COS 424
Foundations of Machine Learning

Assignment 1: Classification of review sentiments

Sentiment Analysis

- ‘Sentiment’ here refers to opinion, assumed to be binary (positive/negative)
- Sentiment analysis uses NLP/ML to extract these opinions from free text
- Applications: analyzing Facebook and Twitter feeds to propagation of sentiment/characterizing change in sentiment over time
- Challenges: brevity, slang, context, negation, sarcasm…

Project Definition

- Dataset: 3000 reviews from Yelp/Amazon/iMDB, split into 2400 in train.txt and 600 in test.txt; each review classified as positive (1) or negative (0)
- Scripts: *preprocessSentences.py* provided as potential starting point for cleaning, tokenizing and extracting bag-of-words
- Task:
    - Using labelled reviews in train.txt, learn a classifier to that distinguishes positive and negative reviews
    - Extract features: script *preprocessSentences.py* provided as potential starting point for cleaning, tokenizing and extracting bag-of-words
    - Train different classifiers: e.g. Naive Bayes/logistic regression/SVM
    - Report performance on 600 held-out reviews in test.txt (ROC curves, etc)

Possible extensions

- More data
  - Twitter: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
  - Movie reviews: https://www.cs.cornell.edu/people/pabo/movie-review-data/
  - Amazon reviews: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
- More features
- More complex/ensemble classifiers
- Binary vs subjective (scalar) sentiment classification
- Beyond classification: unsupervised problems, eg clustering or topic modeling

Deliverables

- Five-page summary:
  - Introduction
  - Description of data, methods [inc 1 page in-depth description of a classifier of interest]
  - Presentation of results
  - Summary, conclusion & extensions
  - Bibliography
- Due: 5pm Feb 28 at CS Dropbox: at https://dropbox.cs.princeton.edu/COS424_S2017/Assignment1