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. These datasets contains various type of features: both continuous and discrete. 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, apart from reporting evaluation metrics upon various models with different datasets, we further investigate the influence of hyper-parameter tuning, feature selection and dataset selection.

1 Introduction

Classification tasks is one of the most common and important task in machine learning community. Given certain features, 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 upone four dataset: Ionosphere Dataset[1], Adult Dataset[2], Iris Dataset[3] and Car Evaluation Dataset[4]. 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

We use four datasets including Ionosphere Dataset[5], Adult Dataset[6], Iris Dataset[7] and Car Evaluation Dataset[8]. 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.

[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`
[5] `https://archive.ics.uci.edu/ml/datasets/ionosphere`
[6] `https://archive.ics.uci.edu/ml/datasets/Adult`
[7] `https://archive.ics.uci.edu/ml/datasets/Iris`
[8] `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`

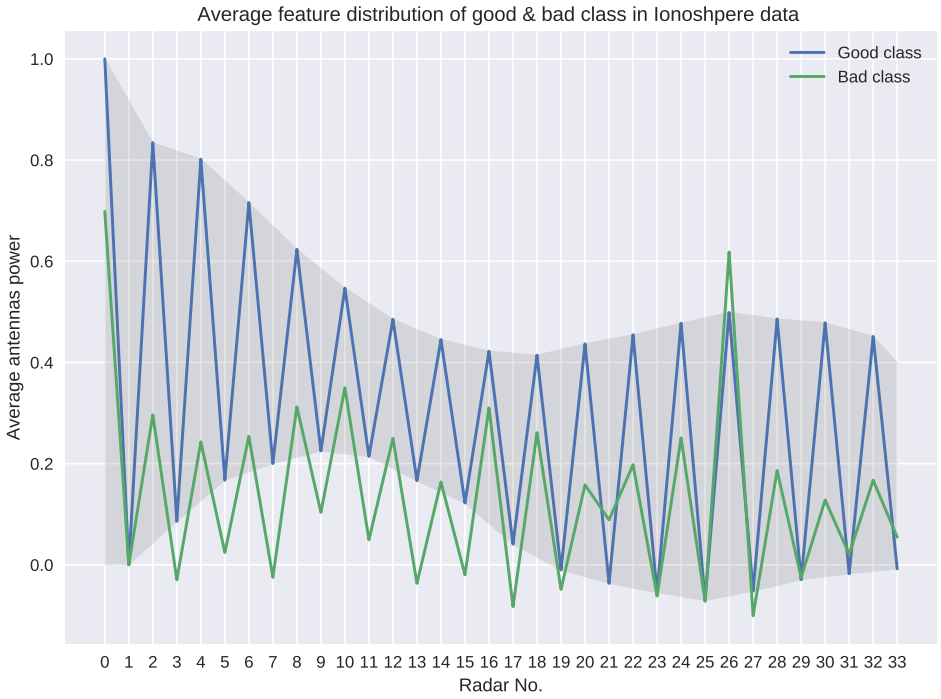

Figure 1: Distribution of average numerical features given "good" or "bad" class in ionosphere dataset

Ionosphere Dataset

Ionosphere dataset contains 34 radar data (real values between -1 and 1)[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 to reduce the dimensions from 34 into 10 to learn the low-rank representation for more efficient and powerful training.

Adult Dataset

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

Iris Dataset

There are continuous real value attributes with 1 decimal precision, whose basic statistic information is 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 discretize the normed real values.

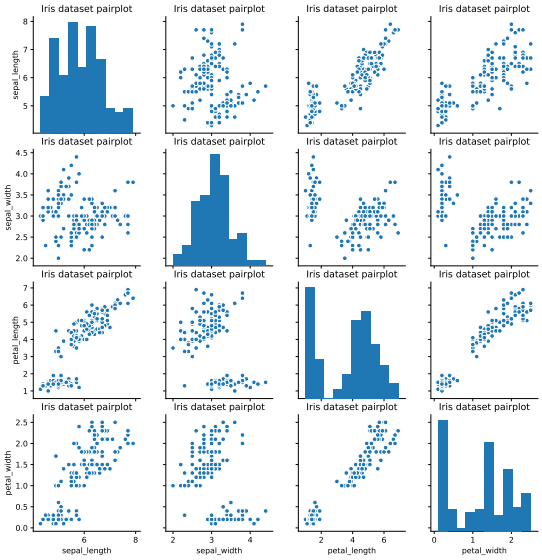

Figure 2: Pairplot of iris data features

	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 some categorical features of cars, e.g. price, door number and capacity, to predict the safety level of the cars [4]. This dataset is quite straightforward, we only transform the raw data into categorical features to predict the categorical safety level.

3 Results

4 Discussion and Conclusion

5 Statement of Contributions

- Dingyi Zhunag
- Tianyu Shi
- Fuyuan Lyu

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.