Final Project Data Memo

Jenny Do (5669841)

April 10, 2022

To make sure that each student is progressing in the final project for this course, and to verify that your data set(s) of choice will work for a machine learning project, everyone is asked to submit a check-in or "data memo."

An Overview of the Dataset

What does it include?

Duolingo surprisingly uses a lot of machine learning in their app.

Duolingo's dataset was used to produce a trainable spaced repetition model for language learning (particularly applications to second language acquisition). As stated on their READ.me, the model "marries psycholinguistic theory with modern machine learning techniques, indirectly estimating the"half-life" of words (and potentially any other item or fact) in a student's long-term memory". More information on the published research (Settles and Meeder, 2016) can be found at the following link: https://aclanthology.org/P16-1174.pdf.

Where and how will you be obtaining it? Include the link and source.

The dataset can be obtained from Harvard Dataverse (under the Duolingo Dataverse) as a gzipped CSV file containing the 13 million Duolingo student learning traces used in experiments by Settles & Meeder (2016). Link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N8XJME.

Below is my successful attempt in loading all 13 million data points into R using the fread() function. Note that though I was successful, for the purposes of this project, the dataset will be reduced to 1,000,000 observations (maybe less if my processor cannot handle it, hoping minimum will be 100,000 observations).

```
# using the fread() function to open the zipped csv file
library(data.table)
dt = fread("settles.acl16.learning_traces.13m.csv.gz")
```

```
# A taste of what the dataset looks like
head(dt)
```

```
##    p_recall  timestamp    delta user_id learning_language ui_language
## 1:      1.0 1362076081 27649635    u:FO                de          en
## 2:      0.5 1362076081 27649635    u:FO                de          en
## 3:      1.0 1362076081 27649635    u:FO                de          en
## 4:      0.5 1362076081 27649635    u:FO                de          en
## 5:      1.0 1362076081 27649635    u:FO                de          en
## 6:      1.0 1362076081 27649635    u:FO                de          en
##                          lexeme_id                   lexeme_string
## 1: 76390c1350a8dac31186187e2fe1e178 lernt/lernen<vblex><pri><p3><sg>
```

```
## 2: 7dfd7086f3671685e2cf1c1da72796d7    die/die<det><def><f><sg><nom>
## 3: 35a54c25a2cda8127343f6a82e6f6b7d         mann/mann<n><m><sg><nom>
## 4: 0cf63ffe3dda158bc3dbd55682b355ae         frau/frau<n><f><sg><nom>
## 5: 84920990d78044db53c1b012f5bf9ab5   das/das<det><def><nt><sg><nom>
## 6: 56429751fdaedb6e491f4795c770f5a4    der/der<det><def><m><sg><nom>
##    history_seen history_correct session_seen session_correct
## 1:            6               4            2               2
## 2:            4               4            2               1
## 3:            5               4            1               1
## 4:            6               5            2               1
## 5:            4               4            1               1
## 6:            4               3            1               1
```

About how many observations? How many predictors? What types of variables will you be working with?

Currently, there are 1,000,000 observations in the dataset used for this project and 12 predictors. More details and replication source code can be found at the following link: https://github.com/duolingo/halflife-regression

```
# take a random sample of size 1000000 from a dataset mydata
# sample without replacement
set.seed(27)
test_sample <- dt[sample(1:nrow(dt), 1000000,
   replace=FALSE),]
head(test_sample)
```

```
##    p_recall  timestamp  delta user_id learning_language ui_language
## 1:        1 1363104673    352  u:fy7_                en          es
## 2:        1 1363029684 123796  u:i-6y                es          en
## 3:        1 1362814150  22331  u:i4kF                de          en
## 4:        1 1362198465 437039  u:ir3L                de          en
## 5:        1 1363041254 332967  u:fOJV                en          es
## 6:        1 1363027700    410  u:dX5P                en          es
##                          lexeme_id              lexeme_string history_seen
## 1: 0d9244f805fd55af1281a1bcb1a2cba6            and/and<cnjcoo>           18
## 2: 389d16a7713404bf18de999b57a54eae          pan/pan<n><m><sg>            4
## 3: 46cd21495035f225d4fe8fc8552badc0 apfel/apfel<n><m><sg><acc>            5
## 4: 768f89b2bf27a38b8ae3e9ac849f5576 käse/käse<n><m><sg><*case>            1
## 5: a617ed646a251e339738ce62b84e61ce        are/be<vbser><pres>           23
## 6: 827a8ecb89f9b59ac5c29b620a5d3ed6          a/a<det><ind><sg>           52
##    history_correct session_seen session_correct
## 1:              17            3               3
## 2:               3            2               2
## 3:               4            1               1
## 4:               1            1               1
## 5:              23            2               2
## 6:              51            1               1
```

The predictors of the dataset are the following:

- p_recall - proportion of exercises from this lesson/practice where the word/lexeme was correctly recalled

- timestamp - UNIX timestamp of the current lesson/practice
- delta - time (in seconds) since the last lesson/practice that included this word/lexeme
- user_id - student user ID who did the lesson/practice (anonymized)
- learning_language - language being learned
- ui_language - user interface language (presumably native to the student)
- lexeme_id - system ID for the lexeme tag (i.e., word)
- lexeme_string - lexeme tag (see below)
- history_seen - total times user has seen the word/lexeme prior to this lesson/practice
- history_correct - total times user has been correct for the word/lexeme prior to this lesson/practice
- session_seen - times the user saw the word/lexeme during this lesson/practice
- session_correct - times the user got the word/lexeme correct during this lesson/practice

Note user_id, learning_language, ui_language, lexeme_id, lexeme_string are catagorical.

(Reference for subsetting the dataset to 1000000: https://www.statmethods.net/management/subset.html)

Is there any missing data? About how much? Do you have an idea for how to handle it?

```
sum(is.na(test_sample))
```

```
## [1] 0
```

Currently I believe there is no missing data. I am assuming that since the dataset was used for experimentation by Settles and Meeder in 2016, it will need minimal cleaning. However, if there turns out to be missing data, those observations will be removed (maybe later readded to see if it makes a significant difference).

An Overview of your Reseach Question(s)

What variable(s) are you interested in predicting? What question(s) are you interested in answering?

I am interested in predicting the variable p_recall (the proportion of exercises from this lesson/practice where the word/lexeme was correctly recalled). The main question I am interested in is – what drives a word to be recalled correctly. Does it have anything to do with the language (for example is Chinese harder to recall than French)/structure of the language? Does it purely have to do with the amount of times the user has seen the word? Does repitition in correctness matter?

Name your response/outcome variable(s) and briefly describe it/them.

The response variable will be p_recall (the proportion of exercises from this lesson/practice where the word/lexeme was correctly recalled).

Will these questions be best answered with a classification or regression approach? Which predictors do you think will be especially useful?

I believe the questions will be best answered in a regression approach (where categorical predictors are converted into a countinous range). The predictors delta (time (in seconds) since the last lesson/practice that included this word/lexeme), learning language(language being learned), lexeme_id (system ID for the lexeme tag (i.e., word)), lexeme_string (lexeme tag - string representation), history_seen (total times user has seen the word/lexeme prior to this lesson/practice), history_correct(total times user has been correct for the word/lexeme prior to this lesson/practice), session_seen(times the user saw the word/lexeme during this lesson/practice) and session_correct(times the user got the word/lexeme correct during this lesson/practice) will be highly useful.

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

The goal of the model is a combination of predictive and inferential. This is because we are aiming to ask "what features are significant" (what really drives a word to be recalled) and are there any relationships between the outcome and the predictors. In addition, we are trying to see what "what combo of features fits best" (from what drives a word to be recalled is there a combination that provides the best amounts of recall).

Proposed Project Timeline

This follows very closely to the timeline that is proposed by the syllabus. I aim to be dedicating Saturdays and whenever I have free time throughout the week to the project.

Week 3	Submit data memo, Start running and writting up descriptive analysis, want to find out if really tidy
Week 4	Run models, run results, work on drafts of paper (making sure to go to office hours)
Week 5	Run models, run results, work on drafts of paper (making sure to go to office hours)
Week 6	Run models, run results, work on drafts of paper (making sure to go to office hours)
Week 7	Run models, run results, work on drafts of paper (making sure to go to office hours)
Week 8	Run models, run results, work on drafts of paper (making sure to go to office hours)
Week 9	Make edits in prep for final
Week 10	Reading through and preparing for turn in
Due on June 6th	

Figure 1: Proposed Project Timeline - Jenny Do

Questions or Comments

Are there any problems or difficult aspects of the project you anticipate?

I am worried that what my project is trying to achieve and the amount of predictors I want to include is too much. In addition, I have concerns on whether or not a million observations will crash my system. As of right now I believe it will work as I can open the dataset with little to no wait time.

Regardless of the concerns I have, I am really excited to be starting this project!