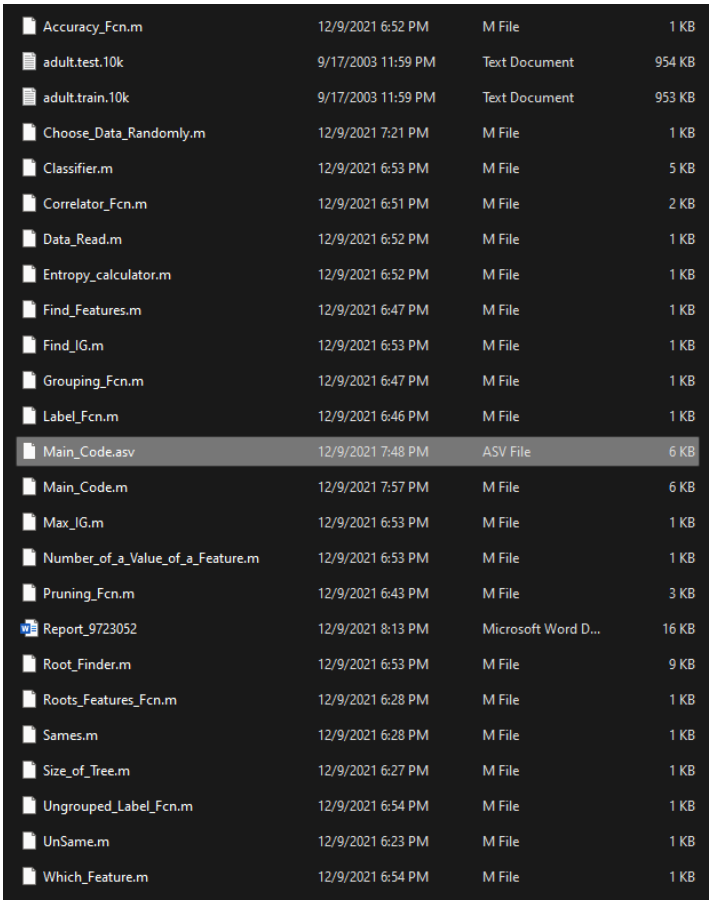
گزارش درخت تصمیم

علیرضا طباطبائیان – الیاس بصیری

ابتدا لازم به ذکر است که دستورات اصلی دارای کامنت و تمامی توابع دارای توضیحات میباشند . لذا به انجام قسمت الف و ب در این گزارش میپردازیم .

جهت اجرای کدها میبایست فایل Main_Code را اجرا نمایید .

بقیه فایل ها ، توابع نوشته شده و مورد استفاده میباشند .



قسمت الف :

تمامی داده های آموزش را استفاده کرده و درخت را ساختیم . سپس با استفاده از داده های تست ، آن را تست کردیم.

در اینجا به علت استفاده از روش اول عمق ، سایز درخت را برابر تعداد مسیر ها تعریف میکنیم.

a) در این گام ، 45 درصد دادگان آموزش را بصورت تصادفی به درخت دادیم و صحت را بر روی کل دادگان تست امتحان کردیم :

تکرار 1 : (سایز درخت یا همان تعداد مسیر ها : 1041)

صحت بر روی 45 درصد دادگان آموزش : 88.60 درصد

صحت بر روی کل دادگان تست : 80.10 درصد

تکرار 2 : (سایز درخت یا همان تعداد مسیر ها : 1015)

صحت بر روی 45 درصد دادگان آموزش : 88.98 درصد

صحت بر روی کل دادگان تست : 80.33 درصد

تکرار 3 : (سایز درخت یا همان تعداد مسیر ها : 1045)

صحت بر روی 45 درصد دادگان آموزش : 88.24 درصد

صحت بر روی کل دادگان تست : 80.69 درصد

میانگین 3 تکرار : (سایز درخت یا همان تعداد مسیر ها : 1033)

صحت بر روی 45 درصد دادگان آموزش : 88.60 درصد

صحت بر روی کل دادگان تست : 80.37 درصد

b) استفاده از 55 درصد دادگان آموزش :

تکرار 1 : (سایز درخت یا همان تعداد مسیر ها : 1179)

صحت بر روی 55 درصد دادگان آموزش : 88.73 درصد

صحت بر روی کل دادگان تست : 80.81 درصد

تکرار 2 : (سایز درخت یا همان تعداد مسیر ها : 1184)

صحت بر روی 55 درصد دادگان آموزش : 88.44 درصد

صحت بر روی کل دادگان تست : 80.31 درصد

تکرار 3 : (سایز درخت یا همان تعداد مسیر ها : 1192)

صحت بر روی 55 درصد دادگان آموزش : 88.11 درصد

صحت بر روی کل دادگان تست : 81.04 درصد

میانگین 3 تکرار : (سایز درخت یا همان تعداد مسیر ها : 1185)

صحت بر روی 55 درصد دادگان آموزش : 88.42 درصد

صحت بر روی کل دادگان تست : 80.72 درصد

b) استفاده از 65 درصد دادگان آموزش :

تکرار 1 : (سایز درخت یا همان تعداد مسیر ها : 1299)

صحت بر روی 65 درصد دادگان آموزش : 88.12 درصد

صحت بر روی کل دادگان تست : 80.24 درصد

تکرار 2 : (سایز درخت یا همان تعداد مسیر ها : 1375)

صحت بر روی 65 درصد دادگان آموزش : 88.49 درصد

صحت بر روی کل دادگان تست : 80.56 درصد

تکرار 3 : (سایز درخت یا همان تعداد مسیر ها : 1386)

صحت بر روی 65 درصد دادگان آموزش : 88.12 درصد

صحت بر روی کل دادگان تست : 80.90 درصد

میانگین 3 تکرار : (سایز درخت یا همان تعداد مسیر ها : 1353)

صحت بر روی 65 درصد دادگان آموزش : 88.24 درصد

صحت بر روی کل دادگان تست : 80.56 درصد

b) استفاده از 75 درصد دادگان آموزش :

تکرار 1 : (سایز درخت یا همان تعداد مسیر ها : 1506)

صحت بر روی 75 درصد دادگان آموزش : 88.05 درصد

صحت بر روی کل دادگان تست : 80.72 درصد

تکرار 2 : (سایز درخت یا همان تعداد مسیر ها : 1511)

صحت بر روی 75 درصد دادگان آموزش : 87.85 درصد

صحت بر روی کل دادگان تست : 80.96 درصد

تکرار 3 : (سایز درخت یا همان تعداد مسیر ها : 1581)

صحت بر روی 75 درصد دادگان آموزش : 88.01 درصد

صحت بر روی کل دادگان تست : 80.81 درصد

میانگین 3 تکرار : (سایز درخت یا همان تعداد مسیر ها : 1532)

صحت بر روی 75 درصد دادگان آموزش : 87.97 درصد

صحت بر روی کل دادگان تست : 80.83 درصد

b) استفاده از 100 درصد دادگان آموزش :

تکرار 1 : (سایز درخت یا همان تعداد مسیر ها : 1898)

صحت بر روی 100 درصد دادگان آموزش : 87.54 درصد

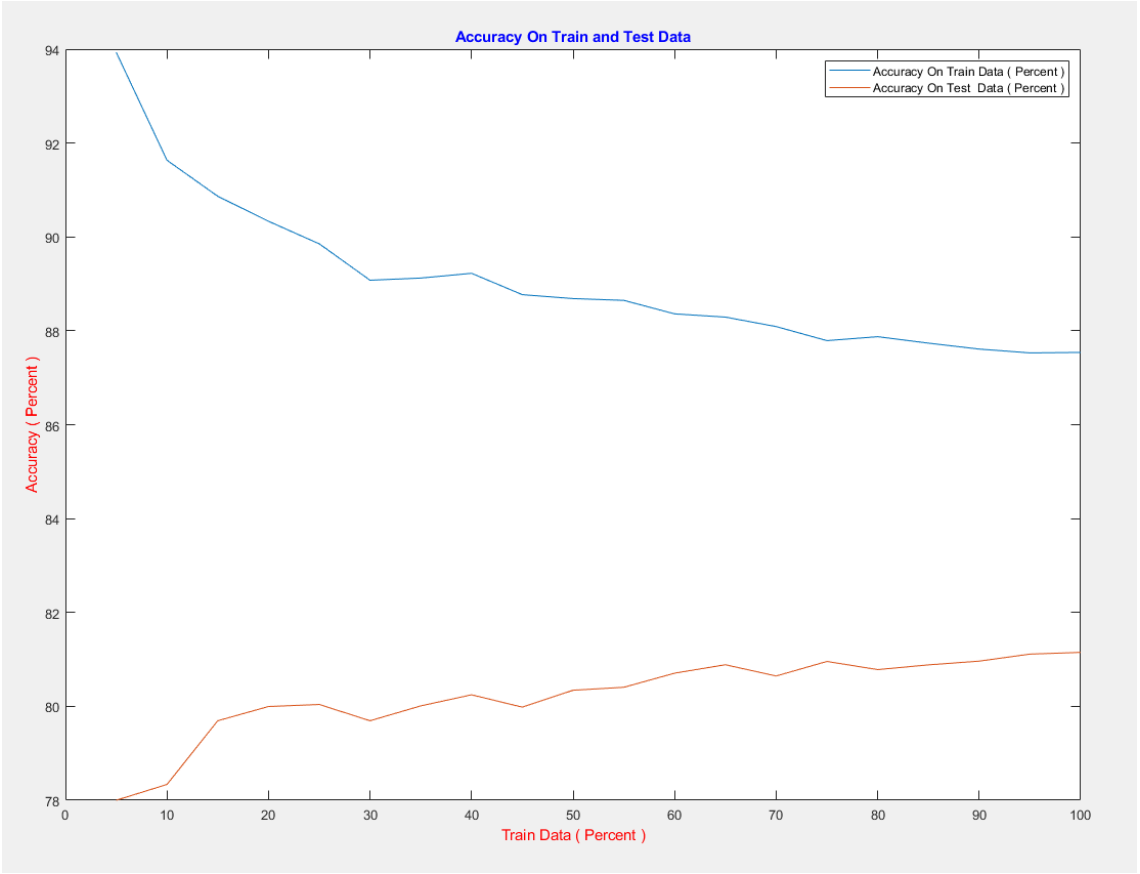
صحت بر روی کل دادگان تست : 81.15 درصد

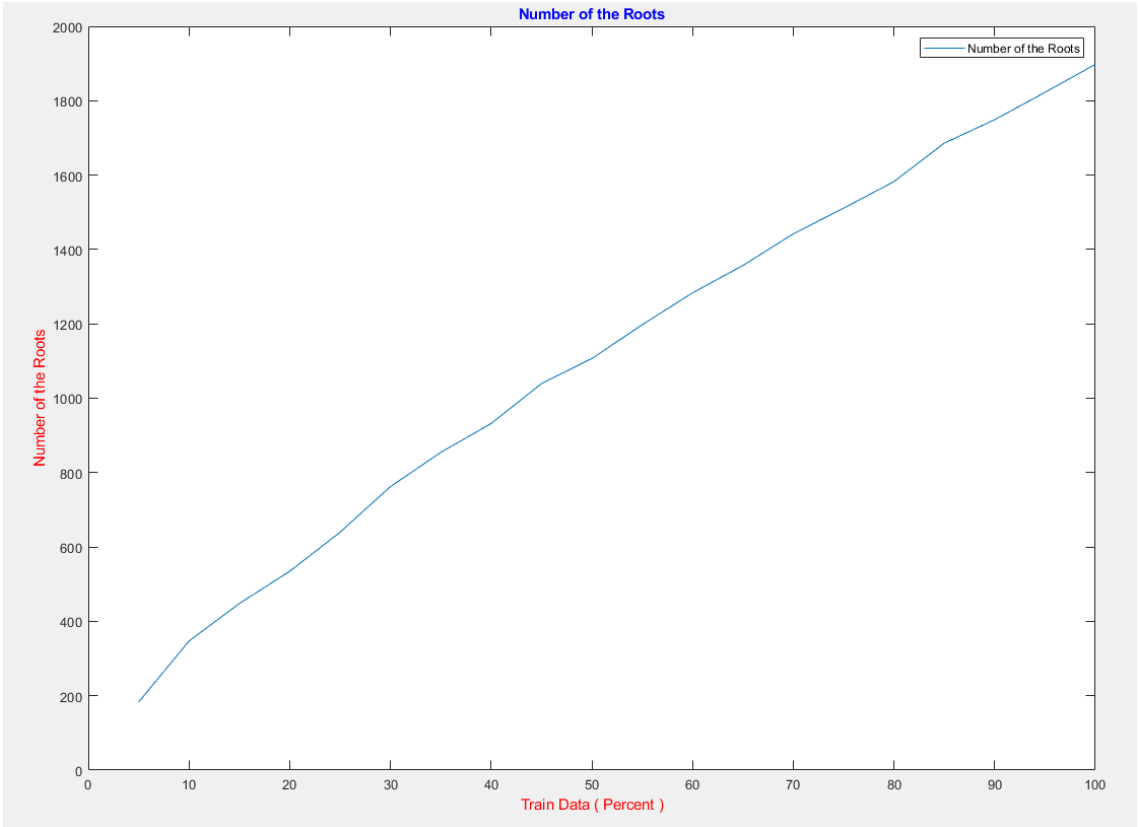
بحث بر روی تعداد دادگان آموزش و تاثیر آن در صحت و سایز درخت :

هر چقدر تعداد دادگان آموزش بیشتر شود ، صحت بر روی مجموعه آموزش کمتر میشود زیرا هنگامی که داده ها کم باشد ، درخت به راحتی overfit میشود اما با دادگان زیاد ، دیگر نمیتواند overfit شود و در نتیجه صحت کمتری بر روی دادگان آموزش دارد .

همچنین با افزایش دادگان آموزش ، صحت بر روی مجموعه تست صعودی است .

افزایش دادگان آموزش به صورت کاملا صعودی باعث افزایش طول درخت میشود .


Accuracy On Train and Test Data
Accuracy On Train Data (Percent)
Accuracy On Test Data (Percent)
Accuracy (Percent)
Train Data (Percent)
94
92
90
88
86
84
82
80
78
0
10
20
30
40
50
60
70
80
90
100


Number of the Roots
Number of the Roots
Number of the Roots
Train Data (Percent)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
80
90
100

قسمت ب : Pruning

اجرای هر قسمت این سوال حدود 10 ساعت نیاز به Run شدن داشت که واقعا فرصت این کار را نداشتم ولی کد ها کامل نوشته شده است و بدون مشکل Run میشود:

(a

Main_Code.m

```
104
105 -  for i = 1 : 50 % Prune the tree
106 -      [Roots , Test_Accuracy , Labels] = Pruning_Fcn(Data_Test,Roots,i,Features,Test_Accuracy,La
107 -      [Grouped_Data_Test , UnGrouped_Data_Test] = Grouping_Fcn(Roots,Data_Test_two,Features) ;
108 -      Roots_Features = Roots_Features_Fcn(Roots,Features) ;
109 -      [Number_of_the_Roots , ~] = size(Roots) ;
110
111 -      for k = 1 : Number_of_the_Roots
112 -          Grouped_Data_Test(k,2) = Labels(k,1) ;
113 -      end
114
115 -      [Row_UnGrouped_Data_Test,~] = size(UnGrouped_Data_Test) ;
116
117 -      for k = 1 : Row_UnGrouped_Data_Test
118 -          for j = 1 : 9
119 -              Data{1,j} = UnGrouped_Data_Test{k,j} ;
120 -          end
121 -          UnGrouped_Data_Test{k,Number_of_features+1} = Ungrouped_Label_Fcn(Data,Roots,Roots_Fea
122 -      end
123
124 -      Test_Accuracy_two  = Accuracy_Fcn(Grouped_Data_Test,UnGrouped_Data_Test,My_Label_Column) ;
125
126 -      i
127 -      Test_Accuracy
128 -      Test_Accuracy_two
129 -      A(1,i) = Test_Accuracy ;
130 -      A(2,i) = Test_Accuracy_two ;
131
132 -  end
133
```

بحث در مورد Overfitting و تاثیر Pruning :

با هرس کردن درخت ، از overfit شدن آن جلوگیری میشود و باعث کاهش صحت بر روی مجموعه آموزش و افزایش صحت در داده های تست میشود که این امر بسیار مطلوب است .