CS534 MACHINE LEARNING

Da Lin 933334812

lind2@oregonstate.edu

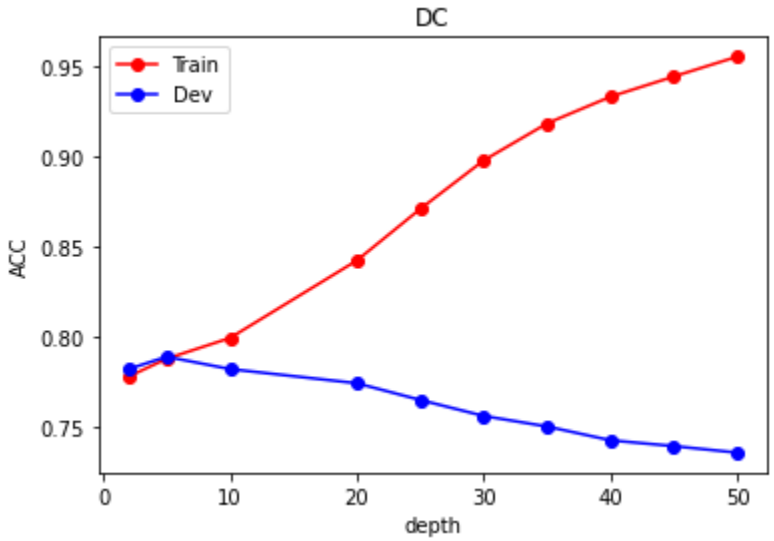
Introduction

In this assignment, students study the attributes of two classifier algorithms: decision tree and random forest. What I have done are: implement these two algorithms in python, plot results as a figure, compare the performance and discuss the problem between different depth. Moreover, the factor of overfitting and underfitting is also talk about in this report.

Part1:

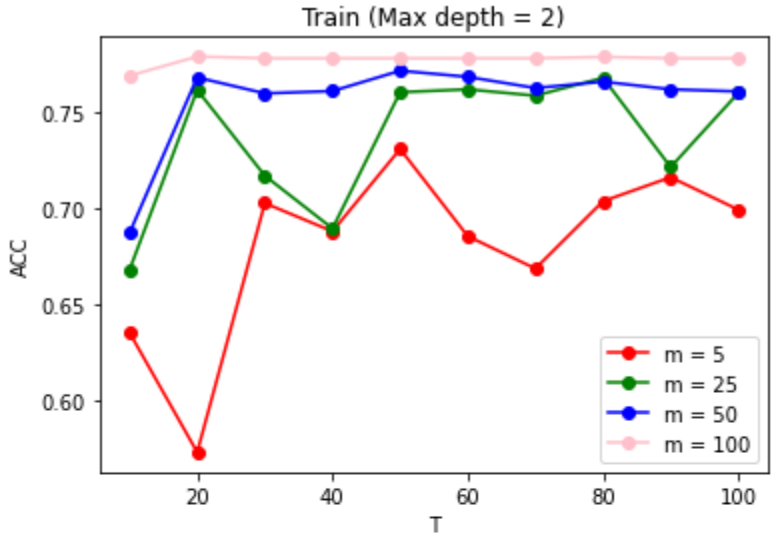
a. Root is feature 2, and the two splits immediately beneath the root are both feature 3.

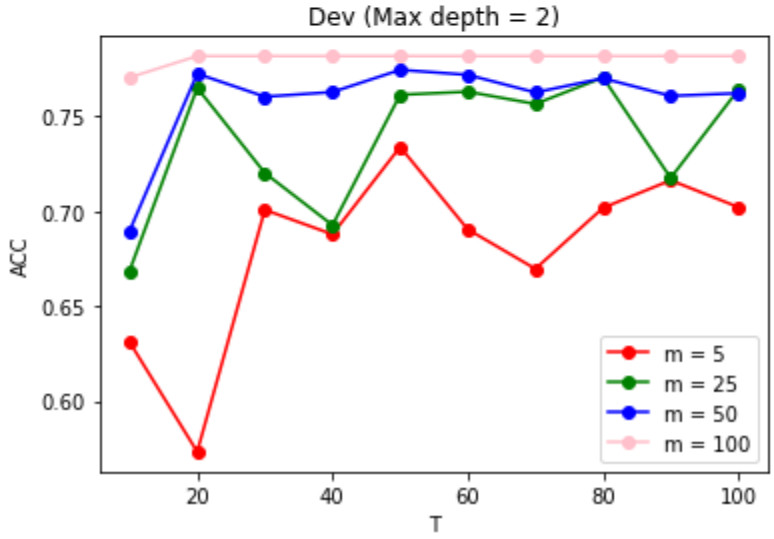
b.

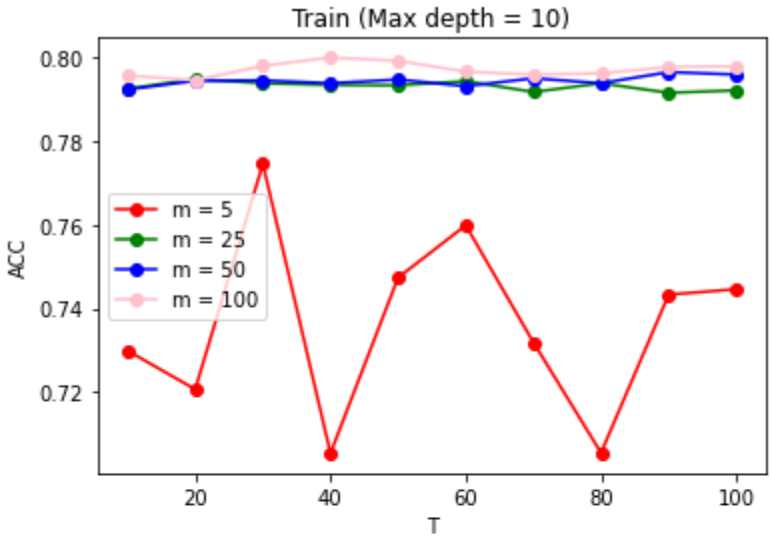


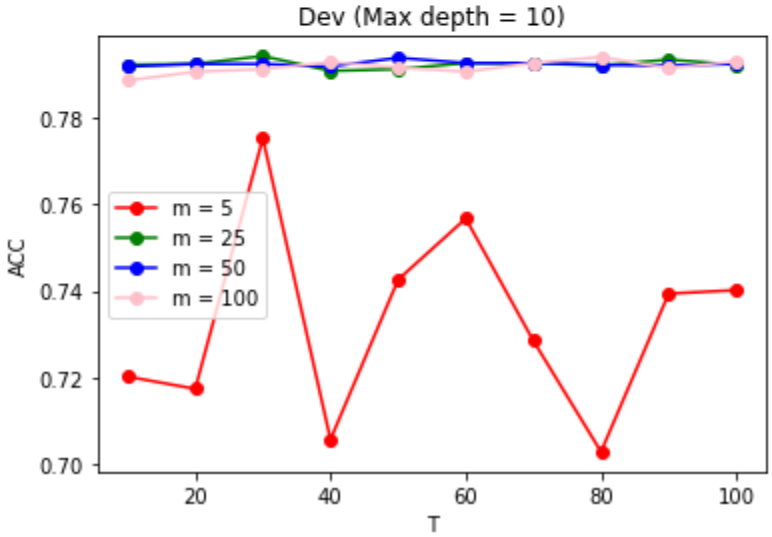
From the figure, we can see train and dev are split at dmax = 10, and the dev accuracy decrease. So, the overfitting start after dmax = 5.

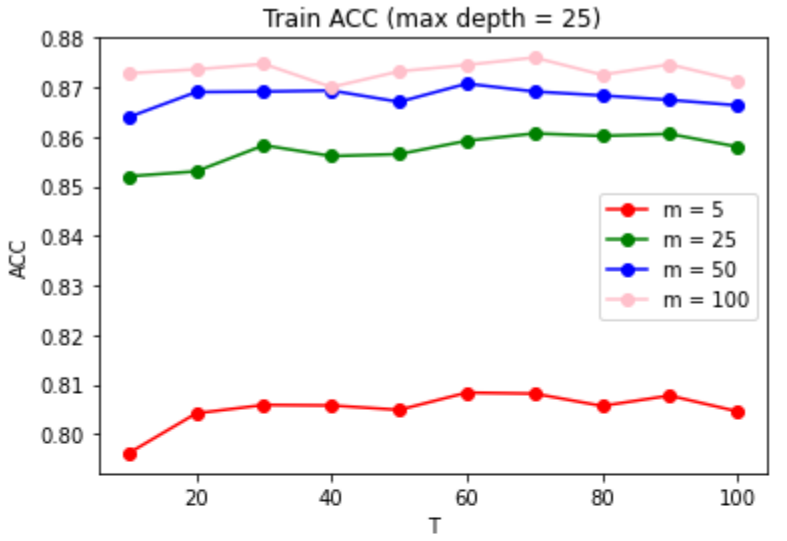
Part2:

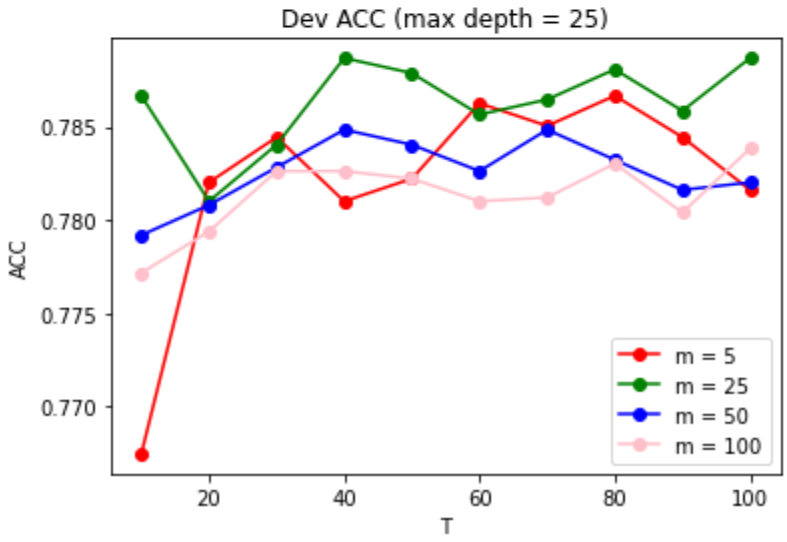

Train (Max depth = 2)
ACC
T
m = 5
m = 25
m = 50
m = 100


Dev (Max depth = 2)
ACC
T
m = 5
m = 25
m = 50
m = 100


Train (Max depth = 10)
ACC
T
m = 5
m = 25
m = 50
m = 100
0.80
0.78
0.76
0.74
0.72
20
40
60
80
100


Dev (Max depth = 10)
ACC
T
m = 5
m = 25
m = 50
m = 100
0.78
0.76
0.74
0.72
0.70
20
40
60
80
100





a. When dmax = 5, the model was underfitting, and when dmax = 25, the model was overfitting. The reason of underfitting is selected features are too small, so we can see the curve of m = 5 have low accuracy. The reason of overfitting is selected features are too much, in this case, the increasing m cause the decreasing accuracy.

b. When dmax = 5, we have large bias and small variance, so it causes underfitting. When dmax = 25, we have large variance and small bias, so it causes overfitting. However, in dmax = 10, the model has a good balance between bias and variance.

   For better performance, I think optimize the parameter like “max_features” and “number_trees” will be helpful, because in problem a, we can see one reason that causes over and underfitting is the number of features.