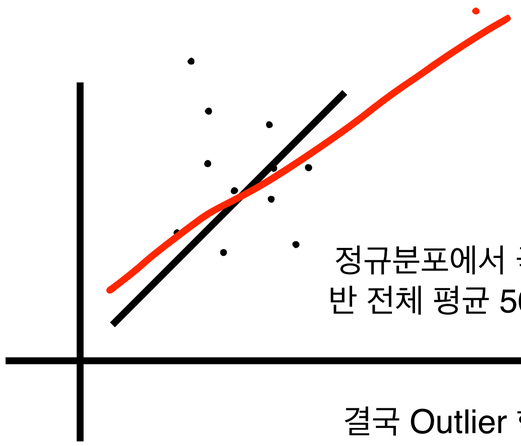
이상치 (Outlier)

240120_파이썬문법응용반_수업자료



왜 Outlier를 제거할까?
정규분포에서 극단에 있는 영향을 주는 값이라고 생각을 하면
반 전체 평균 50~60 대부분 , 혼자만 100점, 또는 0점의 값들

Overfitting
결국 Outlier 학습까지 정확하게 하다보면 과적합이 나온다!
회귀분석에서
회귀계수가 Outlier에 영향을 받아 변하게 된다.

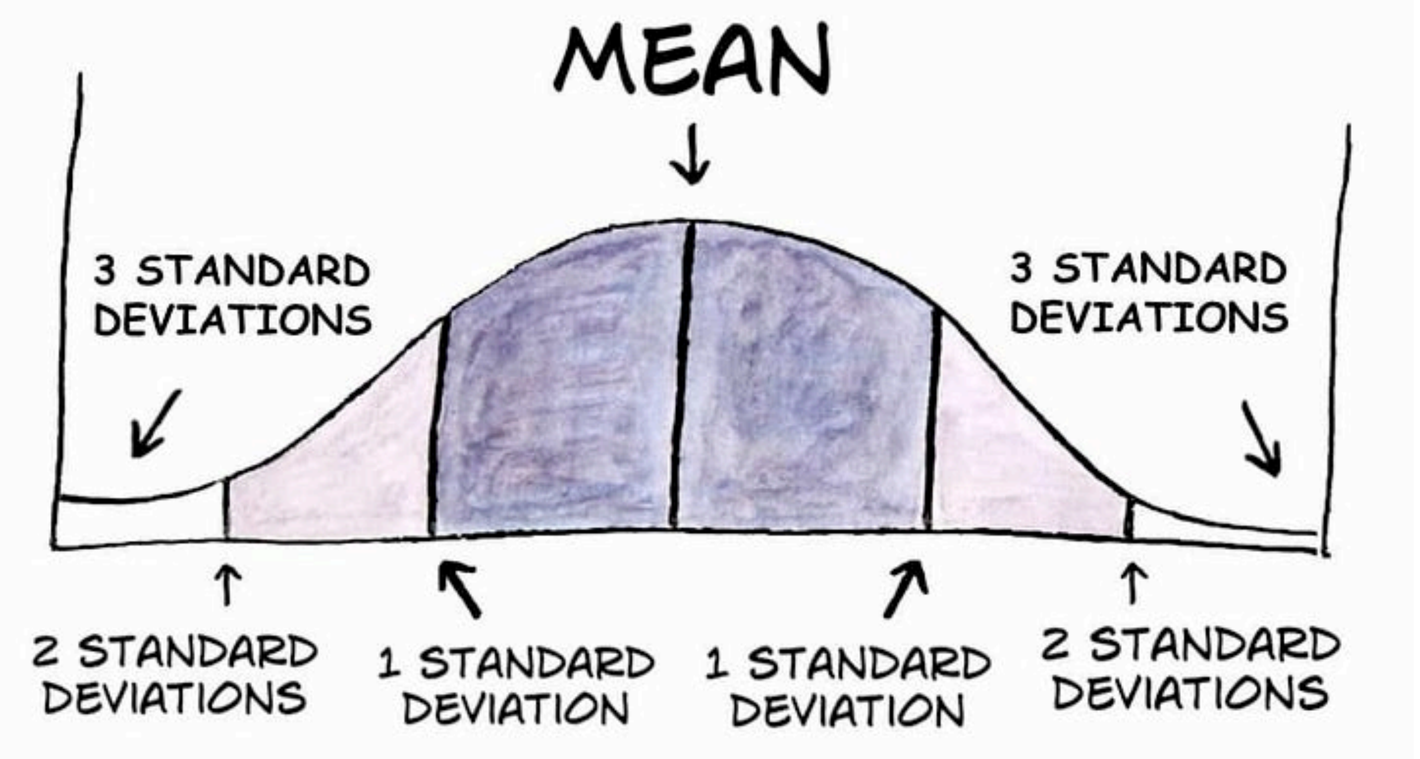
train, test
train 너무 과적합하게 되거나 영향을 받아서 -> 실제 test에 일반화 하지 못하는 성능

Outlier를 단순하게 제거하는 것으로 보면 안 된다.
이상치라는 것이
정말 다양한 것이 있지만
고등학교 시험 성적이 100점 만점
데이터가 대부분 평균이 50~60 인데

Outleir지만 실제 가능한 데이터
단순하게 제거한다는. 건-> 평균이 50~60짜리만 잘 예측하고
평균이 60이상 부터는 잘 예측하지 못하는 모델
100점 평균은 ?
0점 평균은?
두 군집으로 나눠볼 수 있다. 평균적인 점수 학생, 평균 100점인 친구들

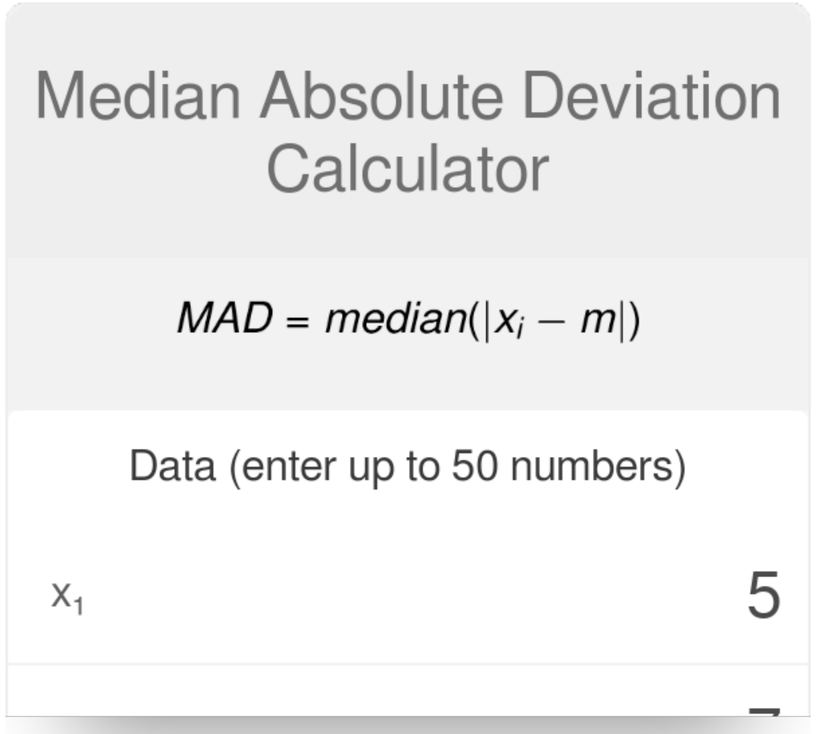
—잘못된 값일 확률이 높음
150점 평균은?

Outlier detection

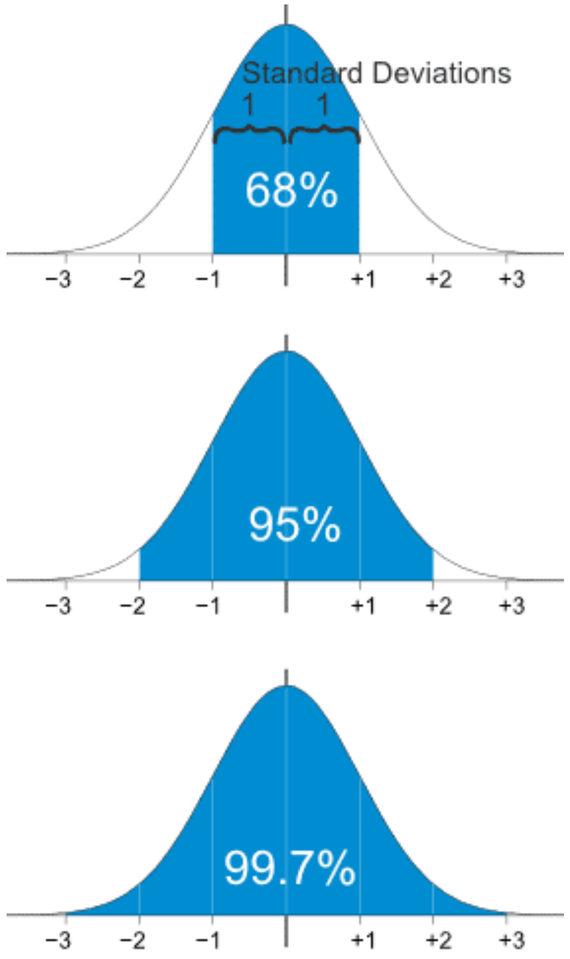


MAD

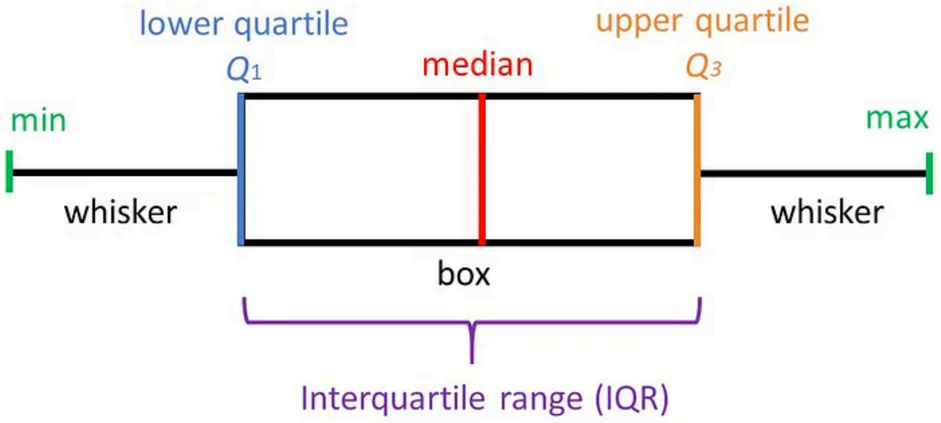
$$MAD = median(|x_i - m|)$$

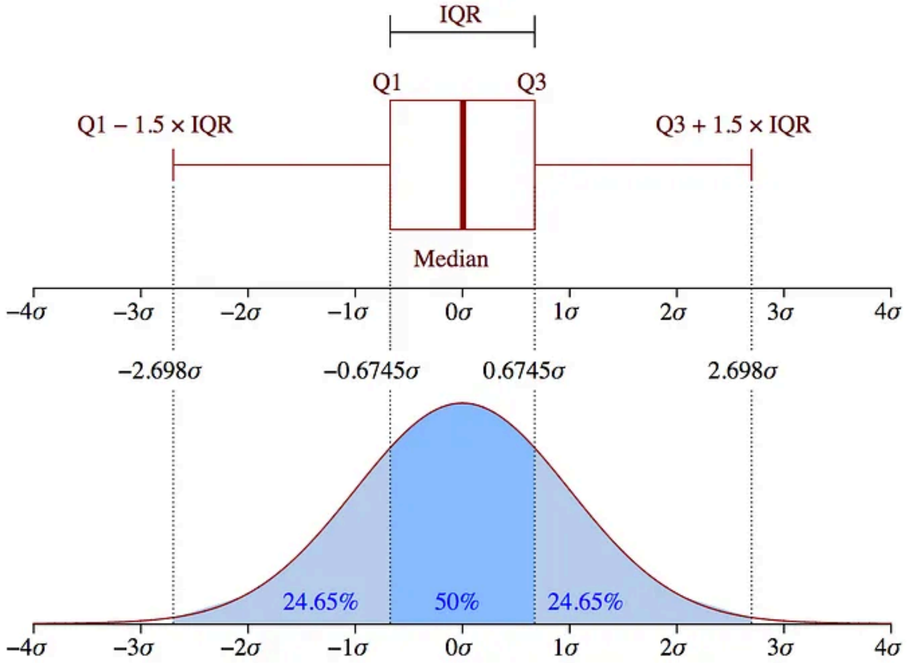


Standard Deviation

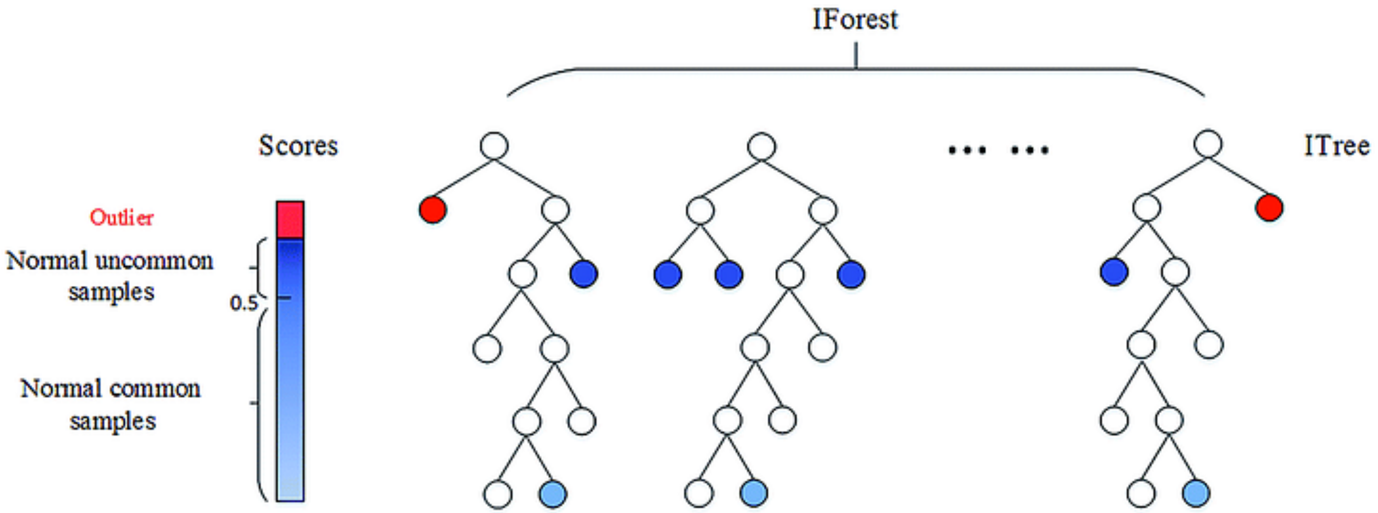


IQR


lower quartile
Q_1
median
upper quartile
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%

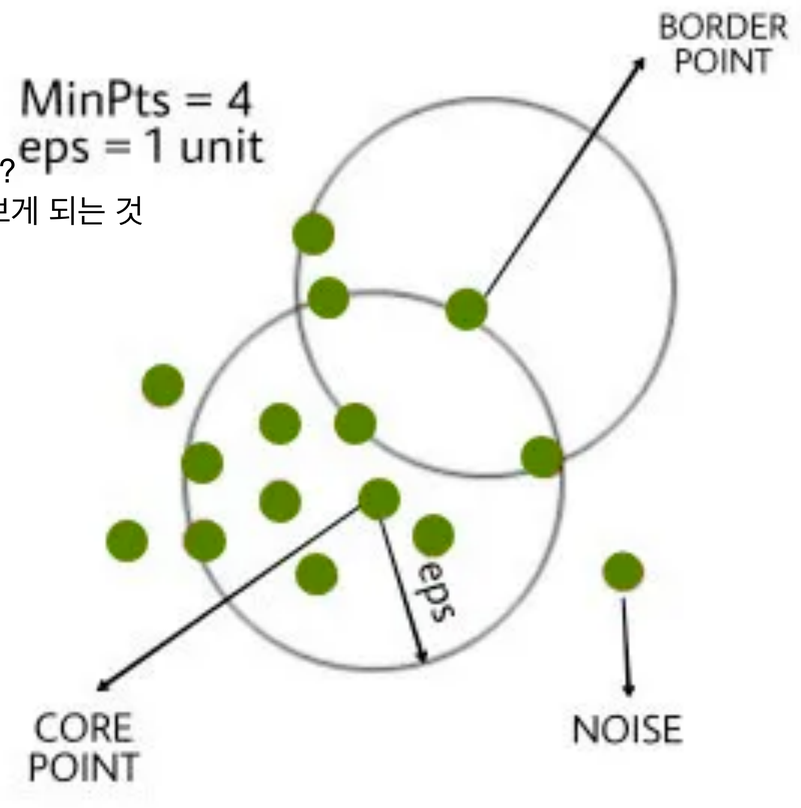
Isolation Forest



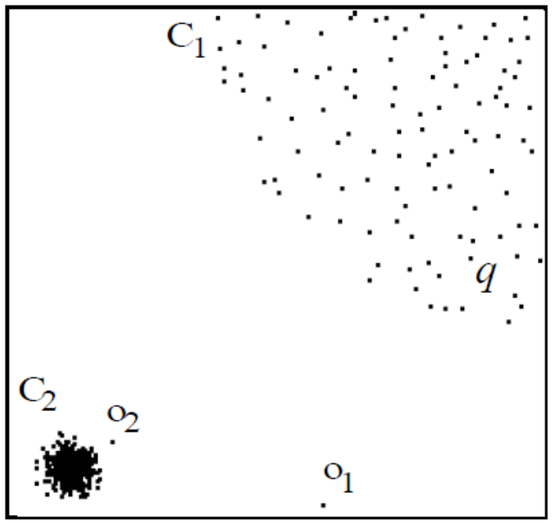
DBScan(Density Based Spatial Clustering of Application with noise)

K-means 거리기반으로 군집을 형성하는 것
밀도기반으로 군집을 형성하고
해당 밀도에 따라 이상치를 판단한가는 것
noise

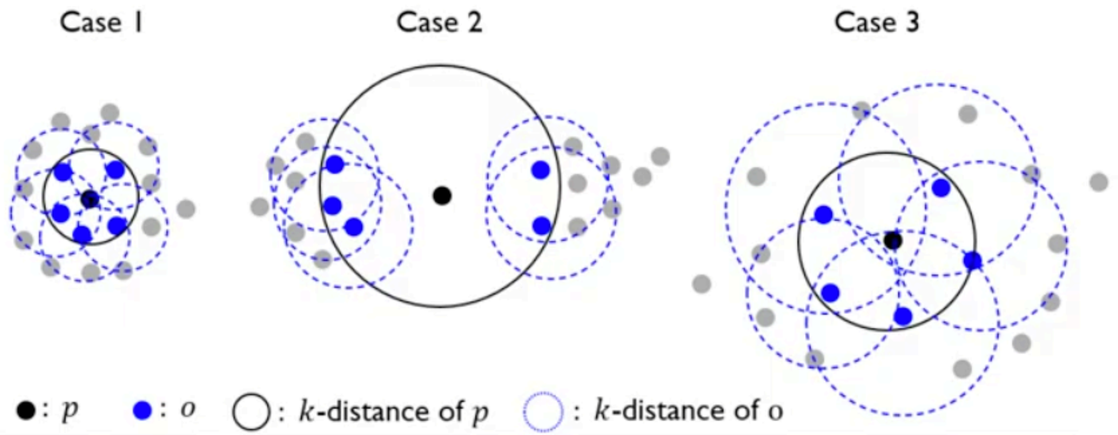
하나의 점 A , 점 A부터 E(eps)거리에 점이 몇 개 있는가?
정해서 그 라인 안에 원하는 점에 개수가 있으면 하나의 군집으로 보게 되는 것



LOF(Local Outlier Factor)



$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} = \frac{\frac{1}{lrd_k(p)} \sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|}$$



Case	$lrd_k(p)$	$lrd_k(o)$	$LOF_k(p)$
Case 1	Large	Large	Small
Case 2	Small	Large	Large
Case 3	Small	Small	Small

데이터 분석가의 도메인 지식을 통한 Outlier 선정