EECS 348 Programming Assignment 4 Reflection

Alan Fu, Jiaming Li

2015-05-18

bayes.py vs bayesbest.py performance

	precision	recall	f-value
bayes.py	0.723427515842	0.842982546918	0.778642558948
bayesbest.py	0.81834775366	0.885399810563	0.850554344177

bayes.py achieves around 70%-80% precision and recall accuracy, which is reasonably good performance. bayes.py works because of the fundamental bayes theorem. bayesbest.py on the other hand performs noticeably better than bayes.py because of the following improvements: first, bayesbest.py assumes the same words in different cases (lowercase, all-capped, first-letter-capped) are still the same word and count for the same key in the dictionary; this assumption is generally correct in modern English writing and allows our dictionaries to hold more relevant values. Second, we assume that different words should have different "weights" in terms of their contribution to the reviews' sentiment; therefore we use gradient ascent method to dynamically alter the word count for words in dictionary so that different words' "weights" can be reflected during classification (for the specific implementation please refer to function train() and gradientDescent() in bayesbest.py).

bayesbest.py may be further-improved by the following extensions:

- Larger training data set: the movie reviews folder contains around 13,000 reviews, each with short to medium length. Considering how many English words there are that might be used in a movie review, 13,000 reviews can fall short in training the dictionaries properly.
- Synonyms: many adjectives mean the same thing (e.g. "awesome", "good", "amazing", and "fantastic" are interchangeable). If we can identify synonyms in a review and use them to count towards the common key word, we can virtually have more training data.

- Grammatical features: naïve bayes classifier ignores grammar completely, which may results in misclassification of reviews that contains subtle and not-so-subtle phrases such as “this movie is a good try but not a good implementation”. Focus on grammar analytics of the reviews might improve classification performance.