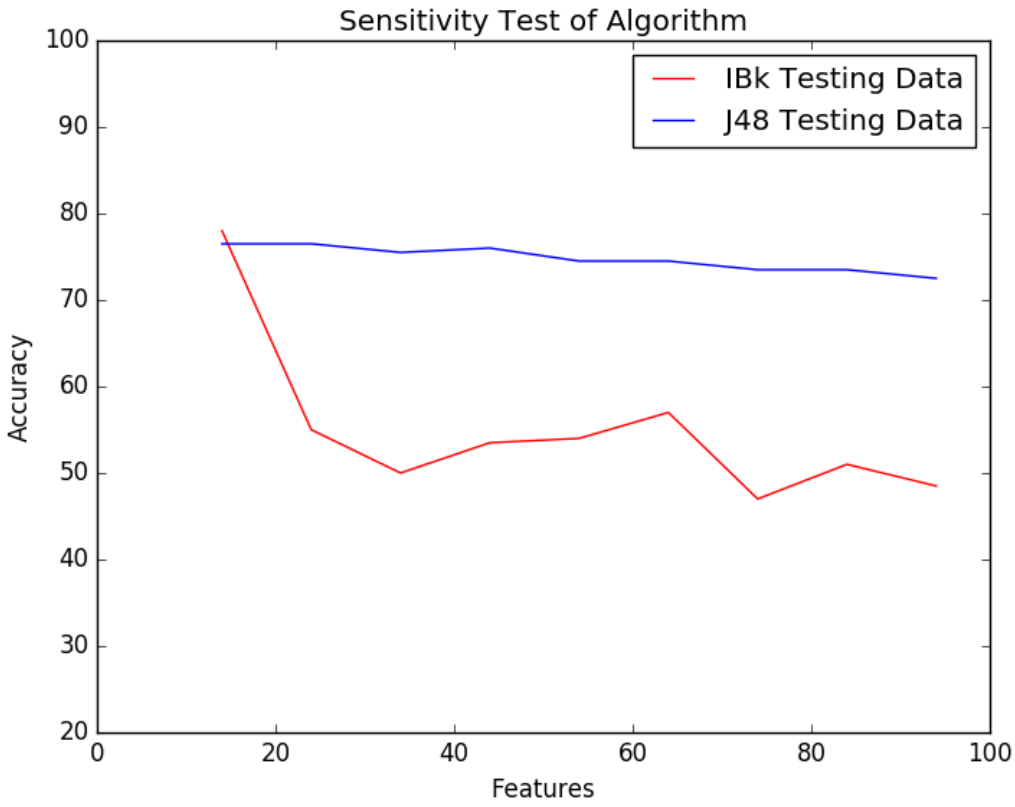
Project 1

Zheyi Yi

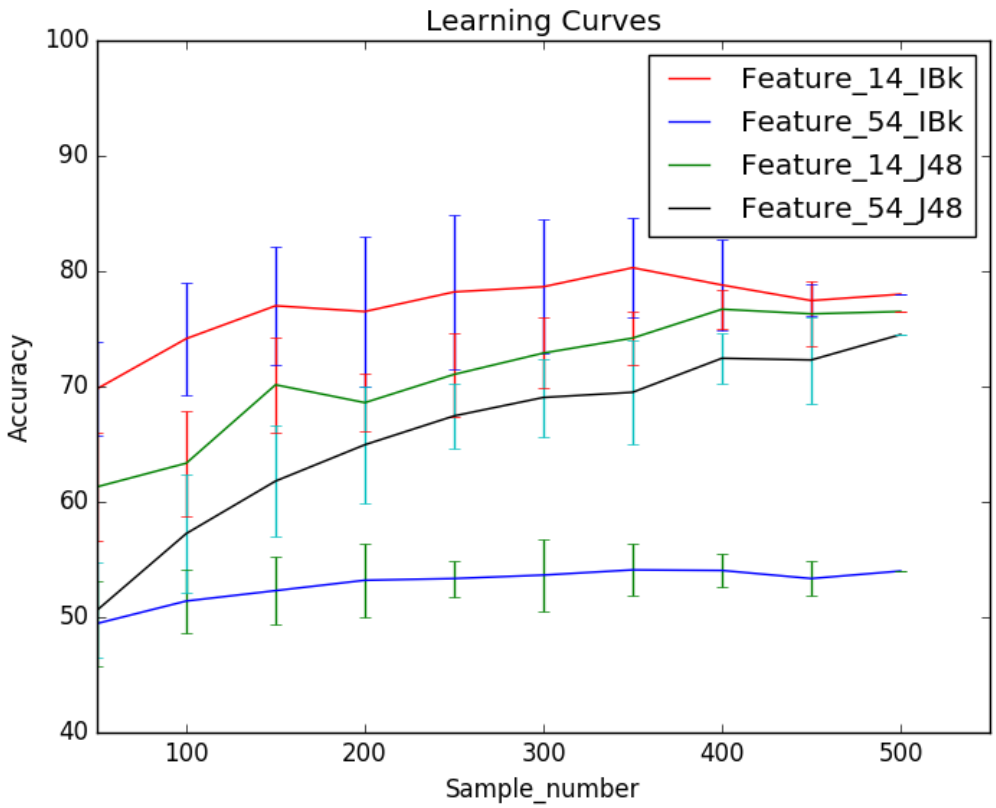
Part 1.



From the graph, there is just a little fluctuation for J48 but big fluctuation for IBk, so J48 algorithm seems to be very sensitive to the number of features however, IBk is not. IBk is sensitive to noise and irrelevant features because even irrelevant features have the same influence on the classification with related features so more features, more inaccurate.

J48 is not sensitive to features because it split based on the gain, unrelated features will get low value of gain and won't split in the beginning, so it won't affect a lot the accurate.

Part 2.



From the conclusion of the graph of the first part, we see IBk is sensitive to the number of features and J48 is not so sensitive compared with IBk. In this graph, we see feature_14_IBk is much more accurate than that of feature_54_IBk. And there are no big difference about accuracy for feature_14_J48 and feature_54_J48. The reasons are the same with my explanation for part 1.

From the graph, we see the accuracy of J48 increase a lot with the increase of number of sample but the accuracy of IBk doesn't. The reason is that KNN algorithm (IBk) only care about the nearest k, (k = 1 in this case), in the beginning, it increase with the number sample increase(from 0 to 150) but later increasing sample actually doesn't affect a lot for this algorithm.

The reason is that for J48, more samples means more accurate to calculate gain to split data so more sample means more accurate.

The deviation of these curves decrease with the increasing number of sample because with more sample, we are more confident to the data.