

6.6 베이즈 정리

베이즈 정리는 데이터를 조건부 확률을 구하는 공식이다. 베이즈 정리를 사용하면 데이터가 주어지기 전의 사전 확률과 데이터가 주어진 이후 어떻게 바뀔지 예상할 수 있다. 따라서 데이터가 주어지기 전에 이미 어느정도 확률값을 예측하고 있을 때, 이를 사전 수집한 데이터와 합쳐서 최종 결과에 반영할 수 있다. 데이터의 갯수가 부족한 경우 아주 유용하다. 데이터를 매일 추가적으로 얻는 상황에서도 매일 전체 데이터를 대상으로 사전 분석작업을 할 필요 없이, 이제 분석결과가 오늘 들어온 데이터를 합쳐서 업데이트하면 그때 유용하게 활용할 수 있다.

베이즈 정리

조건부 확률을 구하는 다음 공식은 베이즈 정리 (Bayesian Rule)라고 한다.

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

(증명)

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

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

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

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

$P(A)$ 는 사전 확률(prior)이라고 하며, 사건 B 가 발생하기 전에 가지고 있던 A 의 확률이다. 만약 사건 B 가 발생하면 이 정보를 반영하여 사건 A 확률은 $P(A|B)$ 라는 값으로 변하게 되며, 이를 사후 확률(posterior)라고 한다.

사후 확률은 사전 확률에 $\frac{P(B|A)}{P(B)}$ 를 곱하면 얻을 수 있다. 여기서 분모값 $P(B|A)$ 를 가능도(likelihood)라고 한다. 분모값 $P(B)$ 는 정규화 상수(normalizing constant) 혹은 증거(evidence)라고 한다.

[베이즈정리]

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

- $P(A|B)$ = posterior, 사건 B 가 발생한 후 갱신된 $P(A_{\text{new}})$
- $P(A)$ = prior, 사건 B 가 발생하기 전에 가지고 있던 사건 A 의 확률
- $P(B|A)$ = likelihood, 사건 A 가 발생했을 때의 사건 B 확률
- $P(B)$ = normalizing constant, or evidence, 확률값의 크기 조정

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

베이즈정리의 확장 1

만약 사건 A_i 가 서로 배타적이고 완전하다고 하자.

- 서로 배타적 (교집합이 없다)

$$A_i \cap A_j = \emptyset$$

- 완전 (합집합 = Ω)

$$A_1 \cup A_2 \dots = \Omega$$

전체확률의 법칙을 이용하여, 다음과 같이 베이즈정리를 확장할 수 있다.

$$\begin{aligned} P(A_1 | B) &= \frac{P(B|A_1) P(A_1)}{P(B)} \\ &= \frac{P(B|A_1) P(A_1)}{\sum_i P(A_i, B)} \\ &= \frac{P(B|A_1) P(A_1)}{\sum_i P(B|A_i) P(A_i)} \end{aligned}$$

이 식은 멀티-클래스 분류 (multi-class classification) 문제에서 베이즈정리가
이렇게 사용되기를 보여주는 수식이다. 멀티클래스 분류 문제는 여러 배타적이고 완전한
사건 중에서 가장 높은 점수를 가진 사건을 고르는 문제다.

예를 들어, B 라는 헌트를 주고 1번부터 4번까지의 보기 중 하나를 골라야 하는 수치선택 문제는
4개의 A_1, A_2, A_3, A_4 중 B 에 대한 조건부 확률이 가장 높은 사건을 고르는 것과
같다. 이 문제를 풀기 위해서는 위의 베이즈정리 확장을 사용하여, 각각의 조건부 확률값을
100%로 만든다.

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{(P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)}$$

normalizing constant

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{\text{normalizing constant}}$$

$$P(A_3|B) = \frac{P(B|A_3)P(A_3)}{\text{normalizing constant}}$$

$$P(A_4|B) = \frac{P(B|A_4)P(A_4)}{\text{normalizing constant}}$$

그리데 분모인 $\sum_i P(B|A_i)P(A_i)$ 수¹은 i 까지 비례하는 항상 증가하므로,
 A_1, A_2, A_3, A_4 중 B 에 대한 조건부 확률이 가장 높은 사건을 가르는 것과 예측이라면
 분자와 같은 비교하여 된다.

$$P(A_1|B) \propto P(B|A_1)P(A_1)$$

$$P(A_2|B) \propto P(B|A_2)P(A_2)$$

$$P(A_3|B) \propto P(B|A_3)P(A_3)$$

$$P(A_4|B) \propto P(B|A_4)P(A_4)$$

확률론적 분류기법의 장단점

- (Pros) 1위가 사용할 수 있을 때, 2위나 3위 등을 대안적으로 사용 가능함
- (cons) 모든 사건 (A_1, A_2, \dots, A_N)의 조건부 확률을 모두 계산해야 함

$A_1 = A$, $A_2 = A^c$ 인 경우 다음과 같은 식이 성립한다. (맞나, 틀려나?)

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B,A) + P(B,A^c)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)(1-P(A))} \end{aligned}$$

⇒ 클래스 2개인 이진 분류 문제에 사용

⇒ $P(A|B) > 0.5 \rightarrow A \text{ else } A^c$

검사시약문제

: 이진 분류 문제의 예

비아즈 정리를 이용하여 다음과 같은 문제를 풀어보자.

- 특정한 병에 걸리지 확인하는 시약

→ 환자에게 테스트한 결과 99% 확률로 양성 반응

→ 병에 걸렸는지 확인하지 않은 어떤 환자가 이 시약을 테스트한 결과,
양성 반응을 보였다면 이 환자가 그 병에 걸려있을 확률은 얼마인가?

* 확률론의 용어로 다시 정리하여 서술

- 환자가 실제로 병에 걸림 = 사건 D
- 환자가 실제로 병에 걸리지 않음 = 사건 D^c
- 시약 테스트 양성 = 사건 S
- 시약 테스트 음성 = 사건 S^c

(주어진 확률값)

병에 걸린 환자에게 시약을 테스트했을 때 양성반응을 보이는 확률 = 99%

~'병에 걸렸다' = 추가정보/조건

$$\therefore P(S|D) = 0.99$$

구해야 하는 것은 이것과 반대로 양성반응을 보이는 환자가 병에 걸려있을 확률.

$$\rightarrow P(D|S)$$

정리하면,

사건

- 병에 걸리는 경우 : 사건 D
- 양성반응을 보이는 경우 : 사건 S
- 병에 걸린 사람이 양성반응을 보이는 경우 : 조건부 사건 $S|D$
- 양성반응을 보이는 사람이 병에 걸려 있을 경우 : 조건부 사건 $D|S$

문제

$$- P(S|D) = 0.99 \text{ 가 주어졌을 때, } P(D|S) \text{ 를 구하라.}$$

베이즈 정리에서

$$P(D|S) = \frac{0.99}{P(S|D)P(D)} \cdot P(S)$$

임D를 알고 있다. 그러나 이식에서 우리가 알고 있는 것은 $P(S|D)$ 뿐이고, $P(D)$ 나 $P(S)$ 는 모르기 때문에 $P(D|S)$ 는 현재로서는 구할 수 없다. 즉, 99%라고 간단히 말할 수 없다.

추가 조사를 통해 정보를 다음과 같이 입수했다고 하자.

- 이 병은 전체 인구 중 걸린 사람이 0.2%인 환자병이다.
- 이 병에 걸리지 않은 사람에게 시약검사를 했을 때, 양성반응, 즉 잘못된 결과 (False Positive)가 나타난 확률이 5%다.

↓
확률론적 표현

$$\begin{cases} P(D) = 0.002 \\ P(S|D^c) = 0.05 \end{cases}$$

이 문제는 피검사자가 병에 걸렸는지, 걸리지 않았는지를 알아보는 이진 분류 문제이므로, 이에 해당하는 베이즈 정리의 확장을 사용하면 다음과 같이 확률을 구할 수 있다.

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

$$= \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^c)P(D^c)}$$

$$= \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^c)(1 - P(D))}$$

양성인 사람이 실제 병에 걸렸을 확률

$$\frac{0.99 \cdot 0.002}{0.99 \cdot 0.002 + 0.05 \cdot (1 - 0.002)}$$

$$= 0.038$$

비이즈 정리의 확장 2

현실 문제에서 베이스 잡리는 단순한 형태로 표현되지 않음
- 여러개의 확률변수!

<updated probability of event A by event BC>

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

위 식에서 $P(A|B, C)$ 는 B 와 C 가 조건인 A 의 확률이다. 즉 $P(A|(B \cap C))$ 의 확률이다. 이 공식을 사건 A 와 C 만 있는 경우에 비교하면, 위 공식을 수립해보게 될 것이다.

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$

(2) 명)

$$P(A, B, C) = P(A|B, C)P(B, C) = P(A|B, C)P(C|B)P(B)$$

$$P(A, B, C) = P(C|A, B)P(A, B) = P(C|A, B)P(A|B)P(B)$$

$$P(A|B, C)P(C|B)P(B) = P(C|A, B)P(A|B)P(B)$$

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

연습문제 6.6.1

다음 식을 증명하라.

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

$$P(A,B,C) = P(A|B,C)P(B|C)P(C)$$

$$P(A,B,C) = P(B|A,C)P(A|C)P(C)$$

$$\therefore P(A|B,C)P(B|C) = P(B|A,C)P(A|C) \cdots ①$$

증명할 때 $P(B|C)$ 은 ~~제각각으로 구하기~~로 구하기,

$$P(A|B,C)P(B|C) = P(B|A,C)P(A|C) \cdots ②$$

$$\textcircled{1} = \textcircled{2}$$

연습문제 6.6.2

다음 식을 증명하라.

$$P(A|B,C,D) = \frac{P(D|A,B,C)P(A|B,C)}{P(D|B,C)} \cdots ①$$

$$P(A,B,C,D) = P(D|A,B,C)P(A|B,C)P(B|C)P(C)$$

$$\begin{aligned} & P(D|A,B,C)P(A|B,C)P(B|C)P(C) \\ & = P(D|A,B,C)P(A|B,C)P(B|C)P(C) \cdots ② \end{aligned}$$

$$\textcircled{1} = \textcircled{2}$$

* Bayes 대현재

조건절에 들어간 과정을 하는 대형은 $P(A|B)$

$$\textcircled{②} \quad P(A|B) = \frac{P(C|A)P(A|B)}{P(C|B)}$$

↓
B, HPPD

simple form

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \rightarrow \text{제법!}$$

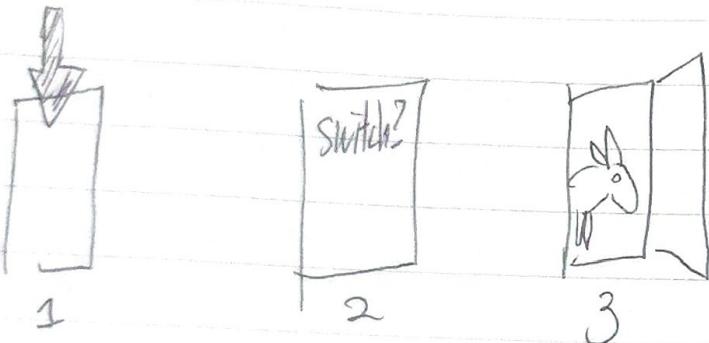
... 복잡한식의 몇몇 그룹을 제거하는 좋은 방법

몬티홀 문제

몬티홀 문제 (Monty Hall Problem)는 다음과 같은 확률문제다.

- 세 개의 문 중에 하나를 선택하여, 문 뒤에 있는 선물을 가지는 게임
- 그게 뒤에 차운 자동차, 나머지 2개에는 염소
- 어떤 사람이 1번 문선택 > 진행자 3번 문 열고 문 뒤에 염소 보여줌
> 1번 대신 2번으로 바꿔겠느냐고 ① 질문

→ 문을 바꿀 것이 자동차 확률에 유리한가?



문의 위치를 0, 1, 2라고 표현하면 다음과 같은 확률변수를 사용하여 이 문제를 풀 수 있다.

1. 자동차가 있는 문을 나타내는 확률변수 C 로, 값은 0, 1, 2를 가질 수 있다.
2. 참가자가 선택한 문을 나타내는 확률변수 X 로, 값은 0, 1, 2를 가질 수 있다.
3. 진행자가 열어줄 문을 나타내는 확률변수 H 로 값은 0, 1, 2를 가질 수 있다.

이 문제는 참가자와 진행자(행위) 조건으로, 자동차의 위치를 결과로 하는 조건부 확률을 두는 문제다. 예를 들어 참가자가 1번 문을 선택하고, 진행자가 2번 문을 열어서 자동차가 없다는 것을 보았으면 조건은 X_1, H_2 가 된다.

X_1, H_2

이 때 자동차는 0번 문 아니면 1번 문 뒤에 있으므로 이런 부류 문제가 된다.
이 문제를 푸는 핵심은 다음 두 가지 사실을 이용하는 것이다.

(1) 자동차를 놓는 진행자는 참가자의 선택을 예측할 수 없고, 참가자는 자동차를 볼 수 있으므로 자동차의 위치와 참가자의 선택은 서로 독립적이다.

$$P(C, X) = P(C) P(X)$$

(2) 진행자가 어떤 문제를 여는가가 자동차의 위치 및 참가자의 선택에 좌우된다. 예를 들어 자동차가 0번 문 뒤에 있고 참가자가 1번 문을 선택하면 진행자는 2번 문을 열 수 밖에 없다.

$$P(H_0 | C_0, X_1) = 0$$

$$P(H_1 | C_0, X_1) = 0$$

$$P(H_2 | C_0, X_1) = 1$$

자동차가 1번 문 뒤에 있는데 참가자가 1번 문을 선택한 경우에는 0번 문과 2번 문을 둘다 열어도 된다. 따라서 진행자가 0번 문이나 2번 문을 열 확률은 0.5다.

$$P(H_0 | C_1, X_1) = \frac{1}{2}$$

$$P(H_1 | C_1, X_1) = 0$$

$$P(H_2 | C_1, X_1) = \frac{1}{2}$$

사실들을 이용하면 참가자가 1번문을 선택하고 진행자가 2번문을 열어서 자동차가
得不到 것을 보면 경우에 1번문 뒤에 차가 있을 확률은 다음처럼 계산할 수 있다.

$$P(C_0 | X_1, H_2) = \frac{P(C_0, X_1, H_2)}{P(X_1, H_2)} = 1.0 \text{ (우리진!)}$$

$$= \frac{P(H_2 | C_0, X_1) P(C_0, X_1)}{P(X_1, H_2)} \quad \text{독립} \\ \rightarrow \frac{P(C_0) P(X_1)}{P(H_2 | X_1) P(X_1)}$$

$$= \frac{P(C_0)}{P(H_2 | X_1)}$$

전체확률의 법칙

$$\rightarrow \frac{P(C_0)}{P(H_2, C_0 | X_1) + P(H_2, C_1 | X_1) + P(H_2, C_2 | X_1)}$$

$$(A, B, C)$$

$$P(A|B, C) P(B)$$

$$\text{when } B, C \text{ 독립}$$

$$\left. \begin{array}{l} \text{연습문제} \\ \{ 0.5-4 중 1 \end{array} \right.$$

$$= \frac{P(C_0)}{P(H_2 | X_1, C_0) P(C_0) + P(H_2 | X_1, C_1) P(C_1) + P(H_2 | X_1, C_2) P(C_2)}$$

"우리가 선택한 경우는 2가지中有 1가지 correct."

$$= \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}}$$

$$= \frac{2}{3} \quad \therefore P(C_0 | X_1, H_2) = \frac{2}{3}$$

이진문제 모제이므로, $P(C_1 | X_1, H_2) = 1 - P(C_0 | X_1, H_2) = \frac{1}{3}$.