HW2: Amazon Review Classification

Published Date:
Feb. 21, 2017, 4:30 p.m.

Deadline Date:
Mar. 7, 2017, 4:30 p.m.

Description:

****
This is an individual assignment and Deadline is 04/07/2017 4:30 PM.
****

Overview and Assignment Goals:

The objective of this assignment are the following:

- Implement the Nearest Neighbor Classification algorithm
- Handle text data (reviews of Amazon baby products)
    - Design and engineer features from text data.
- Choose the best model, i.e., parameters of a nearest neighbor classifier, features and similarity functions

---

Detailed Description:

A practical application in e-commerce applications is to infer sentiment (or polarity) from free form review text submitted for a range of products.

For this assignment, you have to implement a k-Nearest Neighbor Classifier to predict the sentiment for 18506 baby product reviews provided in the test file (test.dat). *Positive sentiment* is represented by a review rating of +1 and *negative sentiment* is represented by a review rating of -1. In test.dat you are only provided the reviews but no ground truth rating. These data will be used for comparing your predictions.

Training data consists of 18506 reviews as well, provided in the file train.dat. Each row begins with the sentiment score followed by the text associated with that rating.

For evaluation purposes (Leaderboard Ranking) we will use the Accuracy metric comparing the predictions submitted by you on the test set with the ground truth. Some things to note:

- The public leaderboard shows results for 50% of randomly chosen test instances only. This is a standard practice in data mining challenges to avoid gaming of the system. The private leaderboard will be released after the submission deadline, based on all the entries in the test set.
- In a 24-hour cycle, you are allowed to submit a prediction file only 5 times.

- The final ranking will always be based on the last submission.

format.dat shows an example file containing 18506 rows alternating with +1 and -1. Your test.dat should be similar to format.dat with the same number of rows (18506), but containing the sentiment score generated by your developed model.

Rules:

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will result in an honor code violation.
- Feel free to use the programming language of your choice for this assignment.
- While you can use libraries and templates for dealing with text data you should implement your own nearest neighbor classifier.

Deliverables:

- Valid submissions to the Leader Board website (TBA, will be posted on Canvas within a week)
- **Canvas Submission of source code and report:**
  - Create a folder called HW2_SJSU-ID
  - Create a subfolder called src and put all the source code there.
  - Create a subfolder called report and place a 2-page, single-spaced report describing details regarding the steps you followed for developing the classifier you used to predict the product review sentiments. Be sure to include the following in the report:
    1. Name and SJSU ID.
    2. Rank & Accuracy score for your submission (at the time of writing the report).
    3. Your approach
    4. Your methodology of choosing the approach and associated parameters.
  - Archive your parent folder (.zip or tar.gz) and submit via Canvas for HW2.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (20%) and ranking submissions (10%). Extra credit (1% of final grade) will be awarded to the top-3 performing algorithms and to the submission with the most interesting solution (to be judged by Prof. Anastasiu). Note that extra credit throughout the semester will be tallied outside of Canvas and will be added to the final grade at the end of the semester.

Files:

- *Train Data:* download
- *Test Data:* download
- *Format File:* download