



# Naïve Bayes

---

Sang Yup Lee



# Naïve Bayes

---

- 분류 문제에 적용되는 알고리즘
- Naïve Bayes 모형은 다음과 같은 베이즈 공식을 사용한 방법으로 이를 이해하기 위해서는 기본적인 확률에 대해서 알아야 함
  - Bayes' Rule

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



# Naïve Bayes

---

- 새로운 관측치에 대해서,
  - 해당 관측치의 종속변수가 특정한 값을 갖을 확률을 베이즈 공식을 이용해서 구한다.
  - 그리고, 확률이 제일 큰 값으로 종속변수의 값을 예측한다.
  - Example
    - $i$ 번째 관측치에 대해서,  $y_i \in \{0,1\}$  인 경우
    - If  $P(y_i = 0) > P(y_i = 1)$ , then  $y_i$  값은 0으로 예측, 그렇지 않으면 1로 예측
  - 별도의 비용함수가 존재하지 않는다.

# Naïve Bayes

i 번째 관측치가 갖는 feature 들의 값

- 종속변수 (Y)가 취할 수 있는 값이 0, 1 두개인 경우, 각 값을 취할 확률은 다음과 같이 표현
  - $P(Y_i = 1 | X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$ 
    - $X_1, X_2, \dots, X_k$  는 데이터에 존재하는 Feature 들
    - 이는 종속변수의 값이 1일 확률이 Feature 들의 구체적인 값에 따라서 달라진다는 것을 의미
  - Example
    - 종속변수: 폐암여부
    - 독립변수: 연령( $X_1$ ), 흡연여부( $X_2$ )
    - 특정 사람에 대해서,  $X_1=60, X_2=1$  인 경우,
    - 우리가 궁금한 것은, whether  $P(Y_i = 1 | X_1 = 60, X_2 = 1) > P(Y_i = 0 | X_1 = 60, X_2 = 1)$



# Naïve Bayes

---

- 설명을 위해 독립변수의 수 = 2 이라고 가정 (즉, k=2)
  - $P(Y_i = 1|X_1 = x_1, X_2 = x_2)$  구하기
    - 베이즈 공식을 사용하면 다음과 같이 표현

$$P(Y_i = 1|X_1 = x_1, X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2 | Y_i = 1)P(Y_i = 1)}{P(X_1 = x_1, X_2 = x_2)}$$

- $P(Y_i = 0|X_1 = x_1, X_2 = x_2)$  구하기
  - 베이즈 공식을 사용하면 다음과 같이 표현

$$P(Y_i = 0|X_1 = x_1, X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2 | Y_i = 0)P(Y_i = 0)}{P(X_1 = x_1, X_2 = x_2)}$$

- 둘중 어느것인 더 큰가를 판단  $\Rightarrow$  해당 값으로 예측



# Naïve Bayes

---

- 종속변수 값의 예측
  - 확률이 높은 값으로 예측
  - 즉,  $P(Y_i = 1|X_1 = x_1, X_2 = x_2)$ 와  $P(Y_i = 0|X_1 = x_1, X_2 = x_2)$ 의 대소 비교
  - 구체적인 값은 중요하지 않다.
  - 베이즈 공식에서의 분모값은 계산하지 않는다.



# Naïve Bayes

---

- 분자의 값 구하기
- $P(Y_i = 1 | X_1 = x_1, X_2 = x_2)$  의 경우

$$P(Y_i = 1 | X_1 = x_1, X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2 | Y_i = 1)P(Y_i = 1)}{P(X_1 = x_1, X_2 = x_2)}$$

- 여기에서 각 Feature 들(즉,  $X_1, X_2$ ) 은 서로 independent하다고 가정

$$P(X_1 = x_1, X_2 = x_2 | Y_i = 1) = P(X_1 = x_1 | Y_i = 1)P(X_2 = x_2 | Y_i = 1)$$

- $P(X_j = x_j | Y_i = 1)$ 와  $P(Y_i = 1)$ 는 학습데이터를 사용해서 구함



# Naïve Bayes

---

- 분자의 값 구하기 (cont'd)
  - 흡연여부 예제
    - $P(Y_i = 1 | X_1 = 60, X_2 = 1)$

$$P(Y_i = 1 | X_1 = 60, X_2 = 1) = \frac{P(X_1 = 60, X_2 = 1 | Y_i = 1)P(Y_i = 1)}{P(X_1 = 60, X_2 = 1)}$$

- 여기에서

$$P(X_1 = 60, X_2 = 1 | Y_i = 1) = P(X_1 = 60 | Y_i = 1)P(X_2 = 1 | Y_i = 1)$$

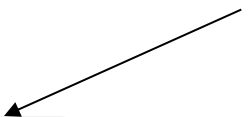




# Naïve Bayes

## ■ Example (학습데이터)

종속변수: 골프 플레이 여부



| Humidity | Windy | Play |
|----------|-------|------|
| high     | false | NO   |
| high     | true  | NO   |
| high     | false | YES  |
| high     | false | YES  |
| normal   | false | YES  |
| normal   | true  | NO   |
| normal   | true  | YES  |
| high     | false | NO   |
| normal   | false | YES  |
| normal   | false | YES  |
| normal   | true  | YES  |
| high     | true  | YES  |
| normal   | false | YES  |
| high     | true  | NO   |



# Naïve Bayes

---

- Example (cont'd)
  - Humidity=normal, Windy=true인 경우, 종속변수는 무엇으로 예측이 되는가?
    - 이를 위해서는 아래 값들을 비교하는 것이 필요

$P(\text{Play}=\text{Yes}|\text{Humidity}=\text{normal}, \text{Windy}=\text{true})$  vs.  $P(\text{Play}=\text{No}|\text{Humidity}=\text{normal}, \text{Windy}=\text{true})$



# Naïve Bayes

---

- Example (cont'd)
  - 각 확률의 계산

$$\begin{aligned} & P(\text{Play}=\text{Yes}|\text{Humidity}=\text{normal}, \text{Windy}=\text{true}) \\ &= \frac{P(\text{Humidity}=\text{normal}, \text{Windy}=\text{true}|\text{Play}=\text{Yes})P(\text{Play}=\text{Yes})}{P(\text{Humidity}=\text{normal}, \text{Windy}=\text{true})} \end{aligned}$$

**vs.**

$$\begin{aligned} & P(\text{Play}=\text{No}|\text{Humidity}=\text{normal}, \text{Windy}=\text{true}) \\ &= \frac{P(\text{Humidity}=\text{normal}, \text{Windy}=\text{true}|\text{Play}=\text{No})P(\text{Play}=\text{No})}{P(\text{Humidity}=\text{normal}, \text{Windy}=\text{true})} \end{aligned}$$



# Naïve Bayes

---

- Example (cont'd)

- 분자의 첫번째 항

$$P(\text{Humidity}=\text{normal}, \text{Windy}=\text{true}|\text{Play}=\text{Yes})$$

- 이는 각 feature들이 독립이기 때문에 다음과 같이 표현

$$P(\text{Humidity}=\text{normal}|\text{Play}=\text{Yes})P(\text{Windy}=\text{true}|\text{Play}=\text{Yes})$$



# Naïve Bayes

---

- Example (cont'd)
  - $P(\text{Humidity}=\text{normal}|\text{Play}=\text{Yes}) = \frac{\#\text{normal}\&\text{Yes}}{\#\text{Yes}} = 6/9$
  - $P(\text{Windy}=\text{true}|\text{Play}=\text{Yes}) = \frac{\#\text{true}\&\text{Yes}}{\#\text{Yes}} = 3/9$
  - $P(\text{Play}=\text{Yes}) = \frac{\#\text{Yes}}{\#\text{Yes} + \#\text{No}} = 9/14$

# Naïve Bayes

- Continuous features  
(독립변수가 연속변수인 경우)
  - 1) 독립변수가 취하는 값을  
기준으로 몇 개의 그룹으로 구분  
=> 범주형변수로 취급
    - 예)  $X \in [0,100]$
  - 2) 특정 확률분포를 사용하여  
확률을 계산
    - 예) 정규분포

| Old X             | New X |
|-------------------|-------|
| $X \leq 25$       | 0     |
| $25 < X \leq 50$  | 1     |
| $50 < X \leq 75$  | 2     |
| $75 < X \leq 100$ | 3     |



# Naïve Bayes

---

- in Python
  - For categorical features
    - Categorical Naïve Bayes 이용
      - CategoricalNB
      - [https://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.CategoricalNB.html#sklearn.naive\\_bayes.CategoricalNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.CategoricalNB.html#sklearn.naive_bayes.CategoricalNB)
      - See "Naive\_Bayes\_example.ipynb"
      - [https://rstudio-pubs-static.s3.amazonaws.com/118220\\_5a7997d6b0aa493c878d661968fc1f08.html](https://rstudio-pubs-static.s3.amazonaws.com/118220_5a7997d6b0aa493c878d661968fc1f08.html)
    - For continuous features
      - GaussianNB 사용 (정규분포 사용)
      - [https://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.GaussianNB.html#sklearn.naive\\_bayes.GaussianNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB)