CMPUT 466 min-project

Marco Mai

Introduction

In real life, every country has a census every few years. This project aims to find out what the most important personality or background takes place in terms of salary.

Problem formulation

To find the answer, a dataset was downloaded from UCI Machine Learning Repository, based on the 1994 Census database. The data contain categorical and numerical data. There are 14 features in the dataset, which are:

```
Age, workclass, fnlwgt, education education-num, marital-status, occupation, relationship,
race, sex, capital-gain, capital-loss, hours-per-week and native-country
```

* For details of each contains please visit */data/old.adult.names , line 64 -77

The output is classified into two classes, people who gain more or less 50k income per year.
There are 48842 instances. 32561 training sets and 16281 testing sets.

To formulate the task in the trainable input data, the One Hot method to convert a new input. There will be 6 algorithms being comparing result in the algorithm, which is logistic regression, Decision Tree, Naïve Bayes Classifier, Random Forest and Weighted Quadratic Discriminant Analysis

Approaches and baselines

The dataset contains categorical data. However, those data can not be directly recognized by the machine. Here thirty-party library Pandas is used to read the file and identify the numerical and categorical columns, data will be spitted into X and target, and using OneHotEncoder to expand those data into 108 features.

While running the different algorithms, if an algorithm has parameter(s) to tune, there will be 50 chances to run, while selecting the best accuracy. (e.g., Changing Max_depth, N_epoch) In those tunning algorithms, the dataset will split a ratio of 9:1 to train and validate the dataset.
In linear and logistic regression mini-batch gradient descent is used to get better accuracy.

After all, the algorithms are completed it will generate a ranking base on the accuracy, and each top 10 most important features will be printed.

Evaluation metric

The measure of success is within the top 10 absolute weight should have more than 6 also appear in the other 3 algorithms.
The goal of the task is to find an algorithm with a relatively short run time while getting relatively good accuracy, with multiple runs used to compare the best algorithm, the result should be stable, so the result will not be an approximation.

Result

The following result in the next is converted from ProgrammeOutput-report.txt.

Translating Report.txt to readable result (underline text indicated text swap for specific One Hot feature Name)

	1	2	3	4	5	6	7	8	9	10
Random Forest	capital-gain	Married-civ-spouse	education- NaN	age	Husband	hours-per-week	fnlwgt	Never-married	capital-loss	Exec-managerial
Decision Tree	Married-civ-spouse	education- NaN	capital-gain	capital-loss	age	hours-per-week	fnlwgt	Wife	Own-child	Sales
Logistic Regression	capital-gain	capital-loss	fnlwgt	age	hours-per-week	education- NaN	Married-civ-spouse	Husband	Never-married	Female
Bayes Classifier (GaussianNB)	capital-gain	capital-loss	age	hours-per-week	education- NaN	fnlwgt	Married-civ-spouse	Husband	Own-child	Married-civ-spouse
Linear Regression	capital-gain	fnlwgt	capital-loss	age	hours-per-week	education- NaN	Married-civ-spouse	Husband	Never-married	Male
Weighted Quadratic Discriminant Analysis(base)	education- NaN	Married-civ-spouse	Husband	HS-grad	Male	Female	Never-married	Not-in-family	Some-college	Bachelors

Fracture Occur Frequency

(6)	(5)	(4)	(3)	(2)	(1)
Married-civ-spouse, education- NaN	capital-gain, age, Husband, hours-per-week, fnlwgt , capital-loss	Never-married	(None)	Own-child Female Male	Exec-managerial Wife, Sales, HS-grad Not-in-family Some-college, Bachelors

The number of Features Occur more than three times in other algorithms.

Random Forest	Decision Tree	Logistic Regression	Bayes Classifier (GaussianNB)	Linear Regression	WQDA (base)
9	7	9	8	9	4

Accuracy and run time for the 6 algorithms.

Algorithm	Run time(seconds)	Accuracy
Random Forest	120.90738	0.86796
Decision Tree	12.53975	0.85636
Logistic Regression	4.24601	0.79563
Bayes Classifier	37.89371	0.79529
Linear Regression	3.19304	0.77933
Weighted Quadratic Discriminant Analysis(baseline)	35.71141	0.35250

The baseline algorithm did purely in the comparison since only 4 of the features Occur more than three times in another algorithm. All other algorithms have twice the accuracy of the base accuracy.

5 algorithms having a similar number in the Number of Features Occur more than three times in another algorithm.

With those algorithms, there is 3 run within 30 seconds, Decision Tree, Logistic Regression and Linear Regression. naive Bayes, linear regression and logistics regression have an accuracy of around 78-80%, and random forest, and Decision tree have an accuracy of around 85-87%. By considering the run time Logistic Regression, Linear Regression, and Decision tree have a run time of fewer than 30 seconds. By considering the criteria in the evaluation metric, the Decision tree is the best algorithm to find the most important features.

By ranking from the decision tree the most important weight related to income class is:
Married-civ-spouse, education-NaN, capital-gain, capital-loss, age, hours-per-week, fnlwgt, Wife, Own-child, Sales

Acknowledge:(Third Party Library)

Database: https://archive.ics.uci.edu/ml/datasets/Adult

Library Used:
Pandas,
category_encoders
numpy,
time,
sklearn