Investigating Fundamental Machine Learning Models

Kenny Oseleononmen
Stanford University
kenny1g@stanford.edu

INTRODUCTION

This paper uses the problem of predicting whether a user will like a restaurant to explore and understand various fundamental machine learning algorithms, namely Logistic Regression, Naive Bayes and Decision Trees.

DATASET AND FEATURES

The origins of the dataset being used is the Yelp academic dataset. From the Yelp dataset, the subset of reviews for businesses that fell under the restaurant category where sampled. Following that, the dataset was further reduced to only include reviews made by the user with the most reviews within this subset. Then, the rows with missing data were removed and finally some feature selection was done and columns that weren't labeled 'attributes' where removed
After this processing we are left with the following features:

1. RestaurantsTakeOut
2. RestaurantsReservations
3. BusinessAcceptsCreditCards
4. GoodForKids
5. Caters
6. HasTV
7. BikeParking
8. RestaurantsGoodForGroups
9. RestaurantsDelivery
10. OutdoorSeating

A Principal Component Analysis (PCA) was then done on the dataset. Results of the PCA in 2 and 3 dimensions are shown below with the different colors representing the labels of the examples

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

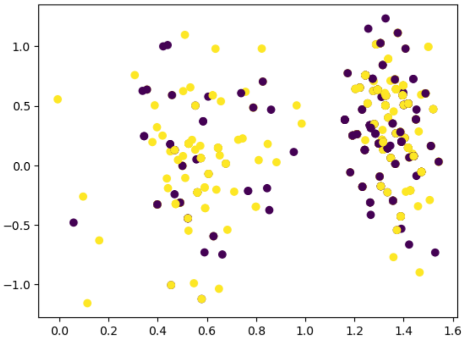

Figure 1. Dimension reduction with 2 Principal Components

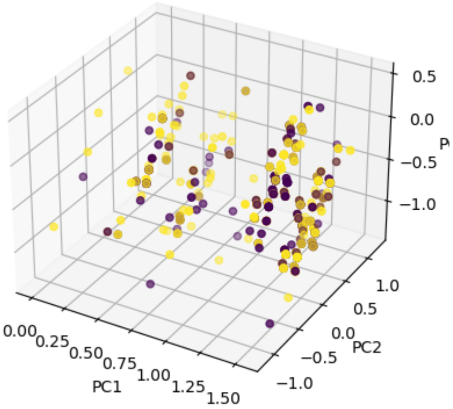

Figure 2. Dimension reduction with 3 Principal Components

The final dataset had 686 examples and this was split into a training and testing set of size 548 and 138 respectively

METHODS

Evaluation

We evaluate the models by calculating accuracy precision and recall.
Given the True Positives, False Positives, True Negatives and

False Negatives of the predictions we have the following

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

Logistic Regression

The first model used is a Logistic Regression fitted using Newton's method.

Naive Bayes

Decision Trees

RESULTS AND DISCUSSION