Homework 10 Possible Answers

Abra Brisbin

2021-08-04

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

Data Set: Load the **OJ** data set, which is in the **ISLR** library. Set the random seed to 10.

Use 5-fold CV with `caret` to build an artificial neural network with 1 hidden node and no weight decay, to model Purchase as a function of LoyalCH, SalePriceMM, and PriceDiff. Tell `caret` to center and scale the data.

Possible Answer:

```
data_used = OJ
set.seed(10)

ctrl = trainControl(method = "cv", number = 5)
fit_OJ = train(Purchase ~ LoyalCH + SalePriceMM + PriceDiff,
                     data = data_used,
                     method = "nnet",
                     tuneGrid = expand.grid(size = 1, decay = 0),
                     preProc = c("center", "scale"),
                     trace = FALSE,
                     trControl = ctrl)

fit_OJ
```

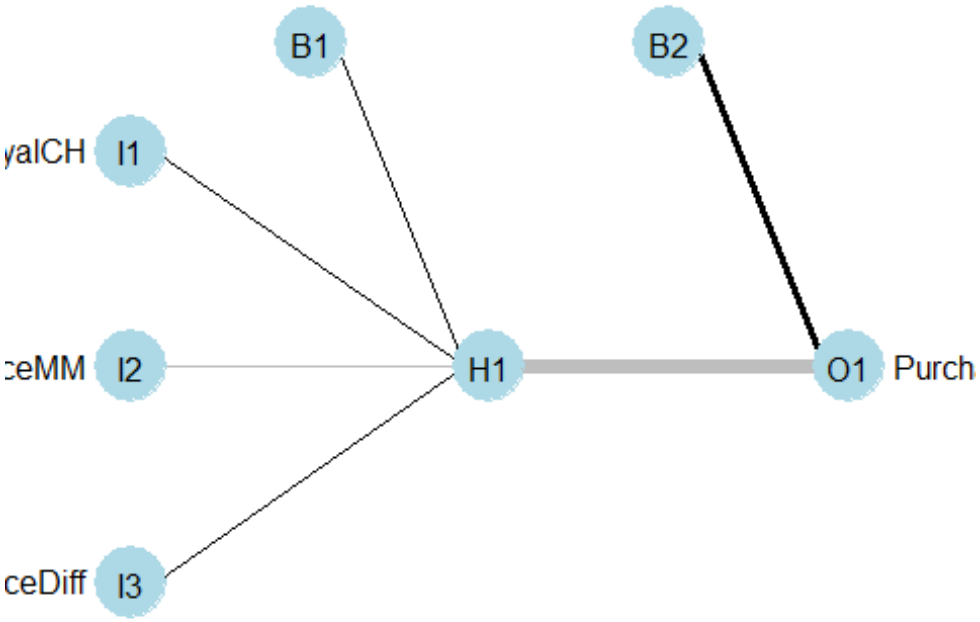
```
## Neural Network
##
## 1070 samples
##    3 predictor
```

```
##    2 classes: 'CH', 'MM'
##
## Pre-processing: centered (3), scaled (3)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 856, 856, 855, 856, 857
## Resampling results:
##
##   Accuracy   Kappa
##   0.8252672  0.6319247
##
## Tuning parameter 'size' was held constant at a value of 1
## Tuning
##  parameter 'decay' was held constant at a value of 0
```

Question

Make a plot of your neural network. You do not need to label the edges with their weights.

Possible Answer:




```
## a 3-1-1 network with 6 weights
## options were - entropy fitting
##  b->h1 i1->h1 i2->h1 i3->h1
##   0.08   1.30  -0.13   0.65
##  b->o h1->o
##  3.06 -7.46
```

Question

Suppose we classify predicted purchases as "MM"" if the probability of purchasing Minute Maid is > .9, as "CH" if the probability of purchasing Minute Maid is < .1, and NA otherwise. What is the classification error rate among purchases for which we make a (non-NA) prediction? Enter your answer to 3 decimal places.

Possible Answer:

```
preds <- preds %>%
  mutate(strong_class = case_when(CH > .9 ~ "CH",
                                  MM > .9 ~ "MM",
                                  TRUE ~ NA_character_))

conf_mat = table(preds$strong_class, OJ$Purchase)
conf_mat
1 - sum(diag(conf_mat))/sum(conf_mat) # Error rate
```

From Problem 2:

In this problem, you will use an artificial neural network to model the salaries of baseball players.

Question

Data Set: Load the **Hitters** data set in the **ISLR** package.

Remove any rows with missing Salary data.

Create new variables as follows, adding to the data frame in the order listed:

1. **League01**, which equals 0 if **League** = "A" and equals 1 if **League** = "N".

2. **Division01**, which equals 0 if **Division** = "E" and equals 1 if **Division** = "W".

3. **NewLeague01**, which equals 0 if **NewLeague** = "A" and equals 1 if **NewLeague** = "N".

Do not convert the new variables to factors. *Remove* the old variables (**League**, **NewLeague**, and **Division**) from the data frame.

Possible Answer:

```
my_hitters <- Hitters %>%
  filter(!is.na(Salary)) %>%
  mutate(League01    = ifelse(League    == "A", 0, 1),
         Division01  = ifelse(Division  == "E", 0, 1),
         NewLeague01 = ifelse(NewLeague == "A", 0, 1)) %>%
  select(-c(League, NewLeague, Division))
```

Question

Set the random seed equal to 10 again. We will fit an artificial neural network with 5 hidden nodes to model **Salary** as a function of all other variables in the data set. Use `caret` to perform 10-fold cross-validation to select the best decay rate, λ, from the set 1, 1.1, 1.2, …, 1.9, 2.

- Tell `caret` to center and scale the data.
- Use a linear output function.
- To ensure convergence, use `maxit = 2000`.

Possible Answer:

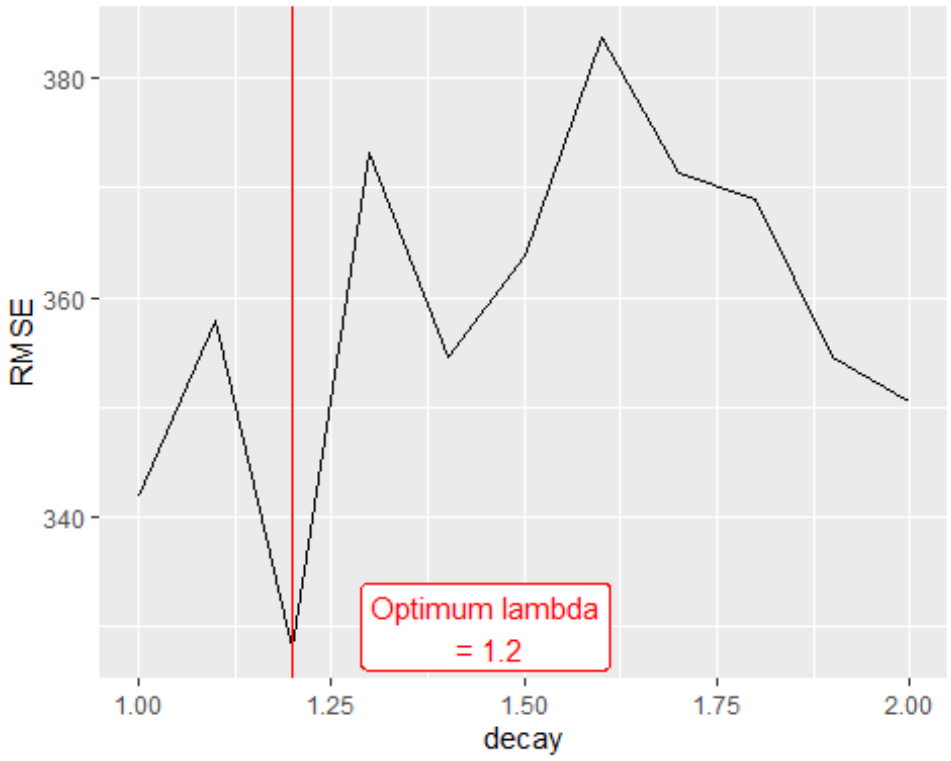
```
data_used = my_hitters
set.seed(10)
ctrl = trainControl(method = "cv", number = 10)
fit_hit = train(Salary ~ .,
                  data = data_used,
                  method = "nnet",
                  tuneGrid = expand.grid(size = 5,
                                         decay = seq(1, 2, by = .1)),
                                         #decay = c(0, .5, 10^(-c(1:7)))),
                  preProc = c("center", "scale"),
                  linout = TRUE,
                  maxit = 2000,
                  trace = FALSE,
                  trControl = ctrl)
```

Question

Make a graph of the RMSE as a function of λ. Add something to your graph (a title, a label, a line, …) to indicate the value of λ that optimizes the RMSE.

- It may be helpful to refer to the `$results` component of your `caret` object.

Possible Answer:



Question

Make a set of example points with realistic values of `Hits` and all other predictor variables held constant at their medians. (It may be helpful to refer to your notes about random forests.)

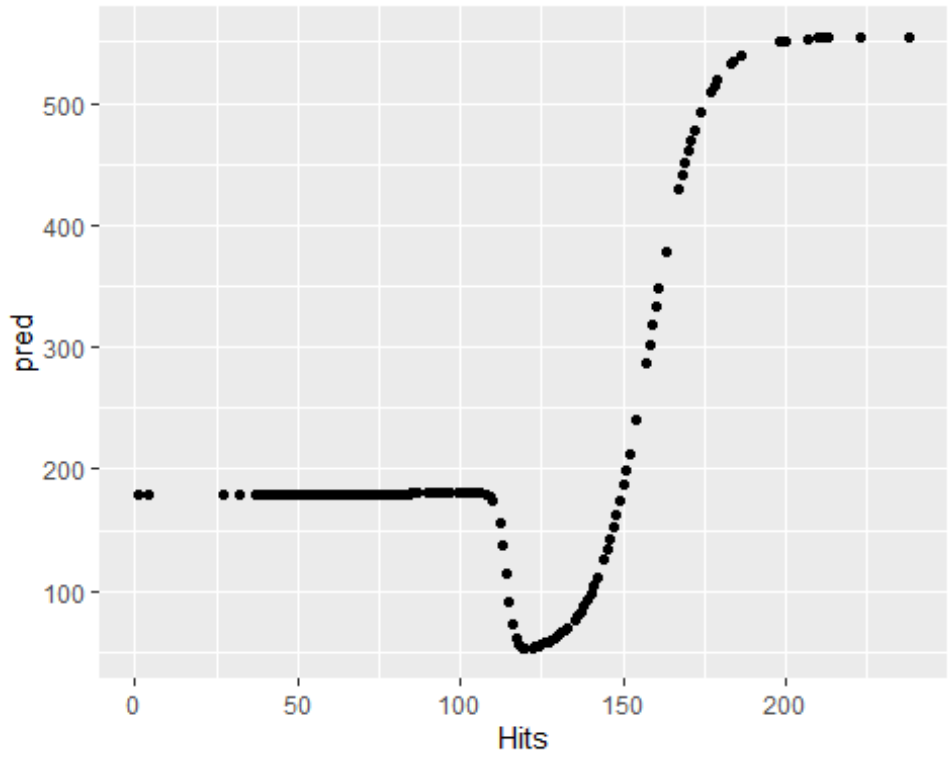
Use the final model from `caret` to predict the salary of the players in your set of example points. Make a graph of predicted salary as a function of `Hits`.

Possible Answer:

```
example_data <- my_hitters %>%
  mutate(across(c(-Hits), median))

example_data <- example_data %>%
  mutate(pred = predict(fit_hit, example_data))

example_data %>%
  gf_point(pred ~ Hits)
```



Question

- On what range of values does the relationship between `Hits` and predicted salary agree with your expectations?

- On what range of values does `Hits` have little effect on predicted salary?

- Suggest a possible explanation for why `Hits` has little effect on predicted salary in this range. (For example, what other factors may have greater influence on salary, for players with `Hits` in this range?)

If you are not familiar with baseball, it may be helpful to read the first two paragraphs of the Wikipedia page on baseball.

Possible Answer:

- For `Hits` between 125 and 200, the relationship between `Hits` and predicted salary is positive, which agrees with my expectations.
- `Hits` has essentially no effect on predicted salary (for players with median values of other variables) when the number of hits is less than 100.

- One possible explanation is that the group of players with 0-100 hits in the season includes some players who are weaker overall, who are paid less, and some players who are weak or average hitters, but who are outstanding at fielding the ball, so they

are paid more. For these players, variables such as Errors, PutOuts, and Assists may be more informative for predicting salary.

Question

If your goal was to optimize the performance of this model, what would you try next? Suggest **two** ideas.

Possible Answer: I would start by building a set of predictor variables with reduced correlations between them, either by using Principal Components Analysis or by regressing each predictor variable on the others and extracting the residuals. This is likely to be helpful, because we saw in the Decision Trees homework assignment that several of the predictors are highly correlated.

Other possible answers could involve using cross-validation to select the number of hidden nodes; duplicating the data set and adding noise; or using early stopping. (Note that early stopping is more challenging to implement with nnet() than the other approaches.)