AUTOMATING REVIEWS IN MEDICINE

The medical literature is enormous. Pubmed, a database of medical publications maintained by the U.S. National Library of Medicine, has indexed over 23 million medical publications. Further, the rate of medical publication has increased over time, and now there are nearly 1 million new publications in the field each year, or more than one per minute.

The large size and fast-changing nature of the medical literature has increased the need for reviews, which search databases like Pubmed for papers on a particular topic and then report results from the papers found. While such reviews are often performed manually, with multiple people reviewing each search result, this is tedious and time consuming. In this problem, we will see how text analytics can be used to automate the process of information retrieval.

The dataset consists of the titles (variable *title*) and abstracts (variable *abstract*) of papers retrieved in a Pubmed (http://www.ncbi.nlm.nih.gov/pubmed) search. Each search result is labeled with whether the paper is a clinical trial testing a drug therapy for cancer (variable *trial*). These labels were obtained by two people reviewing each search result and accessing the actual paper if necessary, as part of a literature review of clinical trials testing drug therapies for advanced and metastatic breast cancer.

PROBLEM 1.1 - LOADING THE DATA (1 point possible)

Load clinical_trial.csv (/c4x/MITx/15.071x/asset/clinical_trial.csv) into a data frame called trials (remembering to add the argument stringsAsFactors=FALSE), and investigate the data frame with summary() and str().

Some students have been getting errors like "invalid multibyte string" when performing certain parts of this homework question. If this is happening to you, use the argument fileEncoding="latin1" when reading in the file with read.csv. This should cause those errors to go away.

We can use R's string functions to learn more about the titles and abstracts of the located papers. The nchar() function counts the number of characters in a piece of text. Using the nchar() function on the variables in the data frame, answer the following questions:

How many characters are there in the longest abstract?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.2 - LOADING THE DATA (1 point possible)

How many search results provided no abstract? (HINT: A search result provided no abstract if the number of characters in the abstract field is zero.)

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.3 - LOADING THE DATA (1 point possible)

What is the shortest title of any article? Include capitalization and punctuation in your response, but don't include the quotes.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.1 - PREPARING THE CORPUS (2 points possible)

Because we have both title and abstract information for trials, we need to build two corpera instead of one. Name them corpusTitle and corpusAbstract.

Following the commands from lecture, perform the following tasks. Make sure to perform them in this order.

1) Convert the title variable to corpusTitle and the abstract variable to corpusAbstract.

2) Convert corpusTitle and corpusAbstract to lowercase.

3) Remove the punctuation in corpusTitle and corpusAbstract.

4) Remove the English language stop words from corpusTitle and corpusAbstract.

5) Stem the words in corpusTitle and corpusAbstract (each stemming might take a few minutes).

6) Build a document term matrix called dtmTitle from corpusTitle and dtmAbstract from corpusAbstract.

7) Limit dtmTitle and dtmAbstract to terms with sparseness of at most 95% (aka terms that appear in at least 5% of documents).

8) Convert dtmTitle and dtmAbstract to data frames.

If the code length(stopwords("english")) does not return 174 for you, then please run the line of code in this file (/c4x/MITx/15.071x/asset/stopwords.txt), which will store the standard stop words in a variable called sw. When removing stop words, use tm_map(corpusTitle, removeWords, sw) and tm_map(corpusAbstract, removeWords, sw) instead of tm_map(corpusTitle, removeWords, stopwords("english")) and tm_map(corpusAbstract, removeWords, stopwords("english")).

How many terms remain in dtmTitle after removing sparse terms (aka how many columns does it have)?

How many terms remain in dtmAbstract?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 2.2 - PREPARING THE CORPUS (1 point possible)

What is the most likely reason why dtmAbstract has so many more terms than dtmTitle?

- ○ Abstracts tend to have many more words than titles
- ○ Abstracts tend to have a much wider vocabulary than titles
- ○ More papers have abstracts than titles

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.3 - PREPARING THE CORPUS (1 point possible)

What is the most frequent word stem across all the abstracts? Hint: you can use colSums() to compute the frequency of a word across all the abstracts.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.1 - BUILDING A MODEL (1 point possible)

We want to combine dtmTitle and dtmAbstract into a single data frame to make predictions. However, some of the variables in these data frames have the same names. To fix this issue, run the following commands:

colnames(dtmTitle) = paste0("T", colnames(dtmTitle))

colnames(dtmAbstract) = paste0("A", colnames(dtmAbstract))

What was the effect of these functions?

- ○ Removing the words that are in common between the titles and the abstracts.
- ○ Adding the letter T in front of all the title variable names and adding the letter A in front of all the abstract variable names.
- ○ Adding the letter T in front of all the title variable names that also appear in the abstract data frame, and adding an A in front of all the abstract variable names that appear in the title data frame.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.2 - BUILDING A MODEL (1 point possible)

Using cbind(), combine dtmTitle and dtmAbstract into a single data frame called dtm. As we did in class, add the dependent variable "trial" to dtm, copying it from the original data frame called trials. How many columns are in this combined data frame?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.3 - BUILDING A MODEL (1 point possible)

Now that we have prepared our data frame, it's time to split it into a training and testing set and to build regression models. Set the random seed to 144 and use the sample.split function from the caTools package to split dtm into data frames named "train" and "test", putting 70% of the data in the training set.

What is the accuracy of the baseline model on the training set? (Remember that the baseline model predicts the most frequent outcome in the training set for all observations.)

Show Answer *You have used 0 of 4 submissions*

PROBLEM 3.4 - BUILDING A MODEL (1 point possible)

Build a CART model called trialCART, using all the independent variables in the training set to train the model, and then plot the CART model. Just use the default parameters to build the model (don't add a minbucket or cp value). Remember to add the method="class" argument, since this is a classification problem.

What is the name of the first variable the model split on?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 3.5 - BUILDING A MODEL (1 point possible)

Obtain the training set predictions for the model (do not yet predict on the test set). Extract the predicted probability of a result being a trial (recall that this involves not setting a type argument, and keeping only the second column of the predict output). What is the maximum predicted probability for any result?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.6 - BUILDING A MODEL (1 point possible)

Without running the analysis, how do you expect the maximum predicted probability to differ in the testing set?

- The maximum predicted probability will likely be smaller in the testing set.
- The maximum predicted probability will likely be exactly the same in the testing set.
- The maximum predicted probability will likely be larger in the testing set.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.7 - BUILDING A MODEL (3 points possible)

For these questions, use a threshold probability of 0.5 to predict that an observation is a clinical trial.

What is the training set accuracy of the CART model?

What is the training set sensitivity of the CART model?

What is the training set specificity of the CART model?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 4.1 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

Evaluate the CART model on the testing set using the predict function and creating a vector of predicted probabilities predTest.

What is the testing set accuracy, assuming a probability threshold of 0.5 for predicting that a result is a clinical trial?

Show Answer *You have used 0 of 4 submissions*

PROBLEM 4.2 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

Using the ROCR package, what is the testing set AUC of the prediction model?

Show Answer *You have used 0 of 3 submissions*

PART 5: DECISION-MAKER TRADEOFFS

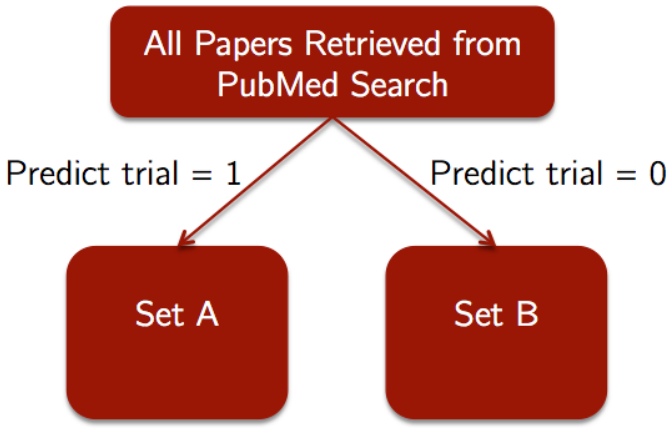
The decision maker for this problem, a researcher performing a review of the medical literature, would use a model (like the CART one we built here) in the following workflow:

1) For all of the papers retreived in the PubMed Search, predict which papers are clinical trials using the model. This yields some initial Set A of papers predicted to be trials, and some Set B of papers predicted not to be trials. (See the figure below.)

2) Then, the decision maker manually reviews all papers in Set A, verifying that each paper meets the study's detailed inclusion criteria (for the purposes of this analysis, we assume this manual review is 100% accurate at identifying whether a paper in Set A is relevant to the study). This yields a more limited set of papers to be included in the study, which would ideally be all papers in the medical literature meeting the detailed inclusion criteria for the study.

3) Perform the study-specific analysis, using data extracted from the limited set of papers identified in step 2.

This process is shown in the figure below.



PROBLEM 5.1 - DECISION-MAKER TRADEOFFS (1 point possible)

What is the cost associated with the model in Step 1 making a false negative prediction?

- A paper will be mistakenly added to Set A, yielding additional work in Step 2 of the process but not affecting the quality of the results of Step 3.
- A paper will be mistakenly added to Set A, definitely affecting the quality of the results of Step 3.
- A paper that should have been included in Set A will be missed, affecting the quality of the results of Step 3.
- There is no cost associated with a false negative prediction.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 5.2 - DECISION-MAKER TRADEOFFS (1 point possible)

What is the cost associated with the model in Step 1 making a false positive prediction?

- A paper will be mistakenly added to Set A, yielding additional work in Step 2 of the process but not affecting the quality of the results of Step 3.
- A paper will be mistakenly added to Set A, definitely affecting the quality of the results of Step 3.
- A paper that should have been included in Set A will be missed, affecting the quality of the results of Step 3.
- There is no cost associated with a false positive prediction.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 5.3 - DECISION-MAKER TRADEOFFS (1 point possible)

Given the costs associated with false positives and false negatives, which of the following is most accurate?

- ○ A false positive is more costly than a false negative; the decision maker should use a probability threshold greater than 0.5 for the machine learning model.
- ○ A false positive is more costly than a false negative; the decision maker should use a probability threshold less than 0.5 for the machine learning model.
- ○ A false negative is more costly than a false positive; the decision maker should use a probability threshold greater than 0.5 for the machine learning model.
- ○ A false negative is more costly than a false positive; the decision maker should use a probability threshold less than 0.5 for the machine learning model.

Show Answer *You have used 0 of 1 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

New Post