Cheng-Kuei Chiu

DSCI - 552: Machine Learning for Data Science
Assignment - 4
Theory Questions

1) How can we detect outliers after hierarchical clustering?

2) In building a regression tree, instead of the mean we can use the median, and instead of minimizing the squared error we can minimize the absolute error. Why does this help in the case of noise?

① After hierachical clustering , we can detect the outlier by seeing that if there's a point really far away from the current clustering , or we can calculate the distance , if it's too far, then it's a out lier

② using the median is because that mean is too sensitive to the noise and outlier a big outlier can shift mean really far away . instead median is more robust to the outlier . Same for the square error. the square error will emphasize the large errors , instead the absolute error will be more robust to the outlier because it treated all errors equally , and less sensitive to the outliers.

and causing
a overfitting
to the outlier