6.6 베이즈 정리

베이즈 정리는 데이터라는 조건이 주어졌을 때의 조건부확률을 구하는 공식이다. 베이즈 정리를 쓰면 데이터가 주어지기 전의 사전확률값이 데이터가 주어지면서 어떻게 변하는지 계산할 수 있다. 따라서 데이터가 주어지기 전에 이미 어느 정도 확률값을 예측하고 있을 때 이를 새로 수집한 데이터와 합쳐서 최종 결과에 반영할 수 있다. 데이터의 개수가 부족한 경우 아주 유용하다. 데이터를 매일 추가적으로 얻는 상황에서도 매일 전체 데이터를 대상으로 새로 분석작업을 할 필요없이 어제 분석결과에 오늘 들어온 데이터를 합쳐서 업데이트만 하면 되므로 유용하게 활용할 수 있다.

베이즈 정리

조건부확률을 구하는 다음 공식을 **베이즈 정리(Bayesian rule)**라고 한다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6.6.1}$$

(증명)

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \rightarrow \quad P(A,B) = P(A|B)P(B) \tag{6.6.2}$$

$$P(B|A) = \frac{P(A,B)}{P(A)} \quad \rightarrow \quad P(A,B) = P(B|A)P(A) \tag{6.6.3}$$

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \tag{6.6.4}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6.6.5}$$

여기에서 $P(A)$는 **사전확률(prior)**이라고 하며 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률이다. 만약 사건 B가 발생하면 이 정보를 반영하여 사건 A의 확률은 $P(A|B)$라는 값으로 변하게 되며 이를 **사후확률(posterior)**이라고 한다.

사후확률값은 사전확률에 $\frac{P(B|A)}{P(B)}$라는 값을 곱하면 얻을 수 있다. 곱하는 $P(B|A)$는 **가능도(likelihood)**라고 하고 나누는 $P(B)$는 **정규화 상수(normalizing constant)** 혹은 **증거(evidence)**라고 한다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6.6.6}$$

- $P(A|B)$: 사후확률(posterior). 사건 B가 발생한 후 갱신된 사건 A의 확률
- $P(A)$: 사전확률(prior). 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
- $P(B|A)$: 가능도(likelihood). 사건 A가 발생한 경우 사건 B의 확률
- $P(B)$: 정규화 상수(normalizing constant) 또는 증거(evidence). 확률의 크기 조정

베이즈 정리는 사건 B가 발생함으로써(사건 B가 진실이라는 것을 알게 됨으로써, 즉 사건 B의 확률 $P(B) = 1$이라는 것을 알게 됨으로써) 사건 A의 확률이 어떻게 변화하는지를 표현한 정리다. 따라서 베이즈 정리는 새로운 정보가 기존의 추론에 어떻게 영향을 미치는지를 나타낸다.

베이즈 정리의 확장 1

만약 사건 A_i가 서로 배타적이고 완전하다고 하자.

- 서로 배타적(교집합이 없다)

$$A_i \cap A_j = \emptyset \tag{6.6.7}$$

- 완전(합집합이 표본공간)

$$A_1 \cup A_2 \cup \cdots = \Omega \tag{6.6.8}$$

전체 확률의 법칙을 이용하여 다음과 같이 베이즈 정리를 확장할 수 있다.

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{\sum_i P(A_i, B)} \\ &= \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_i)P(A_i)} \end{aligned} \tag{6.6.9}$$

이 식은 멀티-클래스 분류(multi-class classification) 문제에서 베이즈 정리가 어떻게 사용되는지를 보여주는 수식이다. 멀티-클래스 분류 문제는 여러 배타적이고 완전한 사건 중에서 가장 확률이 높은 하나의 사건을 고르는 문제다. 예를 들어 B라는 힌트를 주고 1번부터 4번까지의 보기 중 하나를 골라야 하는 4지선다형 문제는 4개의 A_1, A_2, A_3, A_4 중 B에 대한 조건부 확률이 가장 높은 사건을 고르는 것과 같다. 이 문제를 풀기 위해서는 위의 베이즈 정리 확장을 사용하여 4개의 조건부 확률값을 비교하면 된다.

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \tag{6}$$

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \tag{6}$$

$$P(A_3|B) = \frac{P(B|A_3)P(A_3)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \tag{6}$$

$$P(A_4|B) = \frac{P(B|A_4)P(A_4)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \tag{6}$$

그런데 분모에 있는 $\sum_i P(B|A_i)P(A_i)$ 식은 i값이 바뀌어도 항상 같은 값이므로 A_1, A_2, A_3, A_4 중 B에 대한 조건부 확률이 가장 높은 사건을 고르는 것이 목적이라면 분자의 값만 비교하면 된다. 다음 식에서 $\propto$ 기호는 비례한다는 뜻이다.

$$P(A_1|B) \propto P(B|A_1)P(A_1) \tag{6.6.14}$$

$$P(A_2|B) \propto P(B|A_2)P(A_2) \tag{6.6.15}$$

$$P(A_3|B) \propto P(B|A_3)P(A_3) \tag{6.6.16}$$

$$P(A_4|B) \propto P(B|A_4)P(A_4) \tag{6.6.17}$$

$A_1 = A, A_2 = A^C$ 인 경우에는 다음과 같은 식이 성립한다.

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B,A) + P(B,A^C)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)(1-P(A))} \end{aligned} \tag{6.6.18}$$

이 식은 클래스가 2개뿐인 이진 클래스 분류문제에 사용된다. $P(A|B)$의 값이 0.5보다 크면 답은 A이고 반대라면 답은 A^C이다.

검사 시약 문제

베이즈 정리를 이용하여 다음과 같은 문제를 풀어보자.

제약사에서 환자가 특정한 병에 걸린지 확인하는 시약을 만들었다. 그 병에 걸린 환자에게 시약을 테스트한 결과 99%의 확률로 양성 반응을 보였다. 병에 걸린지 확인이 되지 않은 어떤 환자가 이 시약을 테스트한 결과 양성 반응을 보였다면 이 환자가 그 병에 걸려 있을 확률은 얼마인가? 99%일까?

이 문제를 확률론의 용어로 다시 정리하여 서술해보자.

우선 환자가 실제로 병에 걸린 경우를 사건 D라고 하자. 그러면 병에 걸려있지 않은 경우는 사건 D^C 가 된다. 또 시약 테스트에서 양성 반응을 보이는 경우를 사건 S라고 하면 음성 반응을 보이는 경우는 사건 S^C 이다.

현재 주어진 확률값은 병에 걸린 환자에게 시약을 테스트했을 때 양성 반응을 보이는 확률이다. 병에 걸렸다는 것은 추가된 조건 혹은 정보이므로 이 확률은 $P(S|D)$로 표기할 수 있다.

그런데 구해야 하는 값은 이것과 반대로 양성 반응을 보이는 환자가 병에 걸려있을 확률이다. 이 때에는 양성 반응을 보인다라는 것이 추가된 정보이므로 이 확률은 $P(D|S)$로 표기할 수 있다.

- 사건
  - 병에 걸리는 경우 : 사건 D
  - 양성 반응을 보이는 경우 : 사건 S
  - 병에 걸린 사람이 양성 반응을 보이는 경우 : 조건부 사건 $S|D$
  - 양성 반응을 보이는 사람이 병에 걸려 있을 경우 : 조건부 사건 $D|S$
- 문제
  - $P(S|D) = 0.99$가 주어졌을 때, $P(D|S)$를 구하라.

베이즈 정리에서

$$P(D|S) = \frac{P(S|D)P(D)}{P(S)} \tag{6.6.19}$$

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

추가 조사를 통해 필요한 정보를 다음과 같이 입수했다고 하자.

- 이 병은 전체 인구 중 걸린 사람이 0.2%인 희귀병이다.
- 이 병에 걸리지 않은 사람에게 시약 검사를 했을 때, 양성 반응, 즉 잘못된 결과(False Positive)가 나타난 확률이 5%다.

이를 확률론적 용어로 바꾸면 다음과 같다.

$$P(D) = 0.002 \tag{6.6.20}$$
$$P(S|D^C) = 0.05 \tag{6.6.21}$$

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

$$\begin{aligned} P(D|S) &= \frac{P(S|D)P(D)}{P(S)} \\ &= \frac{P(S|D)P(D)}{P(S,D) + P(S,D^C)} \\ &= \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^C)P(D^C)} \\ &= \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^C)(1-P(D))} \\ &= \frac{0.99 \cdot 0.002}{0.99 \cdot 0.002 + 0.05 \cdot (1-0.002)} \\ &= 0.038 \end{aligned} \tag{6.6.22}$$

시약 반응에서 양성 반응을 보이는 사람이 실제로 병에 걸려 있을 확률은 약 3.8%에 불과하다.

피지엠파이를 사용한 베이즈 정리 적용

피지엠파이 패키지는 베이즈 정리에 적용하는 BayesianModel 클래스를 제공한다. 베이즈 정리를 적용하려면 조건부확률을 구현하는 TabularCPD 클래스를 사용하여 사전확률과 가능도를 구현해야 한다. TabularCPD 클래스 객체는 다음과 같이 만든다.

```
TabularCPD(variable, variable_card, value, evidence=None, evidence_card=None)
```

- variable : 확률변수의 이름 문자열
- variable_card : 확률변수가 가질 수 있는 경우의 수
- value : 조건부확률 배열. 하나의 열(column)이 동일 조건을 뜻하므로 하나의 열의 확률 합은 1이어야 한다.
- evidence : 조건이 되는 확률변수의 이름 문자열의 리스트
- evidence_card : 조건이 되는 확률변수가 가질 수 있는 경우의 수의 리스트

TabularCPD 클래스는 원래는 조건부확률을 구현하기 위한 것이지만 evidence=None , evidence_card=None 으로 인수를 주면 일반적인 확률도 구현할 수 있다.

우선 확률변수 X를 이용하여 병에 걸렸을 사전확률 $P(D) = P(X = 1)$, 병에 걸리지 않았을 사전확률 $P(D^C) = P(X = 0)$를 정의한다.

In [1]:

```
from pgmpy.factors.discrete import TabularCPD

cpd_X = TabularCPD('X', 2, [[1 - 0.002, 0.002]])
print(cpd_X)
```

```
+------+-------+
| X(0) | 0.998 |
+------+-------+
| X(1) | 0.002 |
+------+-------+
```

다음으로는 양성 반응이 나올 확률 $P(S) = P(Y = 1)$, 음성 반응이 나올 확률 $P(S^C) = P(Y = 0)$를 나타내는 확률변수 Y를 정의한다.

확률변수 Y의 확률을 베이즈 모형에 넣을 때는 TabularCPD 클래스를 사용한 조건부확률 $P(Y|X)$의 형태로 넣어야 하므로 다음처럼 조건부확률 $P(Y|X)$를 구현한다.

In [2]:

```
cpd_Y_on_X = TabularCPD('Y', 2, np.array([[0.95, 0.01], [0.05, 0.99]]),
                        evidence=['X'], evidence_card=[2])
print(cpd_Y_on_X)
```

```
+------+------+------+
| X    | X(0) | X(1) |
+------+------+------+
| Y(0) | 0.95 | 0.01 |
+------+------+------+
| Y(1) | 0.05 | 0.99 |
+------+------+------+
```

이제 이 확률변수들이 어떻게 결합되어 있는지는 나타내는 확률모형인 `BayesianModel` 클래스 객체를 만들어야 한다.

```
BayesianModel(variables)
```

- `variables` : 확률모형이 포함하는 확률변수 이름 문자열의 리스트

`BayesianModel` 클래스는 다음 메서드를 지원한다.

- `add_cpds()` : 조건부확률을 추가
- `check_model()` : 모형이 정상적인지 확인. `True` 면 정상적인 모형

In [3]:

```
from pgmpy.models import BayesianModel

model = BayesianModel([('X', 'Y')])
model.add_cpds(cpd_X, cpd_Y_on_X)
model.check_model()
```

Out[3]:

True

`BayesianModel` 클래스는 변수 제거법(VariableElimination)을 사용한 추정을 제공한다.
`VariableElimination` 클래스로 추정(inference) 객체를 만들고 이 객체의 `query()` 메서드를 사용하면 사후확률을 계산한다.

```
query(variables, evidences)
```

- `variables` : 사후확률을 계산할 확률변수의 이름 리스트
- `evidences` : 조건이 되는 확률변수의 값을 나타내는 딕셔너리

여기에서는 pgmpy 패키지를 이용하여 베이즈 정리를 적용할 수 있다는 것만 알면 된다. 자세한 내용은 추후 **확률적 그래프 모형(Probabilistic Graphical Model)**에서 다룬다.

In [4]:

```
from pgmpy.inference import VariableElimination

inference = VariableElimination(model)
posterior = inference.query(['X'], evidence={'Y': 1}, joint=False, show_progress=False)
print(posterior['X'])
```

```
+------+----------+
| X    |   phi(X) |
+======+==========+
| X(0) |   0.9618 |
+------+----------+
| X(1) |   0.0382 |
+------+----------+
```

베이즈 정리의 확장 2

베이즈 정리는 사건 A의 확률이 사건 B에 의해 갱신(update)된 확률을 계산한다. 그런데 만약 이 상태에서 또 추가적인 사건 C가 발생했다면 베이즈 정리는 다음과 같이 쓸 수 있다.

$$P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|B)} \tag{6.6.23}$$

위 식에서 $P(A|B,C)$는 B와 C가 조건인 A의 확률이다. 즉 $P(A|(B \cap C))$를 뜻한다.

이 공식을 사건 A와 C 만 있는 경우와 비교하면 위 공식을 쉽게 외울 수 있다.

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \tag{6.6.24}$$

(증명)

$$P(A,B,C) = P(A|B,C)P(B,C) = P(A|B,C)P(C|B)P(B) \tag{6.6.25}$$
$$P(A,B,C) = P(C|A,B)P(A,B) = P(C|A,B)P(A|B)P(B) \tag{6.6.26}$$
$$P(A|B,C)P(C|B)P(B) = P(C|A,B)P(A|B)P(B) \tag{6.6.27}$$
$$P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|B)} \tag{6.6.28}$$

연습 문제 6.6.1

다음 식을 증명하라.

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)} \tag{6.6.29}$$

연습 문제 6.6.2

다음 식을 증명하라.

$$P(A|B,C,D) = \frac{P(D|A,B,C)P(A|B,C)}{P(D|B,C)} \tag{6.6.30}$$

연습 문제 6.6.3

다음 식을 증명하라.

$$P(A,B|C,D) = \frac{P(D|A,B,C)P(A,B|C)}{P(D|C)} \tag{6.6.31}$$

몬티 홀 문제

몬티 홀 문제(Monty Hall problem)는 다음과 같은 확률문제다.

> 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가지는 게임쇼에 참가했다. 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 염소가 있다. 이때 어떤 사람이 예를 들어 1번 문을 선택했을 때, 게임쇼 진행자는 3번 문을 열어 문뒤에 염소가 있음을 보여주면서 1번 대신 2번을 선택하겠냐고 물었다. 참가자가 자동차를 가지려할 때 원래 선택했던 번호를 바꾸는 것이 유리할까?

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

1. 자동차가 있는 문을 나타내는 확률변수 C로 값은 0, 1, 2를 가질 수 있다.
2. 참가자가 선택한 문을 나타내는 확률변수 X로 값은 0, 1, 2를 가질 수 있다.
3. 진행자가 열어준 문을 나타내는 확률변수 H로 값은 0, 1, 2를 가질 수 있다.

이 문제는 참가자와 진행자의 행위를 조건으로, 자동차의 위치를 결과로 하는 조건부 확률을 푸는 문제다. 예를 들어 참가자가 1번 문을 선택하고 진행자가 2번 문을 열어서 자동차가 없다는 것을 보였으면 조건은

$$X_1, H_2 \tag{6.6.32}$$

가 된다. 이때 자동차는 0번 문 아니면 1번 문 뒤에 있으므로 이진 분류 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 없으므로 자동차의 위치와 참가자의 선택은 서로 독립적이다.

$$P(C, X) = P(C)P(X) \tag{6.6.33}$$

(2) 진행자가 어떤 문을 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.

$$P(H_0|C_0, X_1) = 0 \tag{6.6.34}$$
$$P(H_1|C_0, X_1) = 0 \tag{6.6.35}$$
$$P(H_2|C_0, X_1) = 1 \tag{6.6.36}$$

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

$$P(H_0|C_1, X_1) = \frac{1}{2} \tag{6.6.37}$$
$$P(H_1|C_1, X_1) = 0 \tag{6.6.38}$$
$$P(H_2|C_1, X_1) = \frac{1}{2} \tag{6.6.39}$$

이 사실들을 이용하면 참가자가 1번 문을 선택하고 진행자가 2번 문을 열어서 자동차가 없다는 것을 보인 경우에 0번 문 뒤에 차가 있을 확률은 다음처럼 계산할 수 있다.

$$\begin{aligned}
P(C_0 \mid X_1, H_2) &= \frac{P(C_0, X_1, H_2)}{P(X_1, H_2)} \\
&= \frac{P(H_2|C_0, X_1)P(C_0, X_1)}{P(X_1, H_2)} \\
&= \frac{P(C_0)P(X_1)}{P(H_2|X_1)P(X_1)} \\
&= \frac{P(C_0)}{P(H_2|X_1)} \\
&= \frac{P(C_0)}{P(H_2, C_0|X_1) + P(H_2, C_1|X_1) + P(H_2, C_2|X_1)} \\
&= \frac{P(C_0)}{P(H_2|X_1, C_0)P(C_0) + P(H_2|X_1, C_1)P(C_1) + P(H_2|X_1, C_2)P(C_2)} \\
&= \frac{\frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} \\
&= \frac{2}{3}
\end{aligned}$$

$$P(C_0 \mid X_1, H_2) = \frac{2}{3} \tag{6.6.41}$$

이진 분류 문제이므로

$$P(C_1 \mid X_1, H_2) = 1 - P(C_0 \mid X_1, H_2) = \frac{1}{3} \tag{6.6.42}$$

따라서 0번 문 뒤에 자동차가 있을 확률은 1번 문 뒤에 자동차가 있을 확률의 2배이다. 참가자는 선택을 바꾸는 것이 유리하다.