Project Report 1

Dingyi Zhuang, Tianyu Shi, Fuyuan Lyu

February 11, 2020

In this project, we implement two basic machine learning models: logistic regression and Naive Bayes. These models are tested upon four classification datasets: Ionosphere Dataset, Iris Dataset, Car Evaluation Dataset, and Adult Dataset. In the pre-processing phase, we first discretize all the continuous features, handling the missing values and values out of range. In the experiment phase, the Naive Bayes model performs well on larger datasets like Adult Dataset and Car Evaluation Dataset while the logistic regression model performs on smaller datasets, including Iris Dataset and Ionosphere. we further investigate the influence of different learning rates, different stopping criteria, hyper-parameter tuning, feature selection, and dataset selection.

1 Introduction

Classification tasks are one of the most common and important tasks in the machine learning community. Given certain features, the classification task is to categorize which class is most likely to be. In this project, we implement two basic machine learning models: logistic regression and Naive Bayes, and test their performance upon four datasets: Ionosphere Dataset[1], Adult Dataset[2], Iris Dataset[3] and Car Evaluation Dataset[4]. The logistic regression model performs better on Iris Dataset and Ionosphere Dataset, while the Naive Bayes model performs better on Adult Dataset and Car Evaluation Dataset. All the features are discretized during the pre-processing phase. We further investigate the influence of different training techniques, such as hyper-parameter tuning and feature selections, upon the performance of the models.

2 Datasets

The targets of all these four datasets are categorical classification (including binary classification). We exam four datasets to find that no missing values exist. We use *replace* function in *pandas* module to process all the target values into categorical count value. We will briefly describe the specific features and then introduce how to extract/process the features.

Ionosphere Dataset

Ionosphere dataset contains 34 radar data (real values between -1 and 1) with 351 rows[1]. We find their distribution with respect to the "good" or "bad" target, which is reflected in Figure 1. We then remove radar 1 feature which is always 0. We can see that good class and bad class have quite distinct patterns in the antennas power value, which is essential in the following classification. We also fill between the intervals of good-class feature distribution to find that such surface is quite similar to the auto-correlated signals. Therefore, we use Principal Component Analysis (PCA) to reduce the dimensions from 34 into 10 to learn the low-rank representation for more efficient and powerful training.

[1] `https://archive.ics.uci.edu/ml/datasets/ionosphere`
[2] `https://archive.ics.uci.edu/ml/datasets/Adult`
[3] `https://archive.ics.uci.edu/ml/datasets/Iris`
[4] `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`

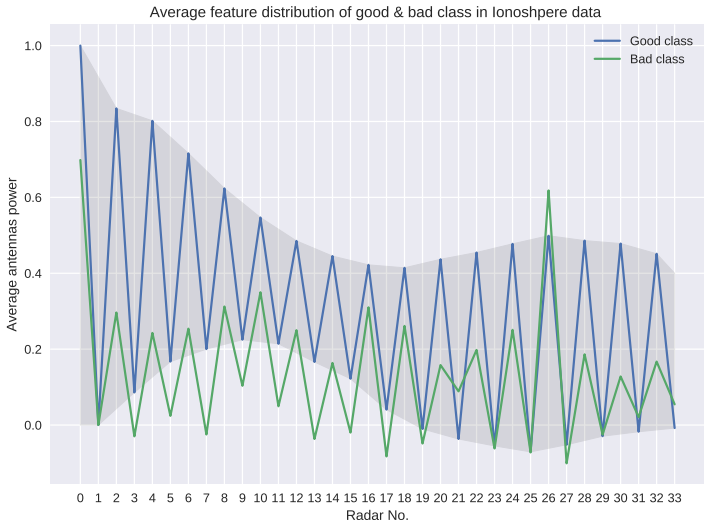

Figure 1: Distribution of average numerical features given "good" or "bad" class in ionosphere dataset

Adult Dataset

Adult dataset aims to predict whether income exceeds \$50K/yr based on census data[2]. There are 32561 records and 14 features included with datatypes of continuous count values, continuous real values and categorical/binary values. We min-max normalize continuous values (both count and real), discretize normalized real values into 10 categories and leave everything else untouched.

Iris Dataset

There are 4 continuous real value attributes, whose basic statistic information are listed in Figure 2 and Table 1, where *sl,sw,pw,pw* stand for *sepal length, sepal width, petal width, petal length* [3]. Linear correlation between features can be found, e.g. petal length & petal width and petal length & sepal length. We also min-max, normalize and then discretize the normed real values.

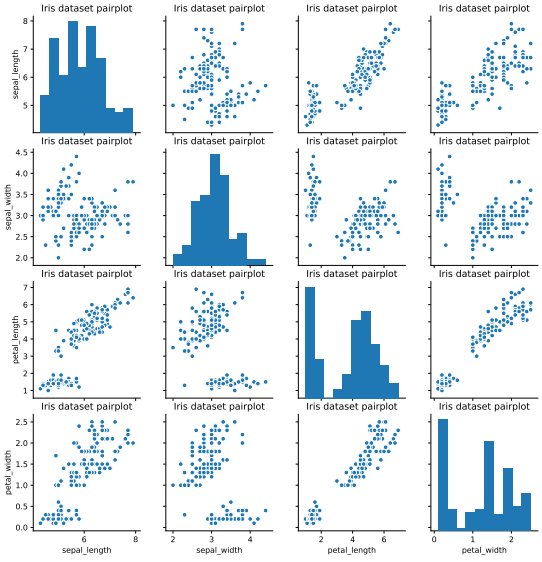

Figure 2: Pairplot of iris data features

	sl	sw	pl	pw
count	150	150	150	150
mean	5.84	3.05	3.75	1.19
std	0.82	0.43	1.76	0.76
min	4.3	2.0	1.0	0.1
25%	5.1	2.8	1.6	0.3
50%	5.8	3.0	4.3	1.3
75%	6.4	3.3	5.1	1.8
max	7.9	4.4	6.9	2.5

Table 1: Iris data description

Car Evaluation Dataset

In Car Evaluation Dataset, we use 1728 records from 6 categorical features of cars, e.g. price, door number and capacity, to predict the safety level of the cars [4]. This dataset is straightforward, we only transform the raw data into categorical features to predict the categorical safety level.

3 Results

Accuracy comparison on four datasets

From Table 2 , for accuracy of each model, we find that our implemented logistic regression model can work well on small datasets (iris and ionosphere datasets) but worse in larger dataset (car and adult datasets) compared with *sklearn* model. In the contrary, naive bayes work well on large dataset but worse in small dataset.

	iris	car	adult	ionosphere
logistic regression(ours)	**0.75**	0.65	0.66	**0.80**
logistic regression(sklearn)	0.93	0.87	0.82	0.82
naive bayes(ours)	0.34	**0.69**	**0.76**	0.67
naive bayes(sklearn)	0.74	0.69	0.75	0.81

Table 2: Accuracy of our implemented model in four datasets comparing to sklearn module

Furthermore, we compare the different evaluation metrics for binary classification, where ionosphere dataset and adult dataset are selected. Details can be found in Table 3, where **o** stands for our model and **s** for results after running *sklearn* module.We found that our model have lower precision and recall on class 1 for large dataset (adult data), which can explain why the model fail to show the performance like *sklearn* model. We plan to investigate more in the characteristic of the features in large dataset.

	ionosphere/adult dataset						
	precision(o)	precision(s)	recall(o)	recall(s)	f1(o)	f1(s)	support
class 0	1.00/0.93	0.90/0.83	0.48/0.49	0.48/0.95	0.65/0.64	0.86/0.89	33/4937
class 1	0.69/0.35	0.85/0.71	1.00/0.88	1.00/0.39	0.81/0.51	0.88/0.51	37/1575
macro avg	0.84/0.64	0.88/0.77	0.74/0.68	0.74/0.67	0.73/0.57	0.87/0.70	70/6512
weighted avg	0.83/0.79	0.87/0.80	0.76/0.58	0.76/0.81	0.74/0.61	0.87/0.79	70/6512

Table 3: Different evaluation metrics of our implemented model in binary-target datasets comparing to sklearn module

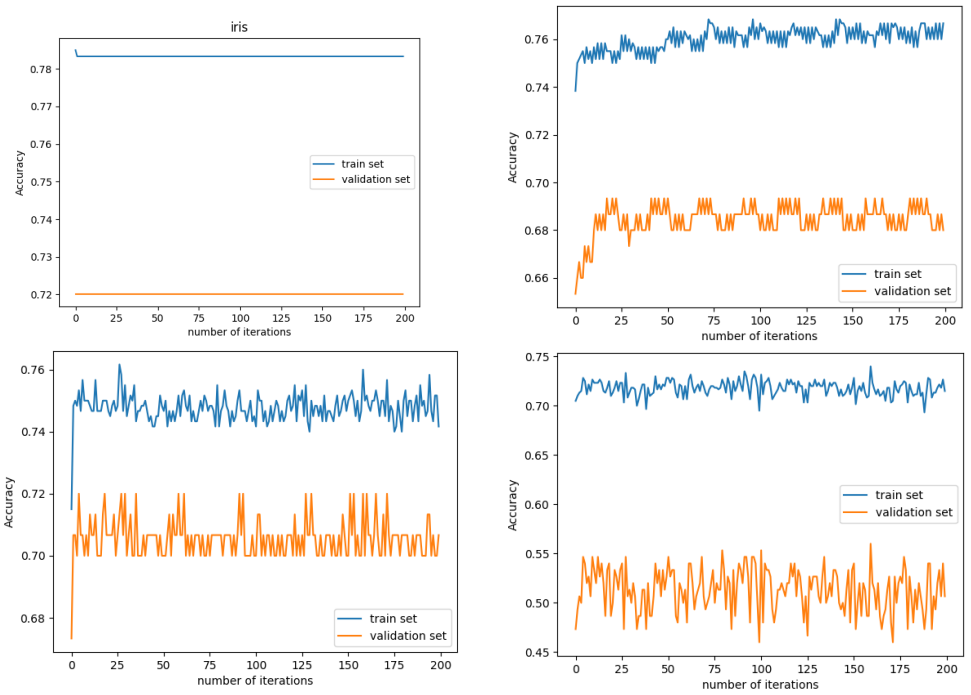

Figure 3: Learning rate test, from left top to right top, from left bottom to right bottom, the learning rates are: 0.001, 0.01, 0.1 and 1

Learning rate test

We mainly test learning rates with stopping threshold 0.05 in four datasets, a specific case on iris dataset is shown in Figure 4. For small learning rate, like 0.0001, the accuracy makes no change during the gradient decent, which means that it converges very slow. When the learning rate is very big such as 1 , the accuracy will fluctuate sharply, which means that it will become unstable during gradient descent.

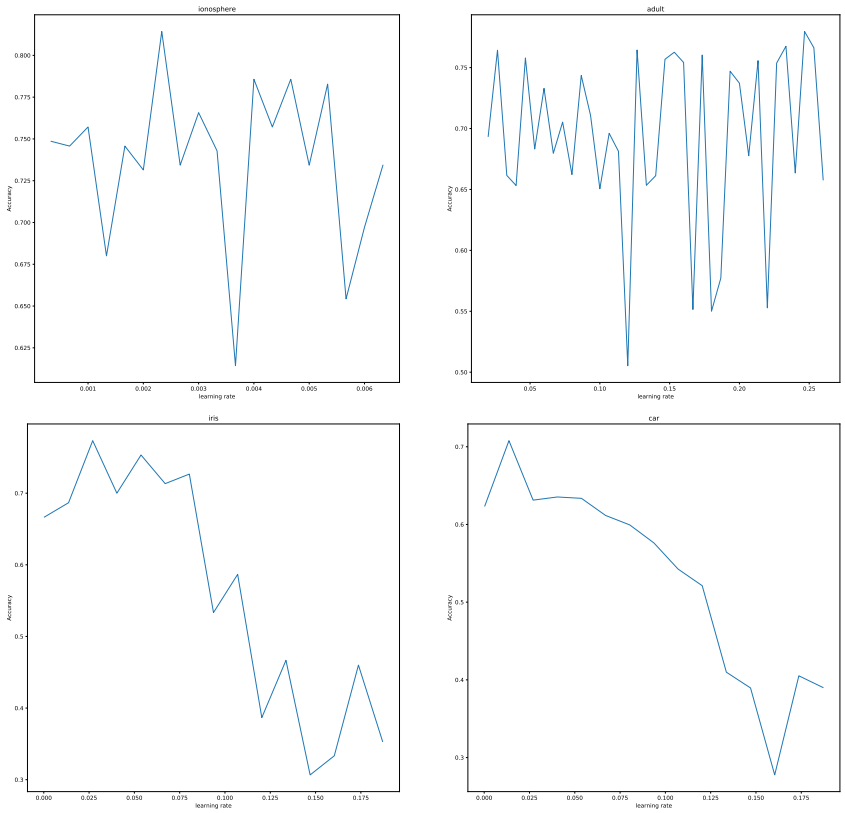

Figure 4: Hyperparameter search for learning rates in different datasets, from left bottom to right bottom, the learning rates are: ionosphere, adult, iris and car datasets

We also find similar effect on four data set, with different learning rate ranges. For example, in the iris dataset, the best learning rate in our experiment should be around 0.01. But in adult data set it should be 0.1. Then, we perform hyper-parameter search in a smaller range of learning rates, which is shown in Figure 5. According to our experiment, we find the best learning rate for the iris dataset should be 0.025, the ionosphere dataset should be 0.0025, the car evaluation dataset should be 0.015 and the adult dataset should be 0.24.

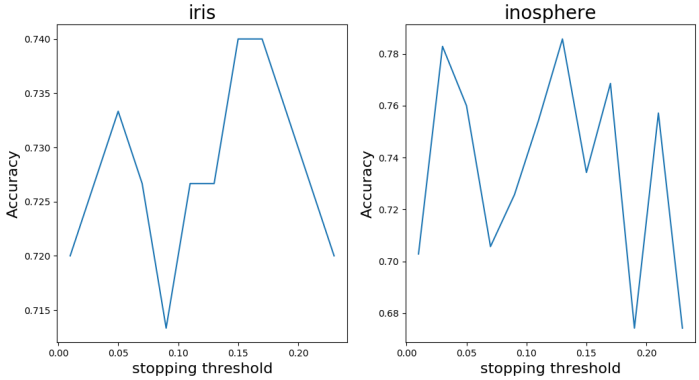

Figure 5: Best stopping criterion for Ionosphere Dataset and Iris Dataset given optimal learning rate

Stopping criteria for logistic regression

We also investigate the stopping criteria on four datasets. The stopping criteria are considered as a threshold for change in the value of cost function. From Figure 5, we find there is not much influence on accuracy for car and adult datasets because these two datasets are very big, our model may need to run max iteration to meet the optimal solution. We find that for iris and ionosphere datasets, there will be optimal criteria, which are 0.15 and 0.14 respectively.

Accuracy as a function of the size of dataset

From Figure 6 we can find that accuracy increases when sample size is roughly less than 200. However, the accuracy becomes stable even the sample size continues growing.

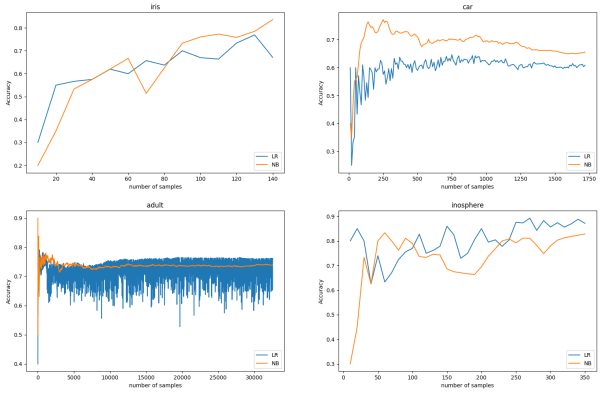

Figure 6: Learning rate test, from left top to right top, from left bottom to right bottom, the learning rates are: 0.001, 0.01, 0.1 and 1

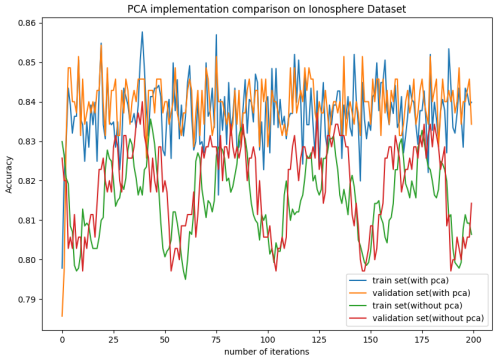

Figure 7: Training and validation accuracy comparison before and after using PCA

Dimension reduction comparison

Because the ionosphere dataset has strong auto-correlation, we use PCA to reduce the dimensions for better training results, which can be reflected in Figure 7. It can be found that accuracy on both training and validation sets perform better if PCA is applied. By using PCA for dimension reduction, we can learn a better representation of features.

4 Discussion and Conclusion

In this report, we implement logistic regression and Naive Bayes models. They are tested on four categorical classification datasets. In the data cleaning part, we check missing values, discretize the real values and reduce dimensions for auto-correlated features. We test the accuracy and impact of learning rate, stopping criteria and dimension reduction, where we all choose an improved strategy. As future work, we still need to look into why our model would perform less accuracy and efficient than *sklearn* module.

5 Statement of Contributions

- Dingyi Zhunag: Data preprocessing, data analysis and report formating.
- Fuyuan Lyu: Model implementation and report writing.
- Tianyu Shi: Running experiment, data visualization and data analysis.

References

[1] Vincent G Sigillito, Simon P Wing, Larrie V Hutton, and Kile B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.

[2] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[3] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[4] Marko Bohanec and Vladislav Rajkovic. Knowledge acquisition and explanation for multi-attribute decision making. In *8th Intl Workshop on Expert Systems and their Applications*, pages 59–78, 1988.