Ling 572 HW2

Daniel Campos dacampos@uw.edu

01/23/2019

1 Q1 Mallet DT Learner

(a) The command lines:

```
mallet import-file –input train.vectors.txt –output train.vectors
mallet import-file –input test.vectors.txt –output test.vectors –use-pipe-from train.vectors
vectors2classify –training-file train.vectors –testing-file test.vectors –trainer DecisionTree –report
test:raw test:accuracy test:confusion train:confusion train:accuracy >de1.stdout 2 >de1.stderr
tail de1.stdout
```

(b) What are the training accuracy and the test accuracy?

Train:0.1781
Test: 0.1

2 Q2 Mallet Different depth

Table 1: Mallet's DT learner with different depths

See table 1.

B) Looking at table 1 we can see that as we train deeper or model learns the distribution of the train set much better but that doesnt necissarily mean that it will perform better on our test set. This, as is common with most ML shows that we must be careful when focusing on performance on the train set because it may mean nothing in terms of true model performance.

Depth	Training accuracy	Test accuracy
1	0.1393	0.1033
2	0.1419	0.12
4	0.1781	0.10
10	0.6285	0.1133
20	0.9970	0.1367
50	1	0.1367

3 Q3 build_dt.sh

See table 2 and table 3

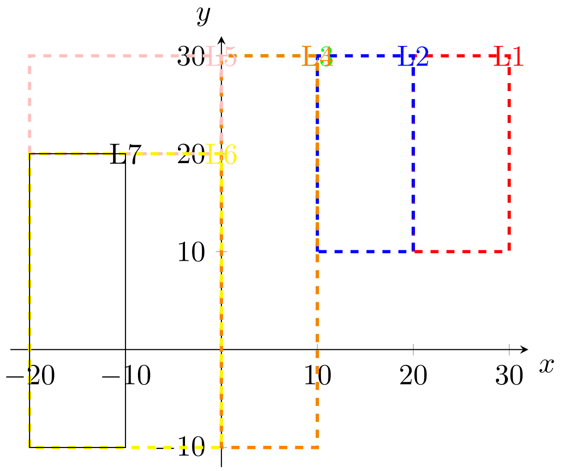
Table 2: build_dt.sh min_gain=0

Depth	Training accuracy	Test accuracy	CPU time (in minutes)
1	0.4530	0.4167	1
2	0.5207	0.53	2
4	0.6377	0.5267	3
10	0.7511	0.5933	10
20	0.8541	0.6733	11
50	0.96333	0.6933	33

Table 3: build_dt.sh min_gain=0.1

Depth	Training accuracy	Test accuracy	CPU time (in minutes)
1	0.6014	0.54	3
2	0.52	0.53	2
4	0.6014	0.54	4
10	0.6014	0.54	3
20	0.6014	0.54	3
50	0.6014	0.54	2

4 Q4



5 Q5: Notes

I believe everything is working great except when the depth of the tree is high the run time is slow. It me a while to move from pandas data frames to numpy arrays and when I did so my programs started going much faster. This assignment really showcases how improtant some of the hyperparameters like max dcepth and min gain are to both affect the runtime and the accuracy.