Algorithms for Decision Making: Exam 1

Jaime Davila

3/24/22

Exam 1 Guidelines

- You have the entire class time (80 minutes) to complete this exam.
- Please make sure to save **immediately** your work in your submit folder. By the end of class time save the last version of your work and **don't** modify it afterwards.
- You are to work completely independently on the exam.
- You are allowed to use your class notes, moodle, worksheets, homeworks, textbooks, plus the "Help" feature from Rstudio.
- You **are not** permitted to do web searches.
- Please silence your cell phone. Place it and any other connected devices in your bag and do not access them for any reason.

For questions that ask for interpretations and explanations, usually no more than a sentence or two is needed for justification. Be thorough, but do not "brain dump". Notice that three sections of this exam are independent and that you can complete later sections successfully whether or not earlier sections are correct.

Do not spend too long on any one question. If you are not sure about an answer, write a note detailing your concern.

PLEDGE: I pledge on my honor that I have neither received nor given assistance during this exam nor have I witnessed others receiving assistance, and I have followed the guidelines as described above.

SIGNATURE:

○ I have intentionally not signed the pledge.

The Weather in Minnesota (40 points)

It is spring in Minnesota and we are interested on predicting what month of the year we are living based on the outside temperature. To do that let's load our testing and training datasets

```
month.levels <-  c("Jan","Feb","Mar")
temp.train.tbl <- read_csv("~/Mscs 341 S22/Class/Data/mn_weather.train.csv") %>%
  mutate(month=factor(month, month.levels))
temp.test.tbl <- read_csv("~/Mscs 341 S22/Class/Data/mn_weather.test.csv") %>%
  mutate(month=factor(month, month.levels))
```

Question 1 (15 points)

Do a boxplot using `temp.train.tbl` with the distribution of the temperature across the first 3 months of the year. Calculate the mean and standard deviation of the temperature for each of the 3 months. Based on this information, which method would you choose between `lda` and `qda` to predict the month based on the temperature? Justify your answer.

Question 2 (15 points)

Using `tidymodels()` create the model you chose in `Question 1` (either `lda` or `qda`) for predicting the month based on the temperature. Using your testing dataset, plot the predicted probability of a temperature corresponding to January, February or March in a single graph. How do you interpret this plot and what are the classification boundaries for each month? Using the information from this plot on which month will your model be making more mistakes?

Question 3 (10 points)

Calculate the sensitivity and specificity of your model and interpret them in the context of your problem. Based on the confusion matrix, on which month does your model make more mistakes?

Wrangling with Terminator 2 (20 points)

`Movielens` is a dataset containing user provided feedback from different movies. You can find more information about this dataset using `?movielens`. We can load this dataset by doing:

```
data(movielens)
movielens.tbl <- tibble(movielens)
```

We are interested in creating a simple recommendation system for the classic 90s movie and one of Prof. Davila's favorites, "Terminator 2: Judgment Day"

Step one (5 points)

What are the movieId and genres corresponding to "Terminator 2: Judgment Day"? Do a histogram of the user ratings for this movie.

Step two (5 points)

Create a table with the average movie ratings for each user. Also, create a table with the average rating for movies with genre "Action|Sci-Fi" for each user.

Step three (10 points)

Create a table `terminator2.movie.tbl` with columns:

- `userId`
- The average rating of all movies for that `userId`.
- The average rating of all movies in the "Action|Sci-Fi" for that userId.
- The rating that user gave to Terminator 2

Make sure to remove rows that have "NA" in either of those columns

Rating Terminator 2 (40 points)

Make sure that the tibble you obtained from the previous point has 237 observations and 4 variables. In case you ran into problems you can use the following code:

```
terminator2.movie.tbl <- read_csv("~/Mscs 341 S22/Class/Data/movies.csv")
dim(terminator2.movie.tbl)
```

```
## [1] 237   4
```

Let's create our testing and testing dataset

```
set.seed(55057)
terminator2.split <- initial_split(terminator2.movie.tbl)
terminator2.train.tbl <- training(terminator2.split)
terminator2.test.tbl <- testing(terminator2.split)
```

Linear models (20 points)

Using `tidymodels()` create a linear model to predict the rating of Terminator 2 based on `avg.rating` and `avg.action.scifi`. How do you interpret the coefficients of your linear model? What is your `rmse` and R^2 in your testing dataset?

KNN models (20 points)

Using `tidymodels()` create a KNN model to predict the rating of Terminator 2 based on `avg.rating` and `avg.action.scifi`. Create a 10-fold cross-validation set and use it to find the optimal `neigbors` by minimizing the `rmse`. Describe in your own words how to construct the 5th training and testing dataset from your cross-validation set. Calculate the `rmse` and R^2 of your model in your testing dataset and compare it with the linear model from the previous point.

```
set.seed(55057)
```

Extra Credit! (5 points)

Knit your document into a PDF and submit it though the Moodle webpage.