투빅스 11기 정규과정

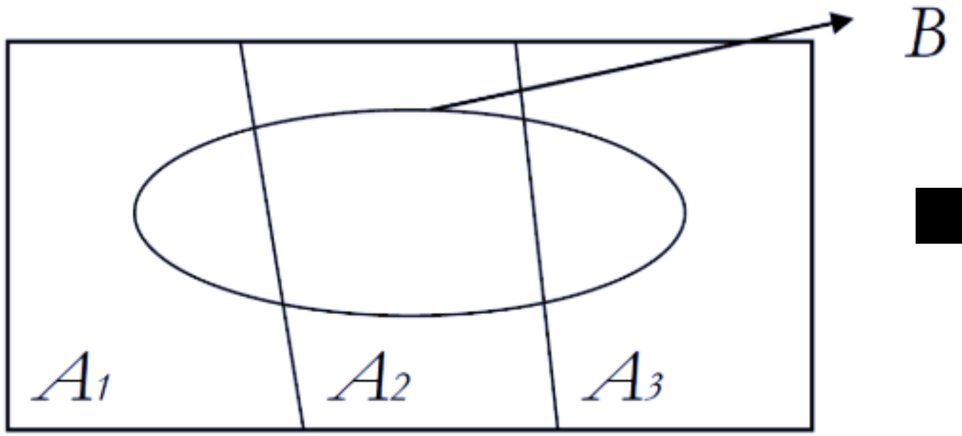
ToBig's 10기 박성진

Naive Bayes - Bayes Theorem

◆ Bayes' theorem & Conditional Probability

- 일반적으로, 사건 A_1, A_2, A_3 가 서로 배반(mutually exclusive)이고 A_1, A_2, A_3 의 합집합이 표본공간(sample space)과 같으면

- 사건 A_1, A_2, A_3 는 표본공간 B의 **분할**이라고 정의.
  우리가 관심있는 사건 B가 나타날 확률을 그림과 식으로 나타내면 다음과 같음



$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

◆ Bayes' theorem & Conditional Probability

- $P(B)$를 조건부확률의 정의를 이용
- 이를 전확률 공식(Law of Total Probability) 또는 베이즈 법칙이라 함.

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

$$P(B) = \sum_{i=1}^{3} P(B|A_i)P(A_i)$$

◆ Bayes' theorem & Conditional Probability

$$P(B) = \sum_{i=1}^{3} P(B|A_i)P(A_i)$$

- 보통 $P(A_i)$는 미리 알고 있다는 의미의 사전확률(Prior Probability)
- $P(B|A_i)$는 가능도(Likelihood Probability)

◆ Bayes' theorem & Conditional Probability

- 우리가 관심있는 사건인 B가 A_1에 기인했을 조건부 확률은?

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)}$$
$$= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)}$$

- $P(A_1|B)$는 사건 B를 관측한 후에 그 원인이 되는 사건 A의 확률을 따졌다는 의미의 사후확률(Posterior Probability)로 정의

- 같은 방식으로 $P(A_2|B)$, $P(A_3|B)$를 구할 수 있음

◆ Bayes' theorem & Conditional Probability

- 쿠키문제

- 쿠키가 들어 있는 그릇 두 개가 있다고 가정.
  첫번째 그릇에는 바닐라 쿠키 30개와 초콜렛 쿠키 10개가 들어있고,
  두번째 그릇에는 두 가지 쿠키가 종류별로 20개씩 들어 있음

- 어떤 그릇인지 보지 않고 한 그릇에서 임의로 쿠키를 집었는데 바닐라 쿠키
  그렇다면 이 때 '이 바닐라 쿠키가 그릇 1에서 나왔을 가능성'은 얼마?

◆ Bayes' theorem & Conditional Probability

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

- $P(H)$: 어떤 쿠키를 골랐던지 상관없이 그릇 1을 골랐을 확률. 문제에서는 그릇을 임의로 선택한 것이므로 0.5라고 가정할 수 있음. 이를 데이터를 관측하기 전의 확률, 즉 사전확률

- $P(D|H)$: 그릇1에서 바닐라쿠키가 나올 확률. 데이터가 가설에 포함될 확률, 즉 가능도

- $P(D)$: 바닐라 쿠키를 고를 확률. 각 그릇을 고를 확률이 동일하고 그릇에 동일한 쿠키가 들어있으므로 어떤 그릇을 택하든 바닐라 쿠키를 고를 확률도 같음.
  이를 어떤 가설에든 포함되는 데이터의 비율, 즉 한정상수(증거, Evidence)

- $P(H|D)$: 바닐라 쿠키가 그릇1에서 나왔을 확률. 우리가 알고 싶은 확률. 이를 데이터를 확인한 이후의 가설 확률, 즉 사후확률

◆ Bayes' theorem & Conditional Probability

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{\frac{1}{2} * \frac{3}{4}}{\frac{5}{8}} = \frac{3}{5}$$

- **왜 이렇게 복잡하게 사후확률을 구하는걸까?**
  실제로 사후확률은 구하기 어려운 경우가 많음. 그에 반해 사전확률은 우리가 이미 알고 있는 값이고, 가능도는 비교적 계산하기 수월. 그래서 Bayes Rule과 Likelihood를 활용해 Posterior을 도출해 내는 것

- 다시 말해, 우리가 알고 싶은 확률을 단박에 계산하기가 까다로울 때 조건과 결과를 뒤집어서 우회적으로 계산하는 것. 이것이 베이즈 모델의 강점

References

◆ Reference

- Ratsgo's blog

Q & A