(https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▾

Courseware (/courses/MITx/15.071x/1T2014/courseware) **Course Info (/courses/MITx/15.071x/1T2014/info)**

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum) **Progress (/courses/MITx/15.071x/1T2014/progress)**

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

SEPARATING SPAM FROM HAM (PART 1)

Nearly every email user has at some point encountered a "spam" email, which is an unsolicited message often advertising a product, containing links to malware, or attempting to scam the recipient. Roughly 80-90% of more than 100 billion emails sent each day are spam emails, most being sent from botnets of malware-infected computers. The remainder of emails are called "ham" emails.

As a result of the huge number of spam emails being sent across the Internet each day, most email providers offer a spam filter that automatically flags likely spam messages and separates them from the ham. Though these filters use a number of techniques (e.g. looking up the sender in a so-called "Blackhole List" that contains IP addresses of likely spammers), most rely heavily on the analysis of the contents of an email via text analytics.

Help

In this homework problem, we will build and evaluate a spam filter using a publicly available dataset first described in the 2006 conference paper "Spam Filtering with Naive Bayes -- Which Naive Bayes?" by V. Metsis, I. Androutsopoulos, and G. Paliouras. The "ham" messages in this dataset come from the inbox of former Enron Managing Director for Research Vincent Kaminski, one of the inboxes in the Enron Corpus. One source of spam messages in this dataset is the SpamAssassin corpus, which contains hand-labeled spam messages contributed by Internet users. The remaining spam was collected by Project Honey Pot, a project that collects spam messages and identifies spammers by publishing email address that humans would know not to contact but that bots might target with spam. The full dataset we will use was constructed as roughly a 75/25 mix of the ham and spam messages.

The dataset contains just two fields:

- **text**: The text of the email.
- **spam**: A binary variable indicating if the email was spam.

IMPORTANT NOTE: This page is Part 1 of a double homework assignment. The second part of the homework assignment is on the next page, so remember to save your work so you can start the next page where you left off here.

PROBLEM 1.1 - LOADING THE DATASET (1 point possible)

Begin by loading the dataset emails.csv (/c4x/MITx/15.071x/asset/emails.csv) into a data frame called emails. Remember to pass the stringsAsFactors=FALSE option when loading the data.

How many emails are in the dataset?

□

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.2 - LOADING THE DATASET (1 point possible)

How many of the emails are spam?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.3 - LOADING THE DATASET (1 point possible)

Which word appears at the beginning of every email in the dataset? Respond as a lower-case word with punctuation removed.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.4 - LOADING THE DATASET (1 point possible)

Could a spam classifier potentially benefit from including the frequency of the word that appears in every email?

○ No -- the word appears in every email so this variable would not help us differentiate spam from ham.

○ Yes -- the number of times the word appears might help us differentiate spam from ham.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.5 - LOADING THE DATASET (1 point possible)

The nchar() function counts the number of characters in a piece of text. How many characters are in the longest email in the dataset (where longest is measured in terms of the maximum number of characters)?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.6 - LOADING THE DATASET (1 point possible)

Which row contains the shortest email in the dataset?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.1 - PREPARING THE CORPUS (1 point possible)

Follow the standard steps to build and pre-process the corpus:

1) Build a new corpus variable called corpus.

2) Using tm_map, convert the text to lowercase.

3) Using tm_map, remove all punctuation from the corpus.

4) Using tm_map, remove all English stopwords from the corpus.

5) Using tm_map, stem the words in the corpus.

6) Build a document term matrix from the corpus, called dtm.

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(corpus, removeWords, sw) instead of tm_map(corpus, removeWords, stopwords("english")).

How many terms are in dtm?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 2.2 - PREPARING THE CORPUS (1 point possible)

To obtain a more reasonable number of terms, limit dtm to contain terms appearing in at least 5% of documents, and store this result as spdtm (don't overwrite dtm, because we will use it in a later step of this homework). How many terms are in spdtm?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.3 - PREPARING THE CORPUS (1 point possible)

Build a data frame called emailsSparse from spdtm, and use the make.names function to make the variable names of emailsSparse valid.

colSums() is an R function that returns the sum of values for each variable in our data frame. Our data frame contains the number of times each word stem (columns) appeared in each email (rows). Therefore, colSums(emailsSparse) returns the number of times a word stem appeared across all the emails in the dataset. What is the word stem that shows up most frequently across all the emails in the dataset? Hint: think about how you can use sort() or which.max() to pick out the maximum frequency.

Show Answer *You have used 0 of 4 submissions*

PROBLEM 2.4 - PREPARING THE CORPUS (1 point possible)

Add a variable called "spam" to emailsSparse containing the email spam labels.

How many word stems appear at least 5000 times in the ham emails in the dataset? Hint: in this and the next question, remember not to count the dependent variable we just added.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.5 - PREPARING THE CORPUS (1 point possible)

How many word stems appear at least 1000 times in the spam emails in the dataset?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.6 - PREPARING THE CORPUS (1 point possible)

The lists of most common words are significantly different between the spam and ham emails. What does this likely imply?

○ The frequencies of these most common words are unlikely to help differentiate between spam and ham.

○ The frequencies of these most common words are likely to help differentiate between spam and ham.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.7 - PREPARING THE CORPUS (1 point possible)

Several of the most common word stems from the ham documents, such as "enron", "hou" (short for Houston), "vinc" (the word stem of "Vince") and "kaminski", are likely specific to Vincent Kaminski's inbox. What does this mean about the applicability of the text analytics models we will train for the spam filtering problem?

○ The models we build are still very general, and are likely to perform well as a spam filter for nearly any other person.

○ The models we build are personalized, and would need to be further tested before use as spam filters for other.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.1 - BUILDING MACHINE LEARNING MODELS (3 points possible)

First, convert the dependent variable to a factor with "emailsSparse$spam = as.factor(emailsSparse$spam)".

Next, set the random seed to 123 and use the sample.split function to split emailsSparse 70/30 into a training set called "train" and a testing set called "test". Make sure to perform this step on emailsSparse instead of emails.

Using the training set, train the following three machine learning models. The models should predict the dependent variable "spam", using all other available variables as independent variables. Please be patient, as these models may take a few minutes to train.

1) A logistic regression model called spamLog. You may see a warning message here - we'll discuss this more later.

2) A CART model called spamCART, using the default parameters to train the model (don't worry about adding minbucket or cp). Remember to add the argument method="class" since this is a binary classification problem.

3) A random forest model called spamRF, using the default parameters to train the model (don't worry about specifying ntree or nodesize). Directly before training the random forest model, set the random seed to 123 (even though we've already done this earlier in the problem, it's important to set the seed right before training the model so we all obtain the same results. Keep in mind though that on certain operating systems, your results might still be slightly different).

For each model, obtain the predicted spam probabilities for the **training set**. Be careful to obtain probabilities instead of predicted classes, because we will be using these values to compute training set AUC values. Recall that you can obtain probabilities for CART models by not passing any type parameter to the predict() function, and you can obtain probabilities from a random forest by adding the argument type="prob". For CART and random forest, you need to select the second column of the output of the predict() function, corresponding to the probability of a message being spam.

You may have noticed that training the logistic regression model yielded the messages "algorithm did not converge" and "fitted probabilities numerically 0 or 1 occurred". Both of these messages often indicate overfitting and the first indicates particularly severe overfitting, often to the point that the training set observations are fit perfectly by the model. Let's investigate the predicted probabilities from the logistic regression model.

How many of the training set predicted probabilities from spamLog are less than 0.00001?

How many of the training set predicted probabilities from spamLog are more than 0.99999?

How many of the training set predicted probabilities from spamLog are between 0.00001 and 0.99999?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 3.2 - BUILDING MACHINE LEARNING MODELS (1 point possible)

How many variables are labeled as significant (at the p=0.05 level) in the logistic regression summary output?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.3 - BUILDING MACHINE LEARNING MODELS (1 point possible)

How many of the word stems "enron", "hou", "vinc", and "kaminski" appear in the CART tree? Recall that we suspect these word stems are specific to Vincent Kaminski and might affect the generalizability of a spam filter built with his ham data.

Show Answer *You have used 0 of 2 submissions*

PROBLEM 3.4 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set accuracy of spamLog, using a threshold of 0.5 for predictions?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.5 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set AUC of spamLog?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.6 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set accuracy of spamCART, using a threshold of 0.5 for predictions? (Remember that if you used the type="class" argument when making predictions, you automatically used a threshold of 0.5. If you did not add in the type argument to the predict function, the probabilities are in the second column of the predict output.)

Show Answer *You have used 0 of 5 submissions*

PROBLEM 3.7 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set AUC of spamCART? (Remember that you have to pass the prediction function predicted probabilities, so don't include the type argument when making predictions for your CART model.)

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.8 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set accuracy of spamRF, using a threshold of 0.5 for predictions? (Remember that your answer might not match ours exactly, due to random behavior in the random forest algorithm on different operating systems.)

Show Answer *You have used 0 of 5 submissions*

PROBLEM 3.9 - BUILDING MACHINE LEARNING MODELS (1 point possible)

What is the training set AUC of spamRF? (Remember to pass the argument type="prob" to the predict function to get predicted probabilities for a random forest model. The probabilities will be the second column of the output.)

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.10 - BUILDING MACHINE LEARNING MODELS (1 point possible)

Which model had the best training set performance, in terms of accuracy and AUC?

○ Logistic regression
○ CART
○ Random forest

Show Answer *You have used 0 of 1 submissions*

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

Obtain predicted probabilities for the testing set for each of the models, again ensuring that probabilities instead of classes are obtained.

What is the testing set accuracy of spamLog, using a threshold of 0.5 for predictions?

Show Answer *You have used 0 of 3 submissions*

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

What is the testing set AUC of spamLog?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.3 - EVALUATING ON THE TEST SET (1 point possible)

What is the testing set accuracy of spamCART, using a threshold of 0.5 for predictions?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.4 - EVALUATING ON THE TEST SET (1 point possible)

What is the testing set AUC of spamCART?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.5 - EVALUATING ON THE TEST SET (1 point possible)

What is the testing set accuracy of spamRF, using a threshold of 0.5 for predictions?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.6 - EVALUATING ON THE TEST SET (1 point possible)

What is the testing set AUC of spamRF?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.7 - EVALUATING ON THE TEST SET (1 point possible)

Which model had the best testing set performance, in terms of accuracy and AUC?

○ Logistic regression
○ CART
○ Random forest

Show Answer *You have used 0 of 1 submissions*

PROBLEM 4.8 - EVALUATING ON THE TEST SET (1 point possible)

Which model demonstrated the greatest degree of overfitting?

○ Logistic regression
○ CART
○ Random forest

Show Answer *You have used 0 of 1 submissions*

IMPORTANT NOTE: This page is Part 1 of a double homework assignment. The second part of the homework assignment is on the next page, so remember to save your work so you can start the next page where you left off here.

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

Show Discussion

New Post

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.