베이즈_통계학

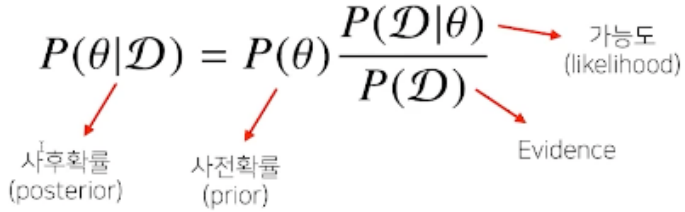
- 데이터가 새로 추가될 때 데이터가 업데이트하는 방식, 이론 설명, 방법론
- 조건부 확률

  $P(A \cap B) = P(B)P(A|B)$

  조건부확률 $P(A \cap B)$: 사건 B가 일어난 상황에서 사건 A가 발생할 확률
- **베이즈 정리**: 조건부확률을 이용하여 **정보를 갱신하는 방법**

  $P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B)\frac{P(A|B)}{P(A)}$

  A라는 새로운 정보가 주어졌을 때 $P(B)$로부터 $P(B|A)$를 계산하는 방법 제공



$$
\begin{aligned}
P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)} \\
&= \frac{P(B \mid A)P(A)}{P(B, A) + P\left(B, \neg A\right)} \\
&= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P\left(B \mid \neg A\right)P\left(\neg A\right)} \\
&= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P\left(B \mid \neg A\right)(1 - P(A))}
\end{aligned}
$$

- 베이즈 정리 EX)

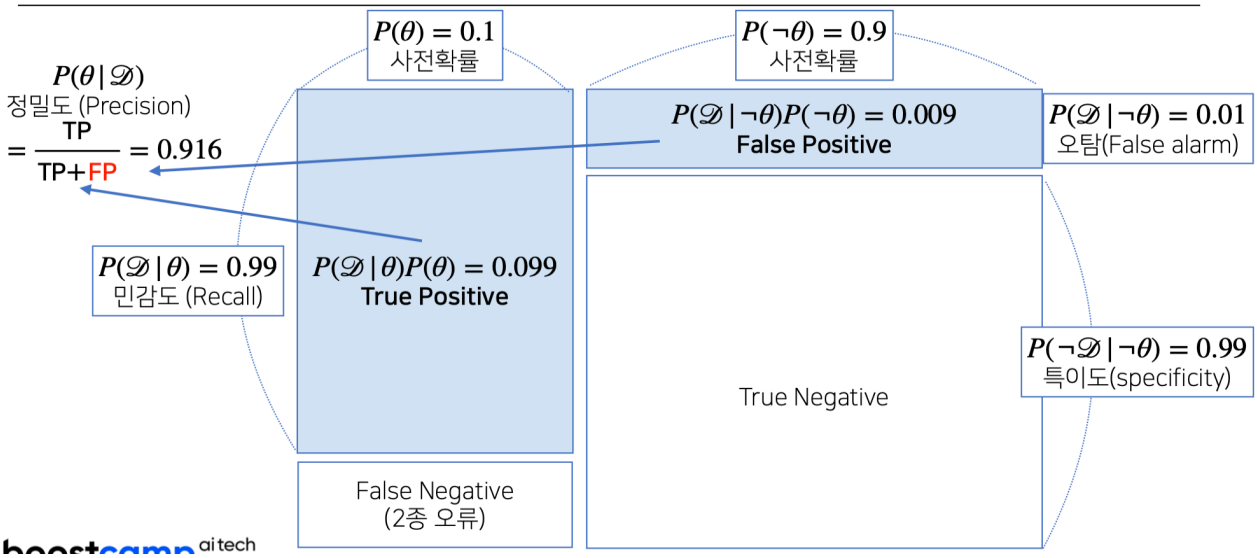
  COVID-99 의 발병률이 10% 로 알려져있다. COVID-99 에 실제로 걸렸을 때 검진될 확률은 99%, 실제로 걸리지 않았을 때 오검진될 확률이 1% 라고 할 때, 어떤 사람이 질병에 걸렸다고 검진결과가 나왔을 때 정말로 COVID- 99 에 감염되었을 확률은?

  $$P(\theta \mid \mathcal{D}) = P(\theta)\frac{P(\mathcal{D} \mid \theta)}{P(\mathcal{D})} \quad P(\theta) = 0.1 \quad \begin{matrix} P(\mathcal{D} \mid \theta) = 0.99 \\ P(\mathcal{D} \mid \neg\theta) = 0.01 \end{matrix}$$

  $$P(\mathcal{D}) = \sum_{\theta} P(\mathcal{D} \mid \theta)P(\theta) = 0.99 \times 0.1 + 0.01 \times \underset{P(\neg\theta)}{0.9} = 0.108$$

  $$P(\theta \mid \mathcal{D}) = 0.1 \times \frac{0.99}{0.108} \approx 0.916$$

  - 오탐율(False alarm)이 오르면 테스트의 정밀도(Precision)가 떨어진다.
- 조건부 확률의 시각화



- 베이즈 정리를 통한 정보 갱신

  베이즈 정리를 통해 새로운 데이터가 들어왔을 때 앞서 계산한 사후확률을 사전확률로 사용하여 **갱신된 사후확률 계산 가능**

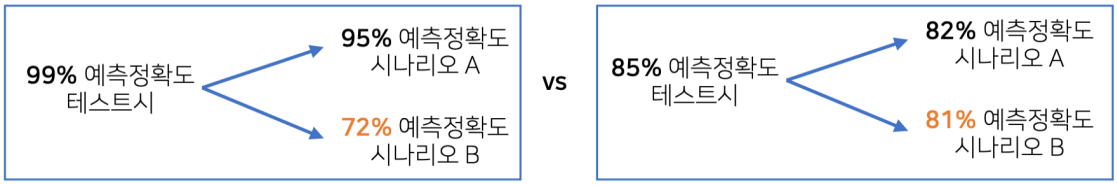
$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

사후확률 (posterior)

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

갱신된 사후확률 (posterior)

- 조건부 확률은 유용한 통계적 해석 제공, but **인과관계(causality)**를 추론할 때 함부로 사용 X
- 데이터가 많아져도 조건부 확률만 가지고 인과관계를 추론하는 것은 불가능
- 인과관계는 데이터 분포의 변화에 강건한 예측모형을 만들 때 필요
- 인과관계만으로는 높은 예측 정확도를 담보하기는 어려움



조건부확률 기반 예측모형 | 인과관계 기반 예측모형

- **조정(intervention)**을 통해 **중첩요인(confounding factor)의 효과를 제거**하고, 원인에 해당하는 변수만의 인과관계를 계산해야 함
- 만약 Z의 효과를 제거하지 않으면 **가짜 연관성(spurious correlation) 나옴**