02/03/18 Homework #1 Harsha Kokel

CS 6375 Machine Learning hxk171430

REPORT

Table 1 shows the test set accuracy for various combinations of L & K values. Accuracy value mentioned is the average of accuracy obtained after pruning the decision tree 10 time with specific L & K.

		Average accuracy on test set			
		Data Set 1		Data Set 2	
L	K	Information gain Heuristics	Variance Impurity Heuristics	Information gain Heuristics	Variance Impurity Heuristics
Pre-Pruning		74.15	68.7	74.5	74.5
10	10	74.31	68.465	75.08333333	75.01666667
10	30	74.2	68.55	74.40740741	75.05555556
100	10	74.74	69	75.45	77.53333333
100	20	74.355	68.865	74.71666667	76.86666667
100	30	73.95	68.21	75.24074074	77.01851852
200	12	74.945	69.48	76.53333333	77.7
200	25	74.6	69.48	75.88333333	77.11666667
500	10	74.91	68.91	77.4	78.3
500	20	74.9	68.68	76.83333333	78.45
700	12	75.705	69.325	76.93333333	78.56666667
700	25	75.08	69.015	76.3	78.5
1000	10	75.18	69.395	76.1	78.78333333
1000	25	74.905	69.635	76.16666667	78.25925926
2000	10	75.185	69.36	77.90740741	78.92592593
5000	25	75.745	69.255	76.2962963	78.16666667

The graph below compares both heuristics.

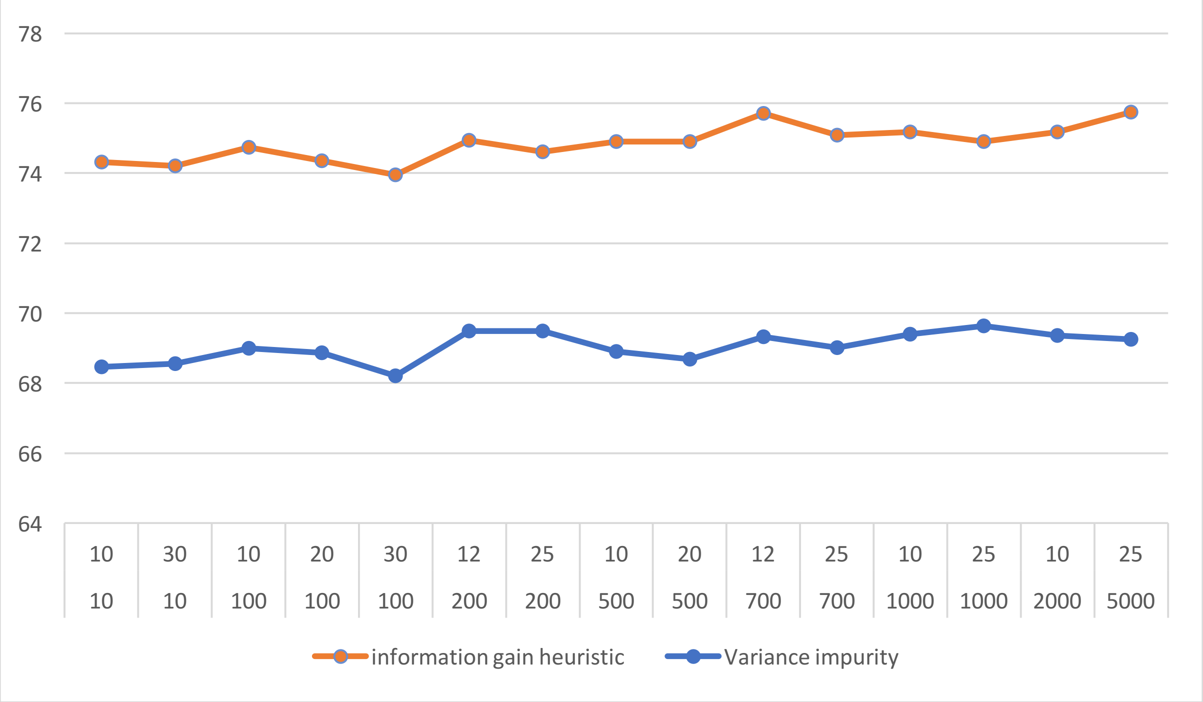


Figure 1 Accuracy plot for Data set 1

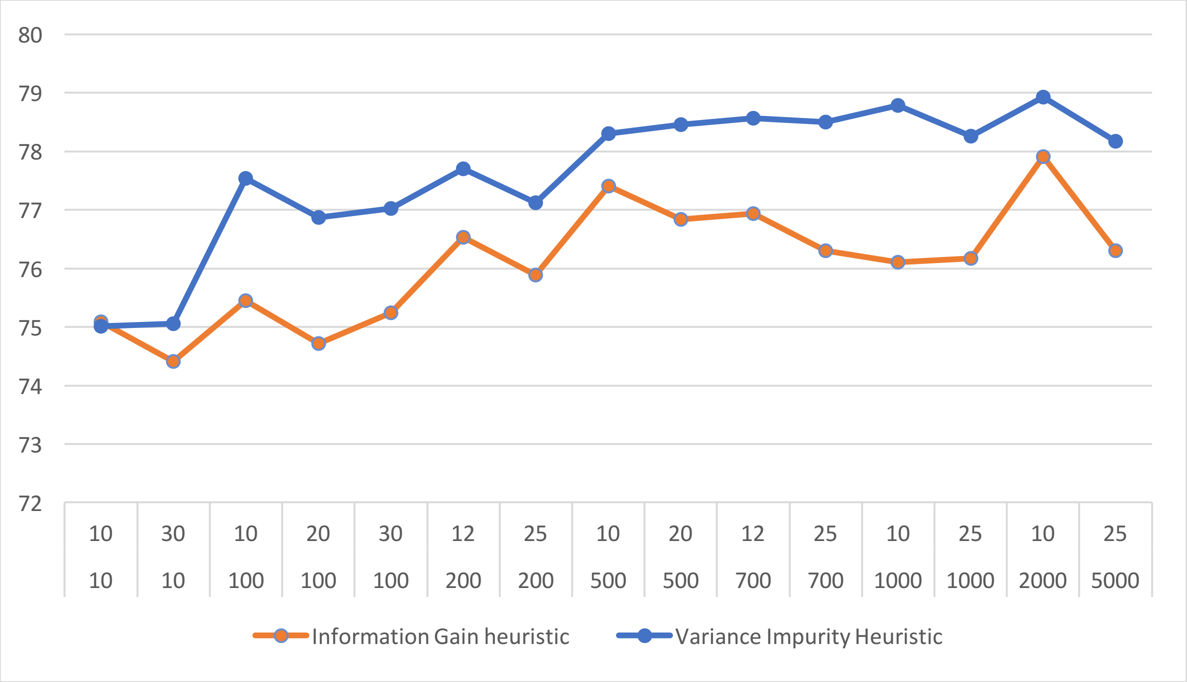
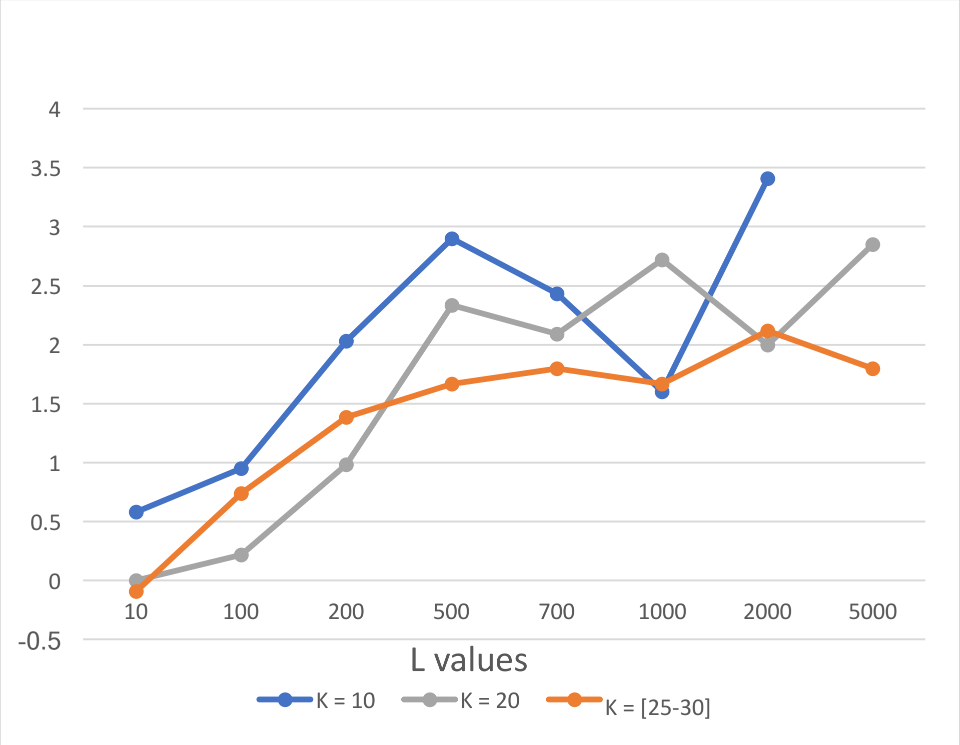


Figure 2 Accuracy plot for Data set 2

I found that Information Gain heuristic performs better on data set 1 where as variance impurity heuristic performs better on data set 2. Also, as L increases the improvement in accuracy becomes constant and depending on number of features in data set the accuracy reduces as value of K goes beyond the number of features. The graph below plots improvement in accuracy against the L value for different k values.


L values
K = 10
K = 20
K = [25-30]