IST 687: Text Mining

Corey Jackson

2020-09-02 19:55:25

Agenda

- Announcements
- Final Project
- Revisit Week 9 Topics/Lab/Homework
- Beakout Rooms
  - Project Update III
  - Complete Lab 10: Text Mining
- Tips for Homework 10

Announcements

- Office Hours: Wed. 6-7pm EDT, after class, by appointment
- Final submission for HW/Lab (incl. Homework/Lab 10)
  Monday, September 14th at 11:59 pm ET
- Mid-term grades/feedback on LMS
  - Your section: N $= 19$, $\mu = 94.2\%$

If you want feedback about specific questions on the mid-term, schedule office hours.

Final Project

- 21% of course grade (7% in-class presentation, 14% project summary description)
- Final Project Documents Due: **Tuesday, September 15th, 11:59pm ET, Submit on LMS**
- All members should participate in presentation and contribute to project report
- Feedback and Evalations
  - Instructor Feedback
  - Audience Feedback
  - Group Evaluation (2% course grade)

Final project deliverables: In-class presentation

- Presentation template (Due: **Tuesday, September 8th, 11:59pm ET, Post slide deck on SLACK**, do not submit on LMS)
    - 8-10 minute presentation
    - 5-10 minutes Q&A

Presentation tips: Defing the audience, provide context/motivation for the problem/research questions, describe the dataset, describe the methods, report the results, and close with major takeaways for the audience

Final Project deliverables: Project Report

- Project Report template (Due: **Tuesday, September 15th, 11:59pm ET**, Submit via LMS)
  - Written description of the project with more details than the presentation (5-10 pages).
  - Code can be in-line or seperate document/Github link.
  - Each member should complete a statement of contribution.

Summary document tips: Write concisely, use a text editor with spell check, make research questions explicit, label figures with captions and reference figures in text.

Week 9

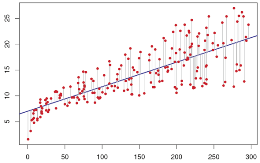
Homework 9 Overview: Support Vector Machines (SVM)

- Creating test/training sets
- Build prediction models for both regression problems having continuious outcomes (e.g., 23.12, 43.54) and classification problems having discrete outcomes (e.g., yes/no or 0/1)
  - **Continuious**: `lm()`, `svm()`, `ksvm()`
  - **Discrete**: `svm()`, `naiveBayes()`, `ksvm()`
- Evaluating model performance by computing error (for continuious) and classification rate (discrete outcomes)
- SVM can “learn”" linear functions.

Week 9 Homework: Computing error

```
lm <- lm(formula = Ozone~.,data=trainData_Corey)
predLm <- predict(lm, testData_Corey)
compTable3 <- data.frame(testData_Corey[,1], predLm)
```

```
##         test     Pred
## 67 40.00000 55.65515
## 61 42.12931 52.98951
## 49 20.00000 18.16127
```



Computing the the Root Mean Squared Error (RMSE)

```
##         test     Pred       diff
## 67 40.00000 55.65515 -15.655153
## 61 42.12931 52.98951 -10.860198
## 49 20.00000 18.16127   1.838726
```

```
sqrt(mean((compTable2$test-compTable2$Pred)^2))
```

```
## [1] 23.78958
```

- The model with the lowest RMSE is the model that best predicts the dependent variable

Modeling a discrete outcome

Step 1. Need to convert the continuious ozone varible to a discrete outcome variable `air$goodOzone <- ifelse(air$Ozone< mean(air$Ozone), 0, 1)`
`air$goodOzone <- as.factor(air$goodOzone)` # convert from numeric to factor

Step 2. Again, create test and training datasets

Step 3. Train model (same as above, with new dependent variable `goodOzone`)

Computing a confusion matrix for discrete outcomes

Step 4. Evaluate the model using predict and compute model accuracy

```
goodPred <- predict(nb, testData_Corey)
compGood1 <- data.frame(testData_Corey[,6], goodPred)
colnames(compGood1) <- c("test","Pred")
```

```
##   test Pred
## 1    0    1
## 2    1    1
## 3    0    0
## 4    1    1
## 5    1    1
## 6    1    1
```

Computing classification rate

- Compare the actual and predicted values

```
compGood1$result <-
ifelse(compGood1$test==compGood1$Pred,1,0)
```

```
##   test Pred result
## 1    0    1      0
## 2    1    1      1
## 3    0    0      1
## 4    1    1      1
## 5    1    1      1
## 6    1    1      1
```

- Compute agreement `sum(compGood1$result)/dim(compGood1)[1]` which is 0.7647059

Week 10

Week 10: Text mining

- Extracting meaning from text
  - **Word/document frequencies**: e.g., tf/idf (a measure how important a word is to a document), wordclouds
  - **Topic Modeling**: extracting higher groupings from text
  - **Sentiment analysis**: identifying and categorizing opinions expressed in text
- A resource for text mining:Text Mining with R

Lab 10: Text Mining

Lab 10 Overview

- Goal: Obtain experience using standard text mining procedures to obtain insight from text data.
  1. Importing and munging text from a Martin Luther King Speech
  2. Extracting valence (e.g., positive/ negative) sentiment from text.
- Packages needed for today's lab: tm wordcloud

Lab 10 Overview II

- Useful functions for today's lab: `match()`, `readLines()`, `scan()`, `rowSums()`
  - `scan(vector, character(0), sep = "\n")` (Step 1)
  - `readLines(path)` (Step 2)
  - `match(vector, vector, nomatch = 0)` (Step 3 & 4)
- Step 2 #Create a term matrix (Check chapter 14 where sba is transformed)

Lab 10 Overview III

- Creating 25% cutpoints for the corpus (Step 5).
- How to determine which words should be taken in each quarter

```
##  [1] "dream"        "president"    "anger"        "allu
##  [5] "school"       "Washington"   "shop"         "misch
##  [9] "capitol"      "constitution" "black"        "chil
```

```
cutpoint <- round(length(words)/4)
```

```
## [1] 3
```

```
words[1:cutpoint]
```

```
## [1] "dream"     "president" "anger"
```

Lab 10 Overview III

How might we capture the next quarter of the words in the word vector?

```
words[(cutpoint+1):(cutpoint*2)]
```

```
## [1] "allude"      "school"      "Washington"
```

Homework 10 Tips

Homework 10 Tips: Text Mining

- Build on Lab 10 to compute valance scores for the entire speech and 4 quarters

```
##          Word Score
## 1     abandon    -2
## 2   abandoned    -2
## 3    abandons    -2
## 4    abducted    -2
## 5   abduction    -2
## 6 abductions    -2
```

Datasets: MLK Speech and the AFFIN wordlist

Packages needed: `readr`, `tm`
More about AFFIN

Next week

Asynchronous Materials

- No videos/readings
- Continute working on final project

Live Session

- Presentations
- Closing remarks