Final Project Ideas

We are given reviews by users on certain restaurants. Main idea is to use these predict ratings of the comment by the user.

Predict the actual value, or we just do good or bad with some rating like 2.5. Binary classification.

We can also do anything we learned about in this class.

Tools:

* Classification
  + Hinge Loss
  + Logistic Loss
* LDA QDA
* KNN classifier
* Train Validation Split
* Cross Validation
* K Fold Cross Validation
* LOOCV
* Linear Regression
  + Best Subset Selection
  + Forward/Backward selection
  + Shrinkage Methods (Lasso, Ridge)
  + Dimension Reduction (PCA)
* Classification Tree
* Bootstrapping
* Random Forest
* Text Mining
  + Hella Models
* Random Forests
* Support Vector Machines

Tree split on words and then predicting

Pre Process in Python

Take sum

Dimension reduction.

Linear, logistic regression, lasso ridge for words.

t.test in logistic regression that shows you which variables are important,

lasso, self

Report Parts

Yelp Dataset Challenge:

Abstract

1. Introduction
2. Define Background and Problem. Predicting Rating,
3. Define the related work, so basically the numericalization step
4. Describe Data
5. Trimmed Data set
6. Experimental Setup
   1. Prepocessing
      1. Get rid of unnecessary + variable transformations
      2. Separation into Positive and Negative
      3. Separation into Diff
      4. Elite turning into # of years of Elite
      5. Getting rid of useless info like state and town
   2. Model description
      1. Benefits and draw backs
      2. Bag of words, get models from friends
7. Results and Analysis
   1. Talk about the results of each one
   2. RMSE
   3. Comparison
   4. Plots
8. Conclusion and Future Work

Experiment

Conclusion and Summary

EDF: lots of nice plots, features of the data,

Choose k points out of n samples. Agglomerative clustering. Generate clusters in a hierarchical way. Start with n clusters. Each containing 1 data point. Merge the two clusters with minimum dat

Introduction

In today’s consumer focused world, user reviews are crucial aspects of the shopping process that allow people to better understand a businesses’ product and service. In fact, reviews are so important that a whopping 93% of consumers say reviews influence their purchasing decisions (Podium, 2017). However, in addition to consumers, businesses also heavily consider their own reviews to learn about themselves, seeing which areas and products need redevelopment and which are successful.

On most sites, such as Amazon and Yelp, user reviews have two main parts: a free-form text section in which users can voice their opinion on the product or service of the company, and a star rating from one to five. While many consumers will sometimes look to the opinions of others, the hundreds sometimes thousands of comments on the business are often too much to read for those that are making quick purchasing decisions. This is especially true for users making food purchases, who will often decide on which restaurants to eat at just by the average star rating of the business. Many consumers will not even engage with business with less than a 3.3 star rating (Podium, 2017).

Since the star rating is so integral to a businesses profit, the problem of predicting the star rating of reviews given the textual comment of users has become a popular, but difficult issue. Discovering the relationship between review and star can be complicated as users can give the same review and have differing opinions or have the same opinion and have different star reviews. For example, consider the case of two users giving the same star rating of 2/5. User A can give a 2/5 because they enjoyed the food but hated the service of the restaurant. User B can give a 2/5 because they thought well of the service, but disliked the food of the restaurant. As we can see there can be many facets with textual comments that increases the difficulty of setting a star rating to it.

However, while challenging many models have been developed to solve this issue of drawing out the sentiment from a piece of text (also called sentiment analysis). One specific model, first originating all the way back from Zellig Harri’s 1954 article *Distributional Structure*, is called the bag-of-words model. In most applications, the bag-of-words model is used as a tool of feature generation. By transforming textual reviews into quantifiable features, data scientists can then use these numerical values in many machine learning models to formulate relationships with other variables. The most common type of quantitative feature used in bag-of-words model is frequency, which is the number of times a term appears in the text. After defining a dictionary or bag of words that you want to keep track of, you can then create a vector of values that lists the frequency of each of the words in the review. This report will be using the bag-of-words model to transform textual reviews from Yelp into features that can be used to predict/classify the star rating of the review.

Yelp is a popular crowd-sourced review platform with millions of active users who rate