CSE 446: Machine Learning - Project Dataset Writeup

Ryan Thomas (rthomas7), Varun Viswanash (varunv9), Jonathan Leang (jleang)

October 17, 2017

Dataset Description

This dataset contains information on businesses, reviewers, and reviews available on the Yelp website. We thought it would be interesting to have students use all the data other than review score to try and predict the number of stars that each restaurant has. Students will have access to the full text of thousands of reviews, and they'll also be able to see how each of these reviews were rated. This could encourage students to try and parse through the text used in the reviews to find common patterns in reviews of good restaurants vs bad restaurants. Students would also be presented with an interesting filtering challenge, since often there are reviews which are outliers and need to be ignored. We wanted to see who could figure out the best way of determining which reviews were worth filtering out. These factors should keep our dataset/challenge from being too easy. We also believe that our dataset/challenge won't be too difficult because there should be logical, intuitive connections between the factors such as the text of the reviews and how well the restaurant is rated. We know that the data is of high quality because it was curated by Yelp and has been used for interesting competitions in the past.

Dataset Training/Test Division

The data was split 80 to 20 for training versus test data. We decided to do this by the general recommendation in class to have 80% training data, 10% development data, and 10% final test data. The 80-20 split is appropriate for this dataset because the there is ample amounts of data but for the purpose of processability there is not enough to warrant a higher ration (ex: a 90-10 split).

Dataset Specification

Name: yelp
The input data is given as a .json file with features of the following types:

- String: name, neighborhood, address, city, state, postal code
- Float: latitude, longitude
- Integer: number of reviews
- Array of Tuples: The reviews themselves, along with the date of the review and how it was scored by other reviewers as "funny", "cool", and such
- Array of Business categories such as "Mexican", "Burgers", "Gastropubs"
- Map of hours open each day of the week

Ethical Evaluation

We know that the data is ethical to use because Yelp posted it publicly for people like us to use it. The information contained specifies the attributes of businesses and individual reviews that use Yelp as a service to advertise their brand and make their opinion known. Thus the true source of this information is also explicitly made public.