot of the MSE (calculated for the validation set predictions) as a function of k.

Enter your R code and plot on this question on Canvas. (Use **Insert** -> **Image** to insert the plot.)
Possible Answer:

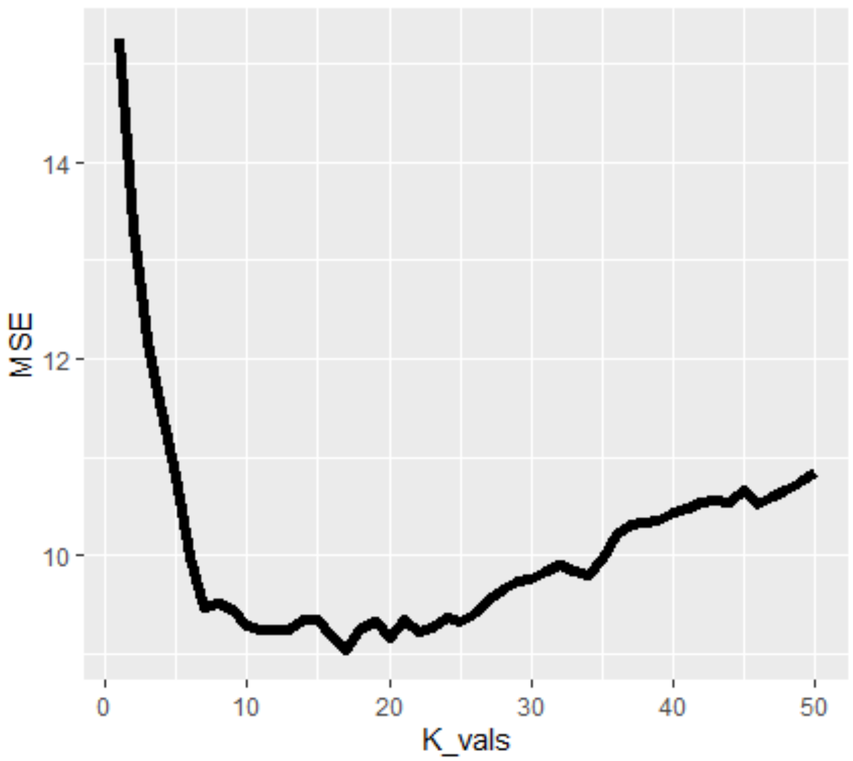
```r
K_vals = 1:50
MSE = numeric(length = length(K_vals))

for(ii in 1:length(K_vals)){
  predictions = knn.reg(train = train_std,
                        test  = test_std,
                        y = Auto$mpg[in_train],
                        k = K_vals[ii])

  MSE[ii] = mean( (predictions$pred - Auto$mpg[!in_train])^2 )

}

gf_line(MSE ~ K_vals, lwd = 2)
```



Question:

In your opinion, which value of k is the best choice? Why?

Possible Answer: The lowest MSE on the validation set occurs at $k = 17$, so this appears to be a good choice. Any k between about 7 and 25 looks reasonable. **Note** that with KNN, *larger* values of k actually produce lower-variance models.

From Problem 2:

Question:

Read the data into R. One-hot encode the variable `Sex`, using `Male` as the default value. (Unfortunately, because this data set is from the US Census, it did not allow allow an option of Intersex, or distinguish between sex and gender.)

Possible Answer:

```
income <- income %>%
  mutate(Sex = ifelse(Sex == "Male", 0, 1))
```

Question:

Use 25-nearest neighbor classification (fit on the training set) to predict whether the income of each individual in the validation set is >50K or <=50K.

Find the confusion matrix. You should be able to produce a matrix table with two rows and two columns, similar to the one below.

Please enter the information as whole numbers. Note carefully the labels for the rows and columns, and be sure to orient your table accordingly.

Possible Answer:

```
predictions = knn(train = x_train,
                  test  = x_test,
                  cl = income$Income[in_train],
                  k = 25)
conf_mat = table(predictions,
                 income$Income[!in_train])
conf_mat
```

Question:

Make a grid of example points with values of education from 1 to 16, ages from 17 to 75, and sex from 0 to 1. Standardize education and age using the original mean and standard deviation of the training set (from question 11).

Create a data frame, x_example, containing the standardized education and age and the unstandardized sex from the example points. The order of the columns should match the order of the columns in x_train.

Possible Answer:

```
educ_to_check = 1:16
age_to_check = 17:75
sex_to_check = c(0, 1)

example_data = expand.grid(educ_to_check,
                           age_to_check,
                           sex_to_check)

example_std = scale(example_data[ , 1:2],
    center = attr(quant_train_std, "scaled:center"),
    scale = attr(quant_train_std, "scaled:scale"))

# Be sure the columns are in the same
# order as the training data
x_example = cbind(example_std,
                  example_data[ ,3])
```

Question:

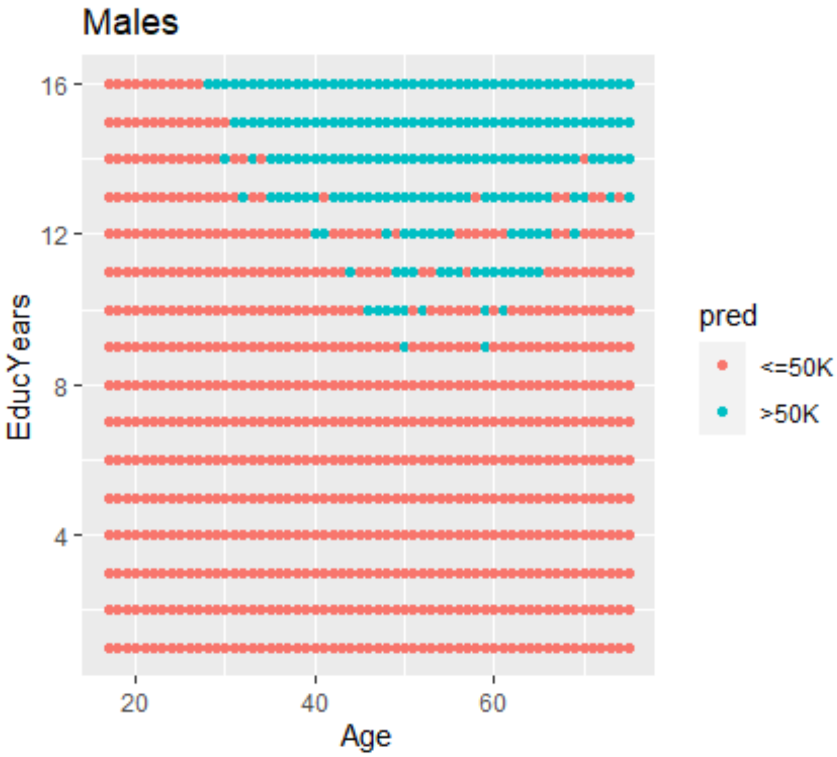
Use 25-nearest neighbors to predict the income classifications of the example points, using the same training data as in question 12. Make graphs showing the relationship between education, age, sex, and predicted income.

Possible Answer:

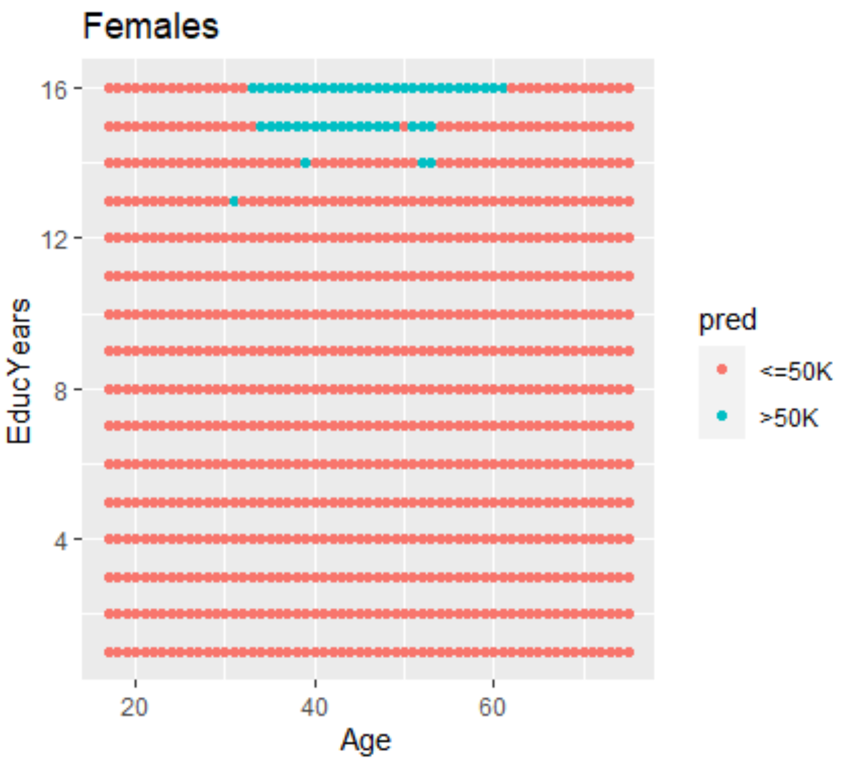
```
predictions = knn(train = x_train,
                  test  = x_example,
                  cl = income$Income[in_train],
                  k = 25)

example_data <- example_data %>%
  mutate(pred = predictions) %>%
  rename(EducYears = Var1,
         Age = Var2,
         Sex = Var3)

example_data %>%
  filter(Sex == 0) %>%
  gf_point(EducYears ~ Age, color =~ pred) %>%
  gf_labs(title = "Males")
```



```
example_data %>%
  filter(Sex == 1) %>%
  gf_point(EducYears ~ Age, color =~ pred) %>%
  gf_labs(title = "Females")
```



Question:

Write 3-6 sentences interpreting the graphs you made in the previous question. (For purposes of interpreting the results, note that the data are from the 1990s.)

Possible Answer: Men age 28 and older with high levels of education (beyond a high school diploma) tend to be predicted to have high incomes (>50K). There is a narrower age band of around 40-60 years old in which men with only a high school diploma, or less, often have high incomes. This suggests that for men, experience in a job can substitute for high levels of education, in terms of obtaining a good income. This effect doesn't hold for men above about age 60, which may indicate that they are partly retired, or possibly subject to age-based discrimination.

In contrast, women are only predicted to have a high income if their age is in the range of about 30-60 *and* they have a high level of education. This suggests that women need *both* experience in a job and high levels of education to earn a good income.