CS403: An Introduction to Machine Learning

Project Report

Restaurant Recommendation System

Team members (Group No. 9)

Nikhil Kumar - 140050037

Krishna Chaitanya - 140050038

Pavan Kumar - 140050045

Sai Teja - 140050059

ACKNOWLEDGMENT

We would like to acknowledge Prof. Ganesh Ramakrishnan for his wonderful lectures and sessions on Machine Learning, and this would not have been possible without the help of the TA's of the course, especially Shubham Jain and Dheeraj.

Introduction

Over the past decade the number of people that are making relying on online recommendation sites like zomato and yelp grew very sharply. Now with a lot of data available it is difficult for the consumers to make an informed choice by just going through the raw data. This is what we intend to achieve based on this project. It is both very time consuming to go through as well as difficult to make sense with so many user reviews available.

We want to build a recommendation system that helps the users choose a particular restaurant based on his preferences and reviews by other users with similar preferences. This saves the consumers the trouble of having to sort and filter the list of restaurants before choosing if he wants to dine in or otherwise.

PROPOSED IDEA FOR THE PROJECT

We wanted to implement three learning algorithms for the prediction of the query's recommendation, Kernel Ridge regression,Linear SVM, Logistic regression. And our idea was to use the validation data to chose with learning algorithm is the best based on the accuracies on the validation datas. To filter out the features which matter and affect the recommendation, we wanted to use Principal Component and Analysis and Random Forest techniques.

DESCRIPTION OF THE DATASET

Our primary dataset is obtained from the Yelp Dataset Challenge data.

This contains the both user level data as well as restaurant level data.For example user level features consist of entries like number of days in yelp, number of fans, number of votes,average stars,elite users, etc.

Restaurant level features are such as binary features for attributes (like Good for, caterers, noise level, reservation etc), categories (Fast food, Bar, Chinese, Italian ,Thai , etc), number of people frequenting the restaurant, etc. Reviews of users to different restaurants is also available.

We processed the dataset and gathered all those features that can be useful in our project that can potentially affect the output using PCA.We also converted categorical variables into boolean variables.

LEARNING ALGORITHMS THAT WE USED IN OUR PROJECT

We experimented with the following algorithms to train the rating predictor classifier:

Kernel Ridge regression

Linear SVM

Logistic regression

We have seen that Logistic Regression gives the best results of the three techniques suggested above.

FEATURE SELECTION TECHNIQUES AND GENERAL DESCRIPTION OF THE WORK DONE IN THE PROJECT

Initially, we have implemented a non user specific algorithm which just recommends (or not) the restaurant to the user based on the restaurant itself, i.e the user's tastes are not taken into consideration while giving the recommendation. We then added new features like averageRatingForEachCuisineForTheUser for a given (user, restaurant, review) tuple for all the cuisines available in the Restaurant categories. Then we decided to cut down on the number of cuisines (so as to avoid too many features) by finding the number of appearances (frequency) of the cuisines and choosing only the top few cuisines where the difference in frequency came out to be too high. We, thus chose 21 features like Chinese, Italian, Thai and so on, and added averageRatingForEachCuisineForTheUser for all these cuisines. We then decided to choose only a few features of the restaurant's details. We achieved this target by doing Principal Component Analysis, and choosing only the features along which account for at least 98 % of the variance. So, we cut down the number of user level features by using the most appearing cuisines and finding their average ratings by the user and adding them to the feature matrix, and then applied PCA on all the features. Then choosing the appropriate axes and projected the data on to required dimensions. We then used Logistic Regression over this feature matrix, as this gives the best results on the validation data.

FUTURE WORK POSSIBLE IN THE PROJECT

We can improve on the accuracy of the predictions by using some better features. We can take into account that the user's friends likings have a good relation with the user's interests, and thus we can find the friend's preferences from the data, and add into our feature matrix. Basically, we can engineer the features better and choose some features which will improve the accuracy of the predictions.

SOME OF THE FUNCTIONS THAT WE USED FOR DIFFERENT ALGORITHMS IN OUR CODE

Linear SVM:

from sklearn.svm import SVR

SVR(kernel='linear', C=<Penalty parameter C of the error term>)--SVR model

fit(X, y)--Fit the model according to the given training data.

predict(X)---Predict using the linear model

Logistic regression:

from sklearn import linear_model

linear_model.LogisticRegression(C=<Penalty parameter C of the error term>)---logistic regression model

fit(X, y)--Fit the model according to the given training data.

predict(X)---Predict using the linear model

Kernel Ridge regression:

from sklearn.kernel_ridge import KernelRidge

KernelRidge(alpha=<penalty>)

fit(X, y)--Fit the model according to the given training data.

predict(X)---Predict using the linear model

CONCLUSION

The accuracy of our prediction on the validation data using PCA is 65.83%.

The accuracy of our prediction on the validation data without using PCA is 63.42%.

References

Yelp Dataset link, Yelp Dataset Challenge,
www.yelp. com/dataset_challenge

Ashish Gandhe Restaurant Recommendation System,
http://cs229.stanford.edu/proj2014/Ashish%20Gandhe,Restaurant%20Recommendation%20System.pdf

Yelp Food Recommendation System, Sumedh Sawant and Gina Pai,
http://cs229.stanford.edu/proj2013/SawantPai-YelpFoodRecommendationSystem.pdf

A Preference-Based Restaurant Recommendation System for Individuals and Groups,
http://www.cs.cornell.edu/~rahmtin/files/YelpClassProject.pdf

Kernel ridge regression python library

http://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html

Linear support vector regression python library

http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html

Logistic regression python library

http://scikit-learn.org/stable/auto_examples/linear_model/plot_iris_logistic.html