



MITx 6.86x

Machine Learning with Python-From Linear Models to Deep Learning

[Help](#)

smitha_kannur ▾

[Course](#)

[Progress](#)

[Dates](#)

[Discussion](#)

[Resources](#)

[Home](#) [Course](#) / [Unit 1 Linear Classifiers and Generalizations \(2 weeks\)](#) / [Project 1: Automatic Review Analyzer](#)

[< Previous](#)



[Next >](#)

9. Feature Engineering

[Bookmark this page](#)

Project due Oct 7, 2020 05:29 IST *Completed*

Frequently, the way the data is represented can have a significant impact on the performance of a machine learning method. Try to improve the performance of your best classifier by using different features. In this problem, we will practice two simple variants of the bag of words (BoW) representation.

Remove Stop Words

1/1 point (graded)

Try to implement stop words removal in your feature engineering code. Specifically, load the file **stopwords.txt**, remove the words in the file from your dictionary, and use features constructed from the new dictionary to train your model and make predictions.

Compare your result in the **testing** data on Pegasos algorithm using $T = 25$ and $L = 0.01$ when you remove the words in **stopwords.txt** from your dictionary.

Hint: Instead of replacing the feature matrix with zero columns on stop words, you can modify the `bag_of_words` function to prevent adding stopwords to the dictionary

Accuracy on the test set using the original dictionary: 0.8020

Accuracy on the test set using the dictionary with stop words removed:

0.808



Submit

You have used 1 of 20 attempts

Change Binary Features to Counts Features

1/1 point (graded)

Again, use the same learning algorithm and the same feature as the last problem. However, when you compute the feature vector of a word, use its count in each document rather than a binary indicator.

Hint: You are free to modify the `extract_bow_feature_vectors` function to compute counts features.

Accuracy on the test set using the dictionary with stop words removed and counts features:

0.77



Try to compare your result to the last problem, and see the discussion in solution after answering the question.

Submit

You have used 1 of 20 attempts

Some additional features that you might want to explore are:

- Length of the text
- Occurrence of all-cap words (e.g. "AMAZING", "DON'T BUY THIS")
- Word embeddings

Besides adding new features, you can also change the original unigram feature set. For example,

- Threshold the number of times a word should appear in the dataset before adding them to the dictionary. For example, words that occur less than three times across the train dataset could be considered irrelevant and thus can be removed. This lets you reduce the number of columns that are prone to overfitting.

There are also many other things you could change when training your model. Try anything that can help you understand the sentiment of a review. It's worth looking through the dataset and coming up with some features that may help your model. Remember that not all features will actually help so you should experiment with some simpler ones before trying anything too complicated.

Discussion

Hide Discussion

Topic: Unit 1 Linear Classifiers and Generalizations (2 weeks);Project 1: Automatic Review Analyzer / 9. Feature Engineering

Add a Post

Show all posts	by recent activity
<p> Remove stop words - how to debug?</p> <p>Am I missing something on removing the stop words? When I removed the stop words from the dictionary, my training accuracy we...</p>	19
<p> I'm still not entirely sure why using count instead of binary representation decreased the accuracy?</p> <p>Even though we removed articles, verbs and pronouns from our bag of words model, why does count decrease accuracy? Is it speci...</p>	1
<p> Help with additional Python resources</p> <p>Could someone please recommend some additional sources to acquire further Python skills that are needed for projects like this on...</p>	3
<p> Poll: Average number of hours to complete Project 1</p> <p>Out of curiosity, how many hours did it take you to complete Project 1? In my case it took me ~15 hours.</p>	19
<p> Question about Recitation</p> <p>The value l2 in the parameter <code>**penalty='l2'</code>, where does it come from?</p>	2
<p> all methods are passing, but returning incorrect validation error</p> <p>All my methods are passing and correct, but the validation error is incorrect. I've also cleared global variables and reran my tests, bu...</p>	3
<p> Dear STAFF</p> <p>For those of us, who didn't manage to complete the section 9. Feature Engineering, would the respective code be available to us?</p>	1
<p> Project 1 is very well engineered</p> <p>Project 1 is very well engineered: Robust code, well organized, very didactic. A model of how questions should be stated. Congratula...</p>	20
<p> Change Binary Features to Counts Features</p> <p>What is meaning of document in this "However, when you compute the feature vector of a word, use its count in each document rat...</p>	8
<p> [To Staff] Problem 8 and 9</p> <p>I have managed to complete problem 8 and 9 by making slight adjustments to existing codes (e.g. <code>utils.py</code>, <code>project1.py</code>). Are we expe...</p>	2
<p> Mathematics for Machine Learning Book</p> <p>Hello For those interested, this books it's a great resource for maths behind machine learning and it has an excellent chapter about...</p> <p> Community TA</p>	7
<p> Tip: Change Binary to Count Features</p> <p>Using the Counter class from collections package made the task really simple for me. Just initialized the Counter class with the wordl...</p>	4
<p> Trouble with Removing stop words -- extract bow features function</p> <p>For removing the stop words, I set a new dictionary of length 12108, then when I go to use it, I get the following error for the extract...</p>	5

< Previous

Next >



edX

[About](#)
[Affiliates](#)
[edX for Business](#)
[Open edX](#)
[Careers](#)
[News](#)

Legal

[Terms of Service & Honor Code](#)
[Privacy Policy](#)
[Accessibility Policy](#)
[Trademark Policy](#)
[Sitemap](#)

Connect

[Blog](#)
[Contact Us](#)
[Help Center](#)
[Media Kit](#)
[Donate](#)



© 2020 edX Inc. All rights reserved.

深圳市恒宇博科技有限公司 [粤ICP备17044299号-2](#)