1 이상치란? # [edit]

← 예측 모델 학습시 성능에 안좋은 영향을 끼칠 수 있기 때문에, 예측 모델 시작지 '이상치 제거'를 보통 수행함

이상치(이상점, outlier)란, 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값을 말한다. 어떤 의사결정을 하는데 필요한 데이터를 분석할 경우 이렇게 이상한 값들에 의해서 의사결정에 영향을 미칠 수 있으므로 제거하는 것이 좋다.

2 사분위수 # [edit]

- 0사분위수(Q0): 최소값
- 1사분위수(Q1): 최소값 ~ 25% 번째 값
- 2사분위수(Q2): 중앙값
- 3사분위수(Q3): 중앙값 ~ 75% 번째 값
- 4사분위수(Q4): 최대값

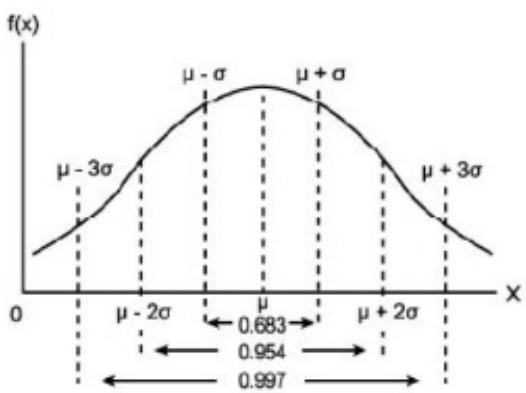
이상치는 보통 다음과 같이 계산된다. 여기서 '사분위범위'란 Q3 - Q1 구간을 말하며, 이 구간에는 50%의 데이터들이 있다.

```
IF 값 < (제1사분위수 - 1.5*사분위범위) OR (값 > 제3사분위수 + 1.5*사분위범위) THEN
  RETURN 이상치
ELSE
  RETURN 보통치
```

3 정규분포 # [edit]

← '정규분포'일 때

이상치는 정말 이상한 값이다. 위의 사분위수로 계산되는 것이 꼭 이상치는 아니다. 정규분포를 이용하여 어느 정도의 값이 이상치인지 직접 판단하여 이상치를 제거할 수도 있다. 일반적으로는 (m - 2σ) ~ (m + 2σ) 또는 (m - 1.5σ) ~ (m + 1.5σ) 구간을 벗어나는 값을 이상치로 판단하는 것이 좋다.(σ: 표준편차, m: 평균)



- 6 sigma - 100백만 중에 3.4
- 5 sigma - 100백만 중에 233
- 4 sigma - 100백만 중에 6210