9. Feature Engineering

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:

0.8080 ✔

Submit You have used 1 of 20 attempts

✔ Correct (1/1 point)

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:

0.7700 ✔

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

✔ Correct (1/1 point)

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 ▼

💬	Change Binary Features to Counts Features I replaced the binary representation in the table for the number of times this word is included in the review. However, as a result, the accuracy became lower. I obtained a...	12
?	Could we have more weeks released in advance? I see the schedule, but the holiday season is coming.	11
?	The total words remain unchanged after loading the stopword_data Hi, for the first question. I loaded the stopwords by writing: stopword_data = utils.load_data('stopwords.txt') and modified the bag_of_words function to: def bag_of_words...	10
?	Try to compare your result to the last problem, and see the discussion in solution after answering the question? Where can I see this discussion? I've succesfully completed the problem, however I'd be happy to read the discussion. The problem is, I don't have a "Show Answer" butto...	1
?	When using Count features Hi, I am using count(word) for every word from word_list in every occurence of reviews in extract_bow_feature_vectors function. I am getting output which is almost 76% o...	1
?	Troubleshooting the code when answers marked wrong by grader for Stop Words I have applied Stop words check criteria on the dictionary and verified that length of dictionary earlier and later got changed. On applying the Pegasos algorithm with its b...	1
?	Change binary to count By count do you the number of times the word has appeared in the entire "reviews" list?	3
💬	Bag of words and stopwords versus negation Consider these two possible reviews: - "I really enjoy this coffee. I would buy it again." - "I really don't enjoy this coffee. I would not buy it again." Using single words rips th... Community TA	1
💬	There are periods in amongst the words. Punctuation is leaking in. There are periods in the dictionary... Highest count of periods in one review in the training data? 580. Somebody wrote an essay... Community TA	1
💬	It is great stuff! thanks	1
💬	Unit 2 Release Hi Admin, How about the unit 2? It's still not yet opened as the course schedule. Thanks	5
💬	parameter initialization I am just wondering if there are some better ways for parameter initialization instead of zeros. More specifically for linear SVM, consider binary classification. Will it help if ...	3

Learn About Verified Certificates