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

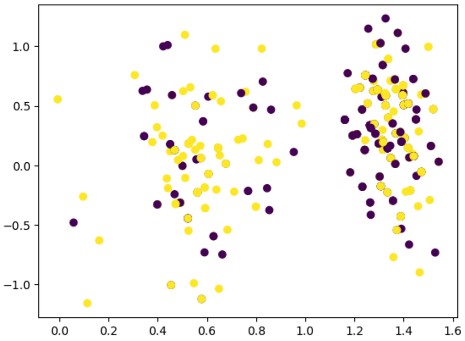

Figure 1. Dimension reduction with 2 Principal Components

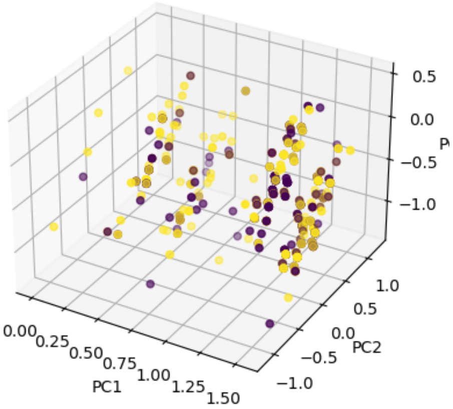

Figure 2. Dimension reduction with 3 Principal Components

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 one of Logistic Regression fitted using Fisher Scoring. Fisher Scoring is the application of newton's method for finding the zero of a function to maximize the logistic regression log likelihood function. It follows that since the maxima of the log likelihood function ℓ corresponds to points where it's first derivative is zero, by finding the zero of the first derivative of the log likelihood $\ell'(\theta)$, we maximize the log likelihood. This gives the update rule

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

Note that the Hessian Matrix H of a function is a matrix made up of the second-order partial derivatives of the function with respect to it's inputs so we have $H = \ell''(\theta)$, also note that in logistic regression, θ is vector valued. With these we get the update rule

$$\theta := \theta - H^{-1}\nabla_{\theta}\ell'(\theta)$$

This is the update rule with which the model was fitted.

Naive Bayes

The second model used is a Naive Bayes model.
Naive Bayes is a Generative Learning Algorithm that derives the posterior distribution on y given x $P(y|x)$ using Bayes Rule. i.e

$$P(y|x) = \frac{P(x|y)\cdot P(y)}{P(x)}$$

Naive Bayes also makes the wrong but useful assumption that all features are conditionally independent given their labels. i.e $P(x_j) = P(x_j|y, x_{j+i})$. Furthermore $P(x)$ is ignored as it is a constant factor for all classes and does not affect the decision boundary.

Following these assumptions, we can expand Bayes Rule further to become:

$$P(y=1|x) = \prod_{j=1}^{d} p(x_j|y=1)\cdot p(y=1) \tag{4}$$

Using the above, the probabilities $P(y=1|x)$ and $P(y=0|x)$ are calculated and the class with the highest probability is chosen as our prediction.

To fit the model, we store the total counts where $x_j = 0 \wedge y = 1$, $x_j = 1 \wedge y = 1$, $x_j = 0 \wedge y = 0$ and $x_j = 1 \wedge y = 0$

Finally, to predict we use these stored counts to calculate the posterior distribution using the equation in 4

Custom Generalized Linear Model

Our label y, thus far has been whether our user with most ratings, Karen, likes a restaurant or not. For the sake of our custom GLM, we will instead predict the rating (1 - 5) the user would have given to the restaurant.

Generalized Linear Models are good at solving real life problems that can be modeled by probability distributions in the exponential family.

The response varibale is the rating the user would have given to the restaurant, because this response variable is continuous it can be modeled as a Gaussian Distribution.

The probability density function of the Gaussian Distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right)$$

Furthermore when constructing a GLM, we make the following assumptions

1. $P(y|X;\theta)$ is modeled after the exponential family
2. The models hypothesis function will be predicting $E[y|x]$
3. the natural parameter η and features X, are related linearly

By modelling the users rating as a Gaussian Distribution, we have satisfied the first assumption.

From the second assumption, we see that

$$h(X) = E[y|X] = \mu$$

From the third assumption, we know that the canonical parameter η is linearly related to X

To identify the canonical parameter η, we write the probability density function of the Gaussian Distribution in exponential family form

$$p(y;\eta,\tau) = b(\alpha,\tau)\exp\left(\frac{\eta^T T(y) - a(\eta)}{c(\tau)}\right)$$
$$p(y;\eta,\tau) = b(\alpha,\tau)\exp\left(\frac{\eta^T T(y) - a(\eta)}{\sigma^2}\right)$$
$$p(y;\eta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$
$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\exp\left((x\eta - \frac{1}{2}\eta^2)T(y)\right)$$

The random component identifies the response variable and a probability distribution for it. We know from assumption 1 above that $P(y|X;\theta)$ is modeled after the exponenetial family. this gives us

$$p(y;\eta) \sim b(y)\exp(\eta^T T(y) - a(\eta))$$

where $T(y)$ is the sufficient statistic for the distribution, $a(\eta)$ is the log partition function and $b(y)$ is the normalization constant. The random component of our model is thus $p(y;\eta) \sim \mathcal{N}(\eta, \sigma^2)$, i.e modeled following the Gaussian Distribution

RESULTS AND DISCUSSION