Assignment 6

Implement random forest in Spark/Scala

Hailun Zhu ID[hailunz]

1. Parameters setting:

Number of the trees : N = 5;

1. Tables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| lineid[int, primary key] | f0[int] | f1[int] | f2[int] | f3[int] | f4[int] | f5[int] | label[int] |

* 1. train

lineid: current line id, unique;

f0~ f5: features;

label: label.

* 1. test

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| lineid[int, primary key] | f0[int] | f1[int] | f2[int] | f3[int] | f4[int] | f5[int] | label[int] |

lineid: current line id, unique;

f0~ f5: features;

label: label.

c) result (performance)

|  |  |  |
| --- | --- | --- |
| timestamp  [timestamp]  primary key | precision  [double] | recall  [double] |

timestamp: indicates the current time of the test. Unique.

precision: accuracy of the whole forest, correct prediction count(true positive + true negative)/ total labels.

Recall: true positive count/ total positive count.

Four tests result:

timestamp | precision | recall

--------------------------+-----------+----------

2015-11-15 17:37:27+0000 | 0.576621 | 0.594308

2015-11-18 15:54:18+0000 | 0.56778 | 0.622766

2015-11-15 17:36:28+0000 | 0.569417 | 0.637326

2015-11-15 18:10:54+0000 | 0.575639 | 0.591661

d) tree (tree object)

Stored in the forestModel folder