Chapter 5

Working with Text

July 20, 2016

- Natural Language Processing (NLP)
- Sentiment analysis (aka opinion mining)
- Document polarity (e.g. positive vs. negative)
- IMDB dataset
  - 50,000 movie reviews labeled as positive/negative
  - Positive: more than six stars on IMDB
  - Negative: fewer than five stars on IMDB
- Predict automatically whether the reviewer liked the movie

Bag-of-words models

- Idea: represent text as numerical feature vectors
- Create a vocabulary (alphabet) of unique tokens (e.g. words)
- Assign an integer index to each token
- Construct a sparse feature vector
- CountVectorizer example

N-grams

- Unigrams and bigrams
- The sun is shining
- Unigrams: the, sun, is, shining
- Bigrams: the sun, sun is, is shining
- CountVectorizer can extract any n-grams
- Tf-idf sometimes works better than row counts

NLP bag-of-tricks

- Data cleaning to remove noisy tokens (e.g. HTML tags)
- Stemming (e.g. running - run)
- Lemmatization (e.g. went - to go)
- Stop-word removal
- Open-source libraries, e.g. NLTK and OpenNLP
- Details in text (should be useful for projects)
- 90% accurate logistic regression example
- Out-of-core learning possible in scikit-learn