CIS 8392
Topics in Big Data Analytics

#Assignment 4

Yu-Kai Lin

Assignment 4

Step 0. Save the following RDS file to your R working directory

- https://dl.dropboxusercontent.com/s/8j886jtexoeiv1v/imdb.rds

Step 1. Preprocess the imdb data using the following code

```
library(keras)
library(tidyverse)
set.seed(123)
n_sample <- 3000; maxlen <- 200; max_features <- 3000

imdb = read_rds("imdb.rds")
c(c(x_train, y_train), c(x_test, y_test)) %<-% imdb # Loads the data

x_train <- pad_sequences(x_train, maxlen = maxlen)
x_test <- pad_sequences(x_test, maxlen = maxlen)

sample_indicators = sample(1:nrow(x_train), n_sample)

x_train <- x_train[sample_indicators,] # use a subset of reviews for training
y_train <- y_train[sample_indicators]  # use a subset of reviews for training
x_test <- x_test[sample_indicators,] # use a subset of reviews for testing
y_test <- y_test[sample_indicators]  # use a subset of reviews for testing
```

Assignment 4

Step 2. Use `x_train` and `y_train` to fit the following deep learning models:

1. Simple RNN
2. LSTM
3. GRU
4. bidirectional LSTM
5. bidirectional GRU
6. 1D convnet

You can decide the parameters for the network structure (e.g., `units`, number of layers, etc) and model training (e.g., `epochs`, `batch_size` and `validation_split`).

However, you need to find parameters such that **the accuracy of each trained model on testing data should be at least 0.6**.

Assignment 4

Step 3. Save the following files

- Save each of these fitted models to an h5 file
- Save the history of each model to an rds file (see `write_rds`)
- Save `x_test` and `y_test` to rds files

Step 4. Save the R code you used for steps 1 to 3 to an R file

Step 5. Compress all the output files from step 3 to a zip file

Assignment 4

Step 6. Use R Markdown to achieve the following:

1. Specify author, date, and title in the YAML metadata of your document
2. Read all the output files from Step 3 (`x_test`, `y_test`, 6 fitted model files, and 6 training history files)
3. Use `x_test` and `y_test` to show the following statistics:
    - Number of reviews in the test set
    - Number of positive reviews in the test set
    - Number of negative reviews in the test set
4. For each model:
    - Show model summary
    - Plot the training history (Do NOT train your model in RMarkdown!)
    - Evaluate the performance of the model using the test set
5. Summarize the performance of different models using a table. Columns include
    - `model_name`
    - `acc`: Overall accuracy of the predictions in the test set
    - `n_tp`: Number of true-positive predictions in the test set
    - `n_tn`: Number of true-negative predictions in the test set
    - `n_fp`: Number of false-positive predictions in the test set
    - `n_fn`: Number of false-negative predictions in the test set
6. Discuss what you found from the table

Assignment 4

Here are some additional notes about writing a RMarkdown report. Violating these rules may lead to a lower grade.

- Put the data in the same folder as your Rmd file. Whenever we run/knit an RMarkdown file, it uses the folder with the Rmd file as the working directory.
- Read the data in your Rmd code chunk using relative path. If you use an absolute path, I will not be able to knit the Rmd file to an html file from my end.
- You will lose 5 points if for any reason (input path, error in code, etc.) the Rmd file cannot be knitted to an html file.
- All tables (any output of a data frame) must be formatted using kable in your R Markdown report.
- Distinguish headings (## heading) and normal text. We should not put all the text in headings.
- Do not put your discussions/explanations in code chunk. Write them as normal text.
- Do not use `include=FALSE` or `echo=FALSE` in your code chunk. I need to read your code. You may use `message=FALSE`, `warning=FALSE` to suppress messages/warnings.
- Do not write an excessively long line of code. Break it into multiple lines to improve readability.

Assignment 4

Step 7. Knit the R Markdown file (.Rmd) to an HTML file

Step 8. The R, Rmd, HTML, zip files must follow the naming rule below:

- `Assignment3-YourLastName.FileExtension`
  - For example:
  - Assignment4-Lin.R
  - Assignment5-Lin.Rmd
  - Assignment6-Lin.html
  - Assignment7-Lin.zip

Step 9. Submit the R, Rmd, html, and zip files (individually) to iCollege

Assignment 4

Due by the beginning of next class

Extra credit: the student who has the best report (determined by the instructor) will be given 5 extra points towards the final grade

- Submissions that are too similar would not be considered for the extra credit
- Accuracy of the models plays a significant role for this extra credit

Grading is based on the following:

- Grading is based on the submitted files on iCollege. Do not wait till the last minutes before the deadline. You will lose 10 points for late submission. You will receive 0 point if you submit your assignment via email.
- Whether all required files were submitted to iCollege on time, following the naming rule
- Whether the Rmd file is syntactically correct and can render the html file
- Whether the report has a professional format and style (succinct and yet provides adequate and clear discussions)
- Whether the report meets the requirements specified in Step 6