Report for MC3P1

Yuanda Zhu

For this project, a random tree learner, a bag learner and an insane learner have been implemented. The random tree learner is built using Cidler Pert method. A few tests have been done on Istanbul.csv data to explore factors that might influence RMSE which indicates overfitting when building the random tree learner.

Part 1: leaf size vs RMSE (without bag learner)

For this part, I vary the leaf size from 1 to 51.

There is overfitting when decreasing the leaf size. Overfitting occurs when leaf size is smaller than 5, where out of sample RMSE increases significantly while in sample RMSE decreases.

Leaf Size	RMSE in Sample	RMSE out of Sample	RMSE sum
1	0.00008	0.00910	0.00910
2	0.00000	0.00864	0.00864
3	0.00372	0.00808	0.00808
4	0.00467	0.00794	0.00794
5	0.00543	0.00766	0.00766
6	0.00563	0.00745	0.00745
7	0.00566	0.00758	0.00758
8	0.00588	0.00761	0.00761
9	0.00578	0.00752	0.00752
10	0.00593	0.00755	0.00755
11	0.00622	0.00752	0.00752
12	0.00619	0.00749	0.00749
13	0.00656	0.00736	0.00736
14	0.00651	0.00777	0.00777
15	0.00636	0.00731	0.00731
16	0.00680	0.00710	0.00710
17	0.00664	0.00762	0.00762
18	0.00669	0.00753	0.00753
19	0.00684	0.00766	0.00766
20	0.00700	0.00767	0.00767
21	0.00693	0.00785	0.00785
22	0.00697	0.00759	0.00759

23	0.00721	0.00730	0.00730
24	0.00695	0.00744	0.00744
25	0.00711	0.00764	0.00764
26	0.00722	0.00771	0.00771
27	0.00719	0.00744	0.00744
28	0.00721	0.00726	0.00726
29	0.00741	0.00768	0.00768
30	0.00704	0.00777	0.00777
31	0.00722	0.00752	0.00752
32	0.00743	0.00772	0.00772
33	0.00715	0.00771	0.00771
34	0.00745	0.00791	0.00791
35	0.00711	0.00768	0.00768
36	0.00728	0.00750	0.00750
37	0.00751	0.00752	0.00752
38	0.00740	0.00791	0.00791
39	0.00734	0.00788	0.00788
40	0.00773	0.00783	0.00783
41	0.00746	0.00823	0.00823
42	0.00725	0.00784	0.00784
43	0.00749	0.00778	0.00778
44	0.00771	0.00764	0.00764
45	0.00740	0.00787	0.00787
46	0.00766	0.00835	0.00835
47	0.00749	0.00826	0.00826
48	0.00737	0.00794	0.00794
49	0.00740	0.00778	0.00778
50	0.00768	0.00759	0.00759
51	0.00752	0.00796	0.00796

Table 1. The relationship between leaf size and RMSE, without bagging.

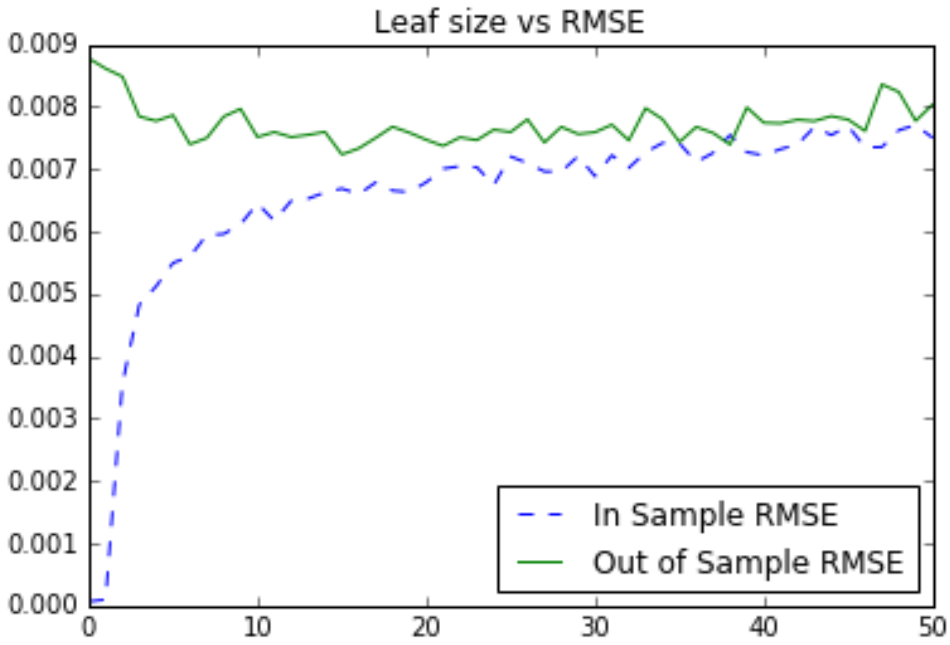

Figure 1. The relationship between leaf size and RMSE, without bagging.

Part 2: leaf size vs RMSE (with bag learner)

For this part, I fix the number of bags to be 20 and vary the leaf size from 1 to 51.

There is no overfitting when decreasing the leaf size. Out of sample RMSE decreases slowly as leaf size decreases till around 30 and remains constant as leaf size keeps decreasing. Thus, bagging does eliminate overfitting.

Leaf Size	RMSE in Sample	RMSE out of Sample	RMSE sum
1	0.00246	0.00610	0.00856
2	0.00256	0.00649	0.00905
3	0.00292	0.00634	0.00926
4	0.00346	0.00589	0.00936
5	0.00389	0.00626	0.01014
6	0.00436	0.00630	0.01066
7	0.00443	0.00644	0.01086
8	0.00470	0.00644	0.01114
9	0.00502	0.00591	0.01093
10	0.00505	0.00614	0.01119
11	0.00515	0.00620	0.01135
12	0.00534	0.00612	0.01146
13	0.00526	0.00625	0.01151
14	0.00548	0.00643	0.01191
15	0.00561	0.00598	0.01159
16	0.00550	0.00626	0.01177
17	0.00556	0.00645	0.01201
18	0.00570	0.00619	0.01189
19	0.00576	0.00658	0.01234
20	0.00584	0.00613	0.01197
21	0.00590	0.00625	0.01215
22	0.00582	0.00657	0.01239
23	0.00584	0.00658	0.01242
24	0.00590	0.00661	0.01251
25	0.00594	0.00669	0.01263
26	0.00611	0.00637	0.01248
27	0.00618	0.00648	0.01265
28	0.00608	0.00667	0.01275
29	0.00602	0.00673	0.01275
30	0.00618	0.00634	0.01252
31	0.00621	0.00665	0.01286
32	0.00612	0.00671	0.01283
33	0.00614	0.00676	0.01290
34	0.00625	0.00653	0.01278
35	0.00630	0.00659	0.01288
36	0.00636	0.00648	0.01285
37	0.00638	0.00640	0.01278
38	0.00639	0.00641	0.01279

39	0.00637	0.00653	0.01290
40	0.00649	0.00652	0.01301
41	0.00650	0.00673	0.01322
42	0.00625	0.00677	0.01301
43	0.00650	0.00668	0.01319
44	0.00654	0.00671	0.01325
45	0.00649	0.00684	0.01333
46	0.00650	0.00724	0.01374
47	0.00663	0.00688	0.01351
48	0.00649	0.00651	0.01300
49	0.00664	0.00688	0.01353
50	0.00664	0.00637	0.01301
51	0.00677	0.00680	0.01357

Table 2. The relationship between leaf size and RMSE, with fixed 20 bags.

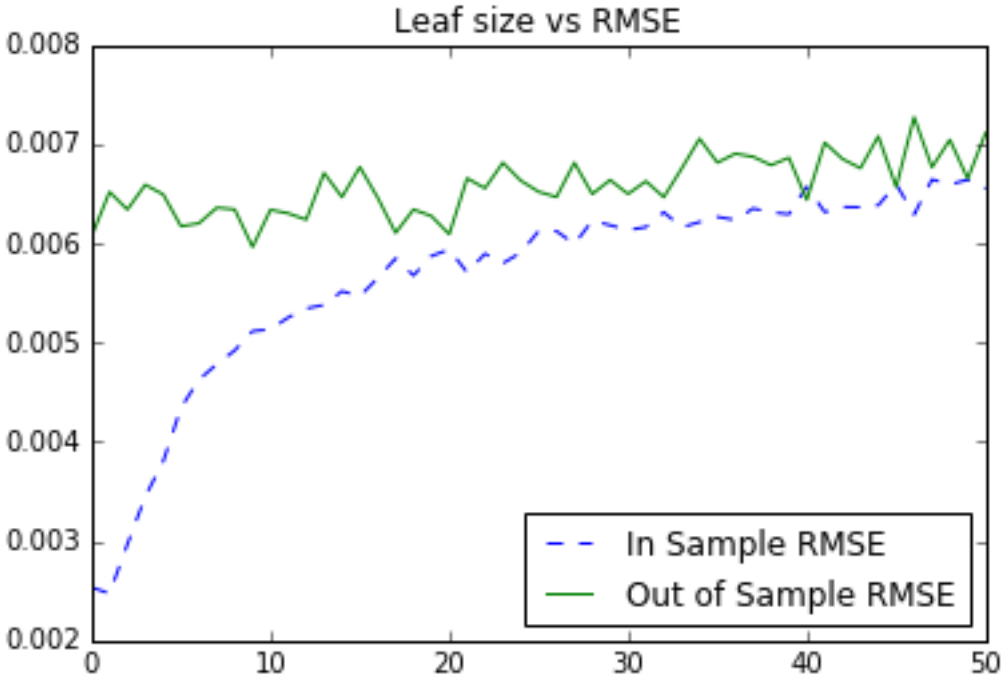

Figure 2. The relationship between leaf size and RMSE, with fixed 20 bags.

Part 3: bag size vs RMSE (with bag learner)

For this part, I fix the leaf size to be 5 and vary the number of bags from 1 to 51.

There is overfitting when decreasing the number of bags. Overfitting occurs when number of bags is smaller than 10, where out of sample RMSE and in sample RMSE both increases significantly.

Number of Bags	RMSE in Sample	RMSE out of Sample	RMSE sum
1	nan	nan	nan
2	0.00640	0.00795	0.01434
3	0.00547	0.00717	0.01264
4	0.00484	0.00683	0.01167
5	0.00478	0.00699	0.01178
6	0.00447	0.00675	0.01123
7	0.00449	0.00644	0.01093
8	0.00470	0.00648	0.01118
9	0.00458	0.00636	0.01094
10	0.00436	0.00656	0.01092
11	0.00451	0.00626	0.01077
12	0.00427	0.00626	0.01054
13	0.00439	0.00630	0.01069
14	0.00439	0.00606	0.01045
15	0.00422	0.00659	0.01081
16	0.00433	0.00644	0.01077
17	0.00433	0.00619	0.01052
18	0.00443	0.00605	0.01049
19	0.00431	0.00626	0.01056
20	0.00424	0.00661	0.01085
21	0.00437	0.00563	0.01001
22	0.00436	0.00600	0.01035
23	0.00438	0.00625	0.01063
24	0.00427	0.00631	0.01058
25	0.00418	0.00627	0.01045
26	0.00409	0.00656	0.01066
27	0.00433	0.00588	0.01021
28	0.00431	0.00596	0.01027
29	0.00428	0.00622	0.01050
30	0.00411	0.00661	0.01072
31	0.00424	0.00626	0.01050
32	0.00418	0.00626	0.01044
33	0.00427	0.00606	0.01033
34	0.00413	0.00632	0.01045
35	0.00429	0.00592	0.01021
36	0.00417	0.00614	0.01031
37	0.00431	0.00597	0.01028
38	0.00429	0.00615	0.01044
39	0.00419	0.00613	0.01032
40	0.00426	0.00593	0.01019
41	0.00427	0.00613	0.01040
42	0.00420	0.00607	0.01027
43	0.00424	0.00609	0.01033
44	0.00422	0.00614	0.01037
45	0.00423	0.00617	0.01040
46	0.00415	0.00643	0.01058
47	0.00429	0.00589	0.01019
48	0.00421	0.00599	0.01020
49	0.00416	0.00620	0.01036
50	0.00419	0.00612	0.01031

51	0.00406	0.00649	0.01055

Table 3. The relationship between number of bags and RMSE, with fixed leaf size of five.

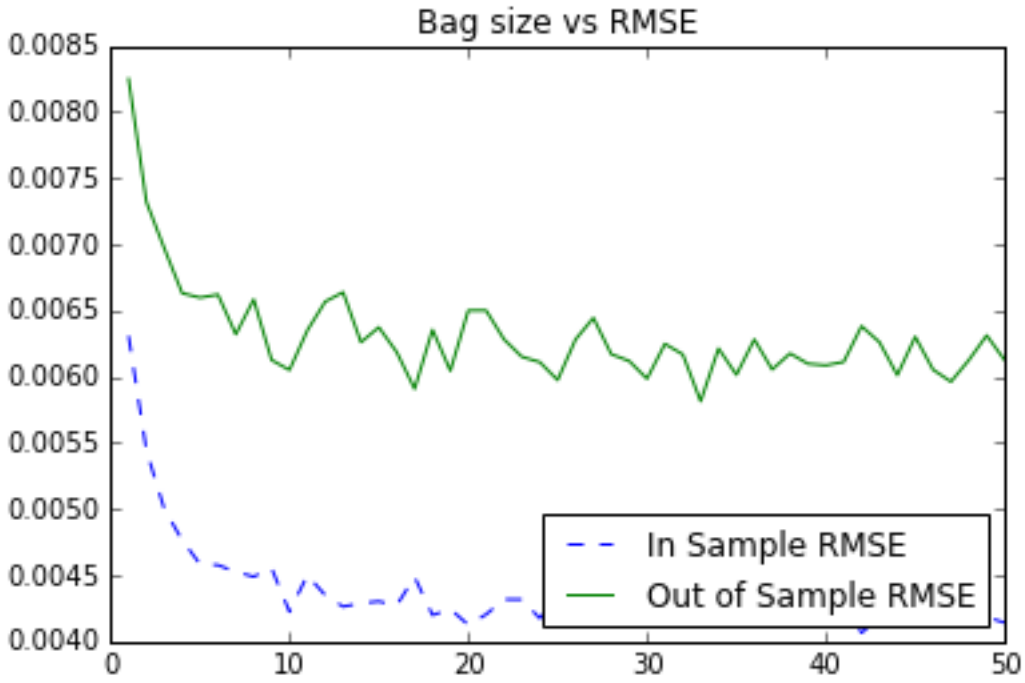

Figure 3. The relationship between number of bags and RMSE, with fixed leaf size of five.

Conclusion

Overfitting occurs as leaf size decreases below 5, when without bagging. Overfitting does not occur with respect to leaf size when incorporating bagging of 20. In other word, bagging does eliminate overfitting. Over fitting occurs as number of bags decreases below 10, with fixed leaf size of 5.