Machine Learning

Assignment One

Abstract

In this project we investigated the performance of two different machine learning classification models, logistic regression and Naive Bayes. We selected four datasets to test the models on, the Ionosphere, Adult, Poker Hands and Credit Approval datasets. Both the Adult and Credit datasets were classified well by both models, obtaining over 74% accuracy and logistic regression outperforming Naive Bayes. The Poker Hands dataset was of interest to us and was selected as we knew it could be easily classified with a brute-force algorithm. This dataset is known to be difficult for classification algorithms and we saw both Naive Bayes and logistic regression perform poorly with less than 50% accuracy. The Ionosphere dataset experienced extremely high accuracy in classification under logistic regression, 94%, however Naive Bayes achieved only 53%, potentially a violation of the Naive Bayes assumption that all features are conditionally independent. We saw a strong correlation between rate of convergence and learning rate in logistic regression. In both models we saw that training on smaller datasets yielded similar results up until the training set size became very small. Removing features saw greater accuracy in datasets with primarily continuous features, but this trend was not observed in mixed or categorical datasets.

Introduction

The purpose of this project was to train two machine learning models using logistic regression and Naive Bayes. Each of these models would be trained on four separate datasets and then tested to have their accuracy assessed. Our datasets included purely continuous data, purely categorical data and two mixed types. Three were binary classification tasks while one required classification into 10 classes. Most importantly we determined that both logistic regression and Naive Bayes performed with similar accuracy on the mixed type datasets and pure classification dataset. We saw logistic regression significantly outperform Naive Bayes on the continuous dataset, however this may have been due to the nature of the dataset. We further determined that logistic regression's learning rate did not impact the accuracy but greatly affected the convergence speed. In both models a smaller training dataset did not affect performance ($p > 0.05$) and resulted in faster training, up until training set size of around 100. Removing a random subset of features was seen to achieve similar performance accuracy on the ionosphere and credit rating data sets, with about half of the features considered.

Datasets

The last two datasets we chose were the poker hands dataset and the credit approval dataset. The latter was chosen because it combines continuous and categorical features, while having binary categorization. The former was chosen because it is a unique type of dataset: one would expect that there are no correlations between any of the features, and so logistic regression would perform very poorly on it. Further, one could easily write an algorithm which could achieve 100% on the poker hands

Group Members: Haoyan Yu, Caitlin Hutnyk, Hanna Navissi

dataset by using the concrete rules of poker hand rating. This dataset has been described as challenging for classification algorithms to achieve high accuracy.

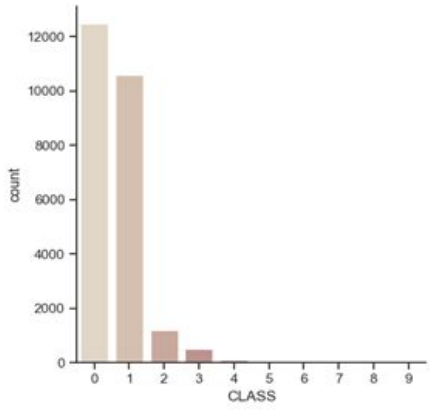
Dataset processing

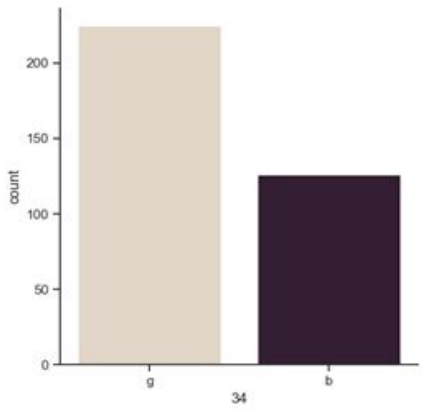
The datasets are processed by first removing instances containing malformed data: like extremely large or missing values, then looking at mean, standard deviation, min, and max of each feature and filter out extremely high and implausible values such as '99999' for age. In order to ease the training step, we have normalized the distribution of numerical features to avoid data overflow issues and to make sure all features were weighted the same. An initial column of ones was added to the datasets for logistic regression in order to allow for the intercept value to be included in the weights array. The training set was separated into continuous and categorical features to be passed to the Naive Bayes model in order to calculate their respective likelihoods. All categorical features were encoded using one hot encoding.

Descriptive Statistics

Class distribution in dataset 1 and 4 are comparatively balanced, however datasets 2 and 3 are not. Especially in dataset 3, over 90% of data are in class 0 and 1, which means we'll achieve higher accuracy in class 0 and 1 and lower in other classes. Similarly, in dataset 2 higher accuracy is expected to occur for classification of <=50K. One possible solution is to undersample majority classes or oversample minority classes.

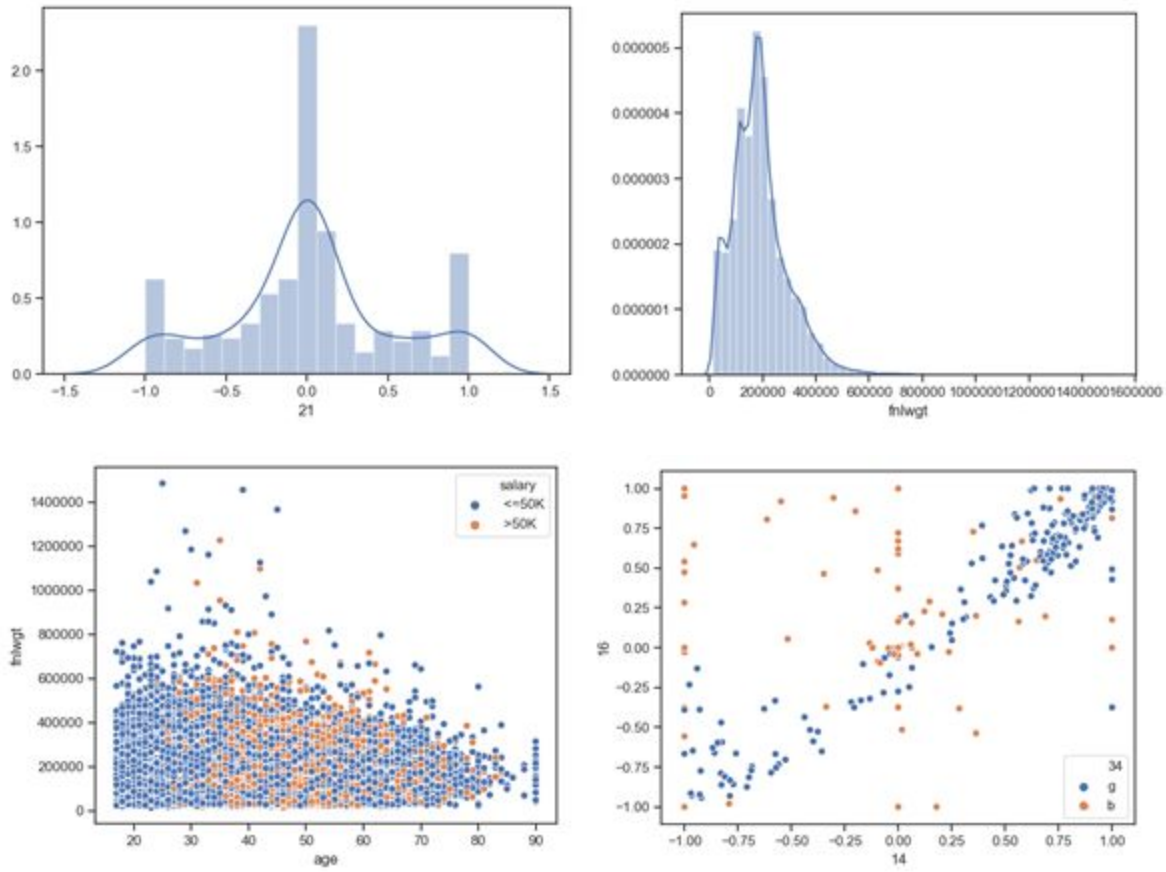
Distribution of numerical features are mostly following skewed or non-skewed Gaussian distributions, except for dataset 3 (Poker Hands), as expected, because the probability of getting each suit and rank is equal when poker sample is unbiased. In dataset 1, we see most features have a Gaussian distribution, however there are still bunches of data that fall into -1 and 1 which may make sense as the characteristic of feature is designed so.





Collinearity (linear correlation) is seen in some of the pairwise features from scatter plots, such as age vs fnlwgt in dataset 2, and feature 14 vs 16 in dataset 1. Which will potentially over or

underestimate the features weight but at this moment we are not dealing with such collinearity issues.



Results

Results Table

The following table displays the performance rates of the two models on the four datasets. The comparative results were very similar for the other factors tested:

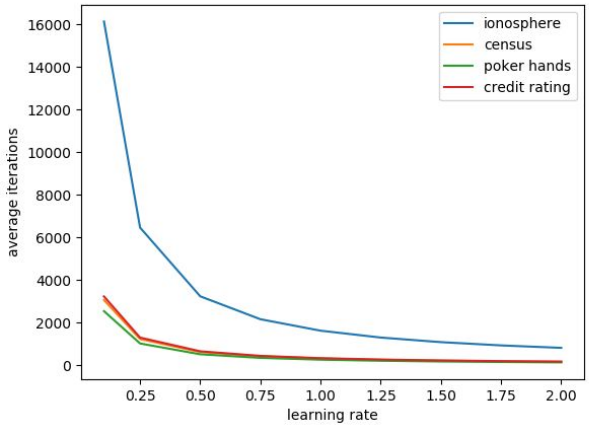
	Ionosphere	Adult	Poker hands	Credit score
Logistic Regression	94%	82%	48%	89%
Naive Bayes	53%	75%	45%	74%

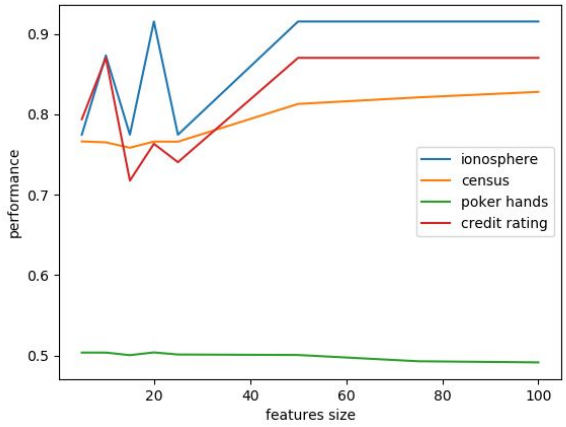
Logistic Regression Results

We tried out three ways to calculate the accuracy of a model: by simply training on the whole training set, then testing on the testing set, by taking the average of the performances on the

validation sets, and by taking the best performing model (of the k made during k-fold cross validation) and taking its accuracy on the testing set. We saw that in general all three methods reported very similar rates.

Further tests were performed altering the learning rate. Convergence was observed to occur more rapidly with a larger learning rate across all datasets. The accuracy was not observed to change with learning rate.





To test the features size vs. accuracy, we randomly dropped features from less to more. With very few features the results were more variable, although in some cases small feature sets achieved results comparable to those achieved with the full set of features.

We also tested using a smaller training set size: less test cases. This yielded poor results for very small training sets (less than half the original size), but other than that the results matched those with all the training examples.

Naive Bayes Results

A model trained using Naive Bayes had an accuracy similar to logistic regression, however significantly underperformed on the Ionosphere dataset. This point is further discussed below.

The Naive Bayes model requires no hyperparameters, instead relying on prior probabilities and likelihoods of data to generate results. As a result, there was little to be done during the validation stage during the k-fold train and validate procedure.

Discussion

Ionosphere Dataset

The Ionosphere dataset produced some unexpected results. While we expected that on a smaller dataset Naive Bayes would perform better than logistic regression, we found this to be untrue. A potential cause could be that the Ionosphere dataset is comprised of pairs of related features, as discussed in the Dataset section. A core assumption of Naive Bayes is that features are unrelated,

which we know to be untrue for this dataset. A second reason for this could be an error in the Gaussian likelihood calculation algorithm, skewing the results of the dataset that is comprised solely of continuous data. The fault could have been masked in datasets two and four, which are mixed, and not detectable at all in dataset three which has purely categorical data.

Learning Rate

The logistic regression model uses a learning rate to inform the next step size in gradient descent. We found that altering the learning rate did not influence the accuracy of the model, however it affected the rate of convergence. Our results affirm our expectation of a changing learning rate. By performing gradient descent on the convex cost function, we can always reach a point close enough to the function minimum that the steps yield improvement smaller than our stopping condition, except when our learning rate is so large that it overshoots. By increasing the learning rate, we make larger changes to the gradient in the direction of the minimum, thereby converging at a faster rate.

Poker Hand Dataset

The Poker Hand dataset is known to be challenging for classifiers. The multitude of different card combinations that comprise the same hand and subsequently same class, make the job of drawing relationships between the two difficult. Both Naive Bayes and Logistic Regression models failed to achieve high accuracy in classification. A naive algorithm can easily be implemented using the rules of poker hand classification which would achieve 100% accuracy, so we wanted to see if machine learning approaches could be nearly as successful. Similarly to the Ionosphere dataset, using Naive Bayes on this dataset encounters the same problem in that features are related but are interpreted as being conditionally independent. That being the case, Naive Bayes still performed relatively well, considering logistic regression only achieved 3% greater accuracy.

Conclusion

The project successfully generated models that were able to classify unlabeled data after a training period. Both models performed well and similarly across all datasets except for the Ionosphere set. Further direction for investigation involves using Naive Bayes on variations of the Ionosphere set that deal with the relationship between the pairs of features, perhaps with preprocessing or learning all the unrelated features together such that the dataset is split in two. The discussion in the datasets section outlines a number of factors in the spread of data that could have resulted in reduced accuracy. Further statistical assessment could be used to identify these areas and reduce their impact.

Contributions

Haoyan Yu: data analysis, report (datasets, results), code to import and process data for use by models

Caitlin Hutnyk: logistic regression, debugging, logistic regression test design

Hanna Navissi: Naive Bayes, debugging, report (discussion, conclusion, editing)