Outlier Detection

Gaussian Distribution

Say we have a data set of received phone calls. Each number represents the number of phone calls received on a given day.

Phone calls = {10, 11, 11, 11, 12, 13, 14, 14, 15, 17, 22}

Q_1= (11 + 11)/2 = 11

Q_2= (12 + 13)/2 = 12.5

Q_3= (14 + 15)/2 = 14.5

IQR=14.5 - 11 = 3.5

We can identify an outlier if it is greater than Q_3+1.5(IQR) or lower than Q_1-1.5(IQR)

14.5 + 1.5(3.5) = 19.75

11 - 1.5(3.5) = 5.75

Therefore 22 is the only outlier, since it is greater than 19.75 and there is no entry less than 5.75.

Non-parametric

*k*NN outliers

Local Outlier Factor (LOF)

Similar to DBSCAN and OPTICS

The local outlier factor is based on a concept of a local density, where locality is given by k nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers.

The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters.