Long-term user rating prediction for business

Dataset description

The dataset to be used is provided by Yelp's dataset challenge, which includes: 1) 1.6M reviews and 500K tips by 366K users for 61K businesses. 2) 481K business attribute. 3) .Social network of 366K users for a total of 2.9M social edges. 4) Aggregated check-ins over time for each of the 61K businesses. The dataset is available for 10 different cities in U.K, Canada, Germany and U.S, including Urbana-Champaign.

Proposed Methods

Objective

The objective of the machine learning task is to predict overall user rating for a business in the future. The feature used for prediction would be based on information such as: 1) the business' own attributes. 2) current and past reviews for the business. 3) ratings and reviews for the surrounding business. etc.

Data preprocess

For training data generations, the idea is to spread data out as time series and draw a time boundary before which the data is used as features; and after which data is used to generate result (rating). For example, looking at business A's rating for 1 month after a pre-set date T and treat it as the correct rating prediction, and all data before date T can be used to generate features.

Learning method

We planned to use a conventional supervised learning method such linear regression. The specific algorithms to test will be determined after initial screening of the test data, and a evaluation of algorithms will be performed to select the most suitable one for this task.

Expected Results

We expect to discover the trends between businesses and their surrounding businesses. We look to predict how well a restaurant will review after opening in based on reviews of other

similar type of business in the area. We also expect to discover effects of a new restaurants reviews to predict future reviews of other restaurants of the same category. For example, opening up a new BBQ restaurant, how the population and the restaurant goer's expectations will change and whether review from a prior restaurant will be good or bad. This determines the underlying quality of a particular restaurant. In addition we hope to see correlation between specific terms in a review and whether a business will do well or not in the long run. We also hope to map correlations between reviews on a similar topic throughout time. One example is a review of service. If a restaurant has bad service when it first opens but corrects to better service later one, we can expect ratings to go up.