anomaly detection

noise vs outliers/anomalies = noise does not always produce unusual values, and noise are not interesting

techniques

statistical approaches = using data distribution, mean, variance

↳ grubb's test = using normal distribution, at each step detect and remove one outlier

↳ likelihood approach = M (majority distribution) + A (anomalous distribution) = D (data)

↳ mostly data distribution unknown

proximity based = points are far away points from other points, distance-based, $O(n^2)$

density based = the outlier score of an object is the inverse of the density around the obj

↳ problems with varying density regions → relative density with respect to k nearest neighbor

clustering based = when a point does not strongly belong to a cluster

reconstruction based = reduce data to lower dimensional, reconstruction error is the difference between original and reduced dimensionality version

one class SVM = constructing SVM model → for one class find a hyper plane max dist to origin

information theoretic approaches = measure how much info decreases when you delete an object

types of outliers

global outlier = point anomaly, when a point deviates from the rest of the data set

contextual outlier = conditional outlier, 50 kg is normal for adults but anomaly for a baby

collective outliers = subset of data objects collectively deviate from the whole dataset, even if the individual data objects may not be outliers