Homework 3 Possible Answers

Abra Brisbin

2021-08-01

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

Read in the data Wisconsin_income.csv. Open the data dictionary in a text editor.

Notice that the following 8 variables are categorical, but are coded as numbers:

- Citizenship
- Class of worker
- Language spoken at home
- Marital status
- Sex
- Disability
- Race
- Hispanic

Tell R to treat them as factors.

Possible Answer:

```
wi = read_csv("../Wisconsin_income.csv")
```

```r
wi <- wi %>%
  mutate(CIT2 = factor(CIT2),
         COW = factor(COW),
         LANX = factor(LANX),
         MAR = factor(MAR),
         SEX = factor(SEX),
         DIS = factor(DIS),
         RAC = factor(RAC),
         Hispanic = factor(Hispanic))
```

Question

Make histograms of people's total earnings, usual hours worked per week, and travel time to work. Which of these 3 variables are likely to benefit from log-transformation? Apply the transformation if appropriate, and enter your R code below.

Possible Answer:

```r
wi <- wi %>%
  mutate(log_earn = log(PERNP),
         log_travel = log(JWMNP)) %>%
  select(-c(PERNP, JWMNP))
```

Question

Set the random seed equal to 3.

Perform 10-fold cross-validation to choose the best size of model (from 1 to 39 variables) based on cross-validation MSE. Record the mean squared error within each fold for each size of variable. **Note**: This step will probably take a few minutes to run!
Enter your R code below.

Possible Answer:

```r
# Define a predict() function for regsubsets objects
predict.regsubsets <- function(object, alldata, subset, id, ...){
    form = as.formula(object$call[[2]])
    mat = model.matrix(form, alldata)
    mat = mat[subset, ]

    if(sum(subset) == 1 | length(subset) == 1){
      # For LOOCV, convert mat to a matrix
      mat = t(as.matrix(mat))
    }

    coefi = coef(object, id=id)
    xvars = names(coefi)
    mat[ , xvars] %*% coefi
} # end function predict.regsubsets
```

```
n = dim(wi)[1]
ngroups = 10
groups = rep(1:ngroups, length = n)

set.seed(3)
cvgroups = sample(groups, n)

nvar = 39
group_error = matrix(NA, nr = ngroups, nc = nvar)
              # row = fold,
              # column = model size (number of variables)

for(ii in 1:ngroups){ # iterate over folds
  groupii = (cvgroups == ii)
  train_data = wi[!groupii, ]
  test_data = wi[groupii, ]


  cv_fit = regsubsets(log_earn ~ .,
                data = train_data, nvmax = nvar)


  for(jj in 1:nvar){ # iterate over model size

    y_pred = predict(cv_fit, alldata = wi,
                     subset = groupii, id = jj)
    # Normally, we'd store this:
    # all_predicted[groupii, jj] = y_pred

    MSE = mean((test_data$log_earn - y_pred)^2)
    group_error[ii, jj] = MSE


  } # end iteration over model size
} # end iteration over folds
```

Question

Estimate the standard error of the cross-validation errors and find the most parsimonious model with a CV error within 1 standard error of the lowest. How many predictor variables are in the most parsimonious model with a CV error within 1 standard error of the lowest?

Possible Answer

```
std_err = apply(group_error, 2, sd)/sqrt(ngroups)
std_err[low_MSE_model]
which(MSE_overall <= MSE_overall[low_MSE_model] +
                     std_err[low_MSE_model])
```

Question 10 (4 points)

Use `regsubsets` to find the best model for the whole data set which has the number of variables you found in the previous question. For each variable included in the model, write a sentence giving a possible explanation for the direction of the association. Refer to variables in plain English.

Note: It may be helpful to refer to the data dictionary and/or a map of Wisconsin, such as https://en.wikipedia.org/wiki/Wisconsin#/media/File:Wisconsin-counties-map.gif.

Example: "Being in a union is positively associated with earnings. This suggests that unions' collective bargaining tends to be successful in convincing employers to offer higher wages."

```
coef(regfit_wi, 5)
```

```
## (Intercept)        COW6        MAR5        SEX2        WKHP   Education
##  8.19757196 -0.63769603 -0.24779279 -0.28139377  0.02987906  0.09413591
```

Possible Answer:

- Being self-employed in a non-incorporated business is associated with lower pay. This could reflect employment in jobs such as babysitting, which may not be highly remunerative, or people whose self-employment businesses have not become sufficiently developed to warrant incorporation.
- Being never married is associated with lower pay. This could be due to an association between age and earnings (younger people earn less due to having less experience, and young people are also more likely to have never married), but it is interesting that this variable appeared in the model instead of age.
- In this data set, females tended to earn less. This model accounts for hours worked, but we can't tell how much of this is due to discrimination vs. how much is due to women choosing lower-paying professions than men.
- Working more hours per week is associated with higher income. This makes sense, because for people who are paid hourly, working more hours directly increases their pay.
- Having more education is associated with higher income. This suggests that earning a college degree or master's degree qualifies one for higher-paying jobs.

From Problem 2:

Question

Write R code to:

- Load the **Auto** data set into R. The data set is in the ISLR library.
- **Tell R to treat the `origin` variable as a factor.**

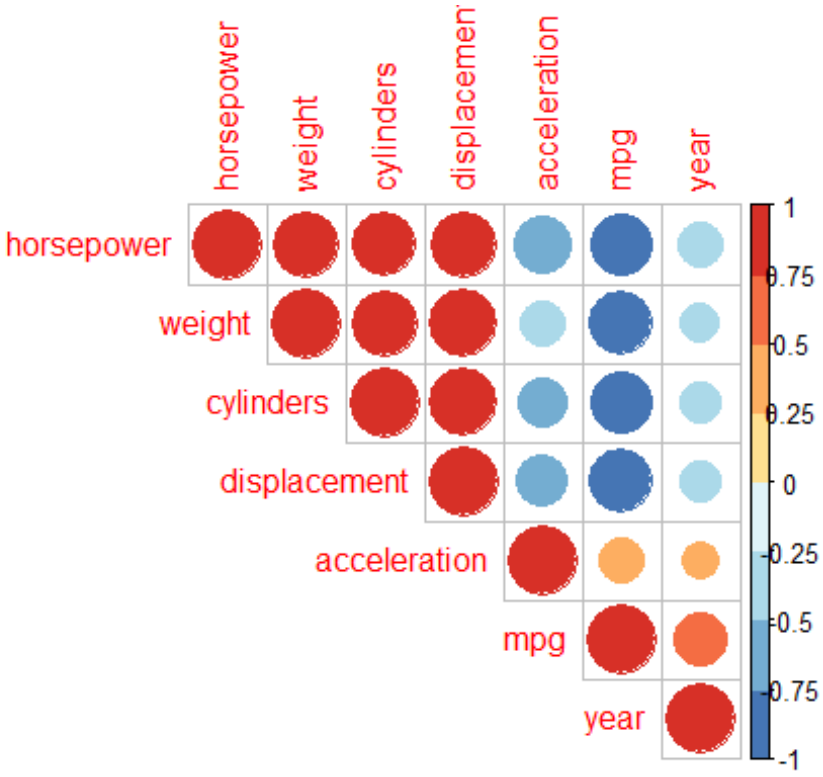
- Create a binary variable that equals "high" for cars with gas mileage (mpg) greater than or equal to the median and "low" for cars with gas mileage below the median. Tell R to treat it as a factor.

Possible Answer:

```
mpg_cutoff = median(Auto$mpg)

Auto <- Auto %>%
  mutate(origin = factor(origin),
         mpg_bin = factor(ifelse(mpg >= mpg_cutoff, "high", "low")))
```

Question

Make a correlation plot of the numeric variables in **Auto**.



Possible Answer:

Question

Remove any variables with VIFs greater than or equal to 10. Set the random seed equal to 3 and perform 10-fold cross-validation. In each phase of the cross-validation, fit the logistic model (excluding name, continuous mpg, and the variable(s) you found in the previous question) and predict the probability of high gas mileage for each data point in the validation set. Store all of the probabilities in a single vector.

Note: Depending on how you set up the formula in the logistic regression, the predict function may give an error, "Factor name has new levels." This is complaining about the

fact that there are models of car in the validation set that weren't included in the training data. But, it's not really a problem, because we're not using name as a predictor variable. You can create a new data frame that excludes name, or you can update the levels of the name factor in the logistic model, as shown here.

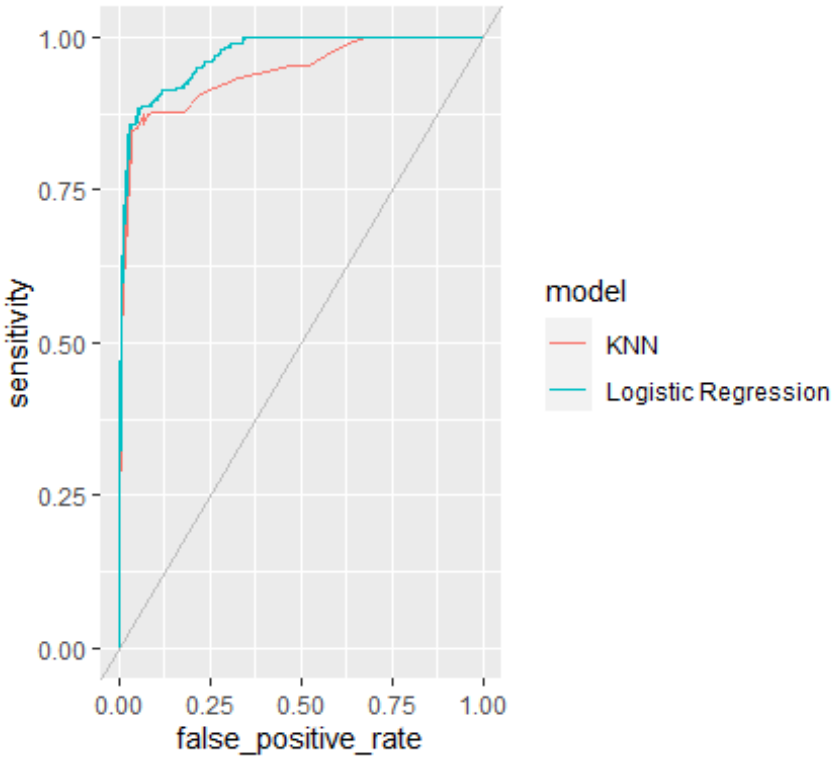
Possible Answer:

```
n = dim(Auto)[1]
ngroups = 10 # using 10-fold cross-validation
groups = rep(1:ngroups, length = n)

set.seed(3)
cvgroups = sample(groups, n)
all_predicted = numeric(length = n)

for(ii in 1:ngroups){
  groupii = (cvgroups == ii)
  train_set = Auto[!groupii, ]
  test_set = Auto[groupii, ]

  model_fit = glm(mpg_bin ~ . - mpg - name,
            data = train_set, family="binomial")
  model_fit$xlevels[["name"]] <- levels(Auto$name)

  predicted = predict(model_fit, newdata = test_set,
                                  type="response")
  all_predicted[groupii] = predicted
}

## Setting levels: control = high, case = low

## Setting direction: controls < cases
```

Question

The file `Homework_3_KNN.csv` (available on Homework 3 on Canvas) contains the data needed to construct a ROC curve for a KNN model of binary gas mileage, using 49 nearest neighbors. Make a graph showing the ROC curve of the logistic regression and KNN models. Write 1-2 sentences comparing the models based on their ROC curves.

Possible Answer:



The two models are both quite good, but logistic regression is slightly better because it has a slightly higher AUC.