Университет ИТМО
Кафедра ИПМ

Машинное обучение
Лабораторная работа 2
«Деревья решения»

Выполнил:
Шаймарданов Руслан
группа P4117
Преподаватель:
Жукова Н. А.

Санкт-Петербург
2017

Выбранный датасет: «Statlog (Shuttle) Data Set»

http://archive.ics.uci.edu/ml/datasets/Statlog+%28Shuttle%29

Количество записей: 14500

Описание:

The shuttle dataset contains 9 attributes all of which are numerical. The first one being time. The last column is the class which has been coded as follows :

1	Rad Flow
2	Fpv Close
3	Fpv Open
4	High
5	Bypass
6	Bpv Close
7	Bpv Open

Алгоритм lab2.py

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
def main():
    dataset = pd.read_csv("shuttle.csv", header=None).values.astype(np.int32,
copy=False)
    count = len(dataset)
    print("Size", "CART\t\t\t", "Random forest", sep="\t")
    for train_part in range(60, 100, 10):
        data_train = dataset[0:int(count*train_part/100)]
        data_test = dataset[int(count*train_part/100 + 1):]
        tree = DecisionTreeClassifier() #CART
        tree = tree.fit(data_train[:, :-1], data_train[:, -1])
        tree = tree.score(data_test[:, :-1], data_test[:, -1])

        forest = RandomForestClassifier(n_estimators=100) #Random forest
        forest = forest.fit(data_train[:, :-1], data_train[:, -1])
        forest = forest.score(data_test[:, :-1], data_test[:, -1])
        if(tree == 1.0):    #formatted output
            tree=str(tree)+'\t'*3
        print(str(train_part)+'%', tree, forest, sep='\t\t')
main()
```

Вывод программы

Size	CART	Random forest
60%	0.999137782376	0.998792895327
70%	0.99931018625	0.999080248333
80%	0.999655053467	0.999310106933
90%	1.0	0.999309868875

Вывод

Данные методы показывают высокую степень точности даже при 60% выборке. Значит, данных достаточно для корректого обучения, но не слишком много, благодаря чему не произошло переобучения. В предложенной обработке метод CART оказался точнее, однако

этим можно пренебречь, так как зачастую происходит совпадение до 3 знака после запятой. Если провести еще несколько итераций запуска программы, можно наблюдать картину, когда Random forest показывает незначительно больший результат. Следовательно, для исходной выборки методы можно считать одинаково полезными.