Report

for

COMP 3314 Assignment 1

Prepared by Shunhua Guo, 3035447635

Logistic Regression & Random Forest

Discuss their implementations, obtained results, advantages and limitations of the implemented methods in their reports.

March 26, 2021

# Implementations

The algorithms are implemented in Jupyter Notebook.

Required Libraries are specified in each notebook.

## Logistic Regression

文本

中度可信度描述已自动生成As the sigmoid function applies to binary classification cases while the Softmax function applies to multi-class classification problems, the Softmax function is implemented below.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成The structure of the implementation is as follows. The main function is ‘fit’ for training the regression using data input, which will run gradient descend for ‘n\_iter’ times.

文本, 信件

描述已自动生成The loss function is negative-log-probability, and gradient is the derivative of it.

## Decision Tree

As stated in the lecture, the following impurity measures or splitting criteria are commonly used in decision trees:

1. Gini impurity (gini),

2. Entropy (entropy),

3. Classification error (classificationError).

图形用户界面, 文本, 应用程序

描述已自动生成And to record Decision Tree training result, I initialize a tree node structure.

For the Decision Tree model, I implement splitting group in a recursively manner.

文本

描述已自动生成The structure of the model class is as follows.

## Random Forest

文本, 应用程序

描述已自动生成The basic logic of random forest is implementing multiple decision trees, with sample bagging strategy.

# Obtained Results

## Logistic Regression on Iris Data

文本

中度可信度描述已自动生成The following is an example training Logistic Regression on Iris Dataset.

图表, 折线图

描述已自动生成图表

描述已自动生成And the below graphs are Loss and Accuracy plot for training and validation set.

## Logistic Regression on Car Data

文本

描述已自动生成 The following is an example training Logistic Regression on Car Dataset.

## Random Forest on Iris Data

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成The following is an example training Decision Tree and Random Forest on Iris Dataset.

## Random Forest on Car Data

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成The following is an example training Decision Tree and Random Forest on Car Dataset.

# Advantages and Limitations of the implemented methods

**3.1 Logistic Regression**

Good at handling data with linear separatable or almost characteristics.

Not able to handle higher-order data, like quadratic relationships.

**3.2 Decision Tree**

Easily overfit the training set, though we can apply pruning or early termination to reduce overfitting problem.

High variance of the model.

Interpretable.

**3.3 Random Forest**

Can solve overfitting problem of Decision tree by average across multiple decision trees, and reduce the variance of the model.

# Parameter analysis

**4.1 Logistic Regression**

图形用户界面

描述已自动生成图表

描述已自动生成图形用户界面

描述已自动生成A simple plotting of different **regularization term** result vs. number of iterations.

From the left to the right, they are No regularization, L1 regularization and L2 regularization cases.

**3.2 Decision Tree & Random Forest**

图表

描述已自动生成It is a heatmap plotting of the accuracy from each impurity or error function after training Car dataset.

图表, 折线图

描述已自动生成This is a graph studying Accuracy vs. max\_depth of Decision Tree.

图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成In the random forest model, I also studied the impact on accuracy of Number of tree estimators in Random Forest (k), Number of sample size in each sample bagging (n) and Number of featurses selected in each sample bagging (d).

# References

**4.1 Datasets**

**[1] Iris dataset**

**(url:** [**https://archive.ics.uci.edu/ml/datasets/Iris**](https://archive.ics.uci.edu/ml/datasets/Iris)**)**

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The 3 plants are Versicolour (class 0), Virginica (class 1) and Setosa (class 2) respectively.

There are 4 input attributes: sepal length, sepal width, petal length and petal width.

There are 100 training samples and 50 testing samples.

**[2] Car evaluation dataset**

**(url: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation)**

This model evaluates cars according to their status and classifies them as unacceptable (class 0), acceptable (class 1), good (class 2) and very good (class 3).

The 6 input attributes are buying price, price of the maintenance, number of doors, number of persons to carry, size of luggage boot, safety of the car.

There are 1209 training samples and 519 testing samples.