

Ling 572 HW2
Michael Roylance, Olga Whelan

Q1 (a) `mallet import-svmlight --input examples/train.vectors.txt --output train.vectors`
`mallet import-file --input examples/test.vectors.txt --output test.vectors --use-pipe-from`
`train.vectors`
`vectors2classify --training-file train.vectors --testing-file test.vectors --trainer DecisionTree`
`> dt.stdout 2>dt.stderr`

(b) train accuracy mean = 0.6377777777777778
test accuracy mean = 0.5233333333333333

Q2 It looks like Mallet DT learner treats the features as binary. After running the commands below, we get the same accuracy results as in Q1.

`./binarize.sh train.vectors.txt train.vectors.bin.txt`
`./binarize.sh test.vectors.txt test.vectors.bin.txt`
`mallet import-svmlight --input train.vectors.bin.txt --output train.vectors.bin`
`mallet import-svmlight --input test.vectors.bin.txt --output test.vectors.bin --use-pipe-from`
`train.vectors.bin`
`vectors2classify --training-file train.vectors.bin --testing-file test.vectors.bin --trainer`
`DecisionTree > dt.stdout.bin 2>dt.stderr.bin`
`diff dt.stdout dt.stdout.bin`

Q3 (a)

Depth	Training accuracy	Test accuracy
1	0.45296296296296296	0.4166666666666667
2	0.5207407407407407	0.5266666666666666
4	0.6377777777777778	0.5233333333333333
10	0.7514814814814815	0.6
20	0.8555555555555555	0.6833333333333333
50	0.9681481481481482	0.7
100	0.9685185185185186	0.7
1000	0.9685185185185186	0.7

(b) We can see that growing the depth of decision tree improves the training accuracy up until 100 nodes deep. The improvement at 100 nodes over 50 nodes is negligible, and there is no improvement at depth of 1000.

But creating a very deep tree overfits the training data and does not generalize it well, i.e. does not improve test accuracy a lot. It can be seen that changing the depth from 20 to 50 nodes brings a more than 10% improvement in training accuracy, but less than 2% improvement in testing.

Q5

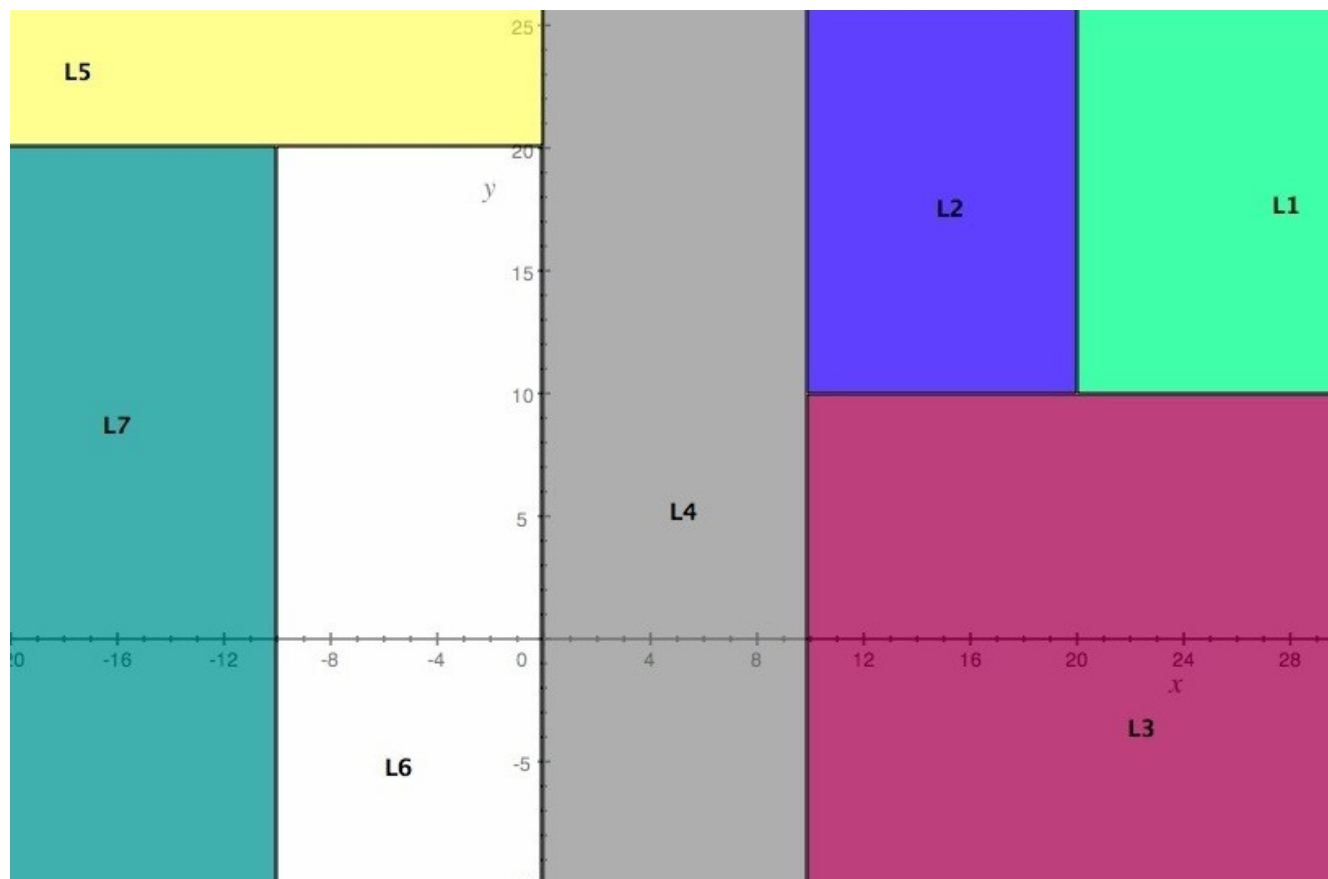
min_gain = 0

Depth	Training accuracy	Test accuracy
1		
2		
4		
10		
20		
50		

min_gain = 0.1

Depth	Training accuracy	Test accuracy
1		
2		
4		
10		
20		
50		

Q6 Here f_1 is x , f_2 is y .



Note: picture is cut off on the top and on the right. It is supposed to extend to $x=30$ and $y=30$.