Solutions to Lab for week 7: Text Classification

Author: Sharon Goldwater
Date: 2017-10-30

This lab is available as a web page[2] or pdf document[3].

Feature extraction

- (Opt) Complete the function...

  See our answer code[4] .

- From train_pos_features, how can you get the feature count of the POS tag 'NNS' in the third training sentence?:

  ```
  j = pos_vocab.index("NNS")
  train_pos_features[2][j]
  ```

- In this toy data set, what is the prior probability of each class if we use Maximum Likelihood Estimation?

  Class 0: 2/5, class 1: 3/5

- What is the equation to compute the feature probabilities using add-alpha smoothing?

  See lecture slides.

- **(Opt)** Complete the function...

  See our answer code[4] .

- Do the results seem reasonable? Do you think those features are salient and discriminative?

  Some of the features seem reasonable, like "furniture" and "fresh". But some of them do not seem either salient or discriminative, like "and" and "are". This is because words that have high probability given the class may or may not help to discriminate between classes.

- Are they similar or different to the most probable features of each class in the Naive Bayes model? Would you expect them to be similar? Why or why not?

  The most influential features seem much more reasonable. That's because these are the features with highest weights, which means they are the ones that help to discriminate between classes.

- Are the most probable features in a generative model likely to be discriminative?

  Not necessarily, because all classes could have the same high probability features. (However it's also possible for them to be discriminative: they could be high probability in all classes, but still have different probabilities. In this case, since they occur frequently, they may have a big influence on the final decision.)

[1] http://creativecommons.org/licenses/by-nc/4.0/.

[2] http://www.inf.ed.ac.uk/teaching/courses/anlp/labs/lab7.html

[3] http://www.inf.ed.ac.uk/teaching/courses/anlp/labs/lab7.pdf

[4] lab7_toy-sol.py

- Will removing stopwords reduce the size of the vocabulary by much?

  No, because the vocabulary size is very large, and there are only a relatively small number of stopwords.

- Will removing stopwords change the performance of the classifiers by much?

  Probably not, because stopwords are not likely to be discriminative for this task.

- How many features are there now? [after variance thresholding] Did this big reduction affect performance much? Can you detect any change in how fast the models run?

  The number of features should be much smaller now. However, the performance only decreases slightly. You should notice some decrease in the time to train the logistic regression model. (It would probably be more noticeable if we were more careful about timing.)