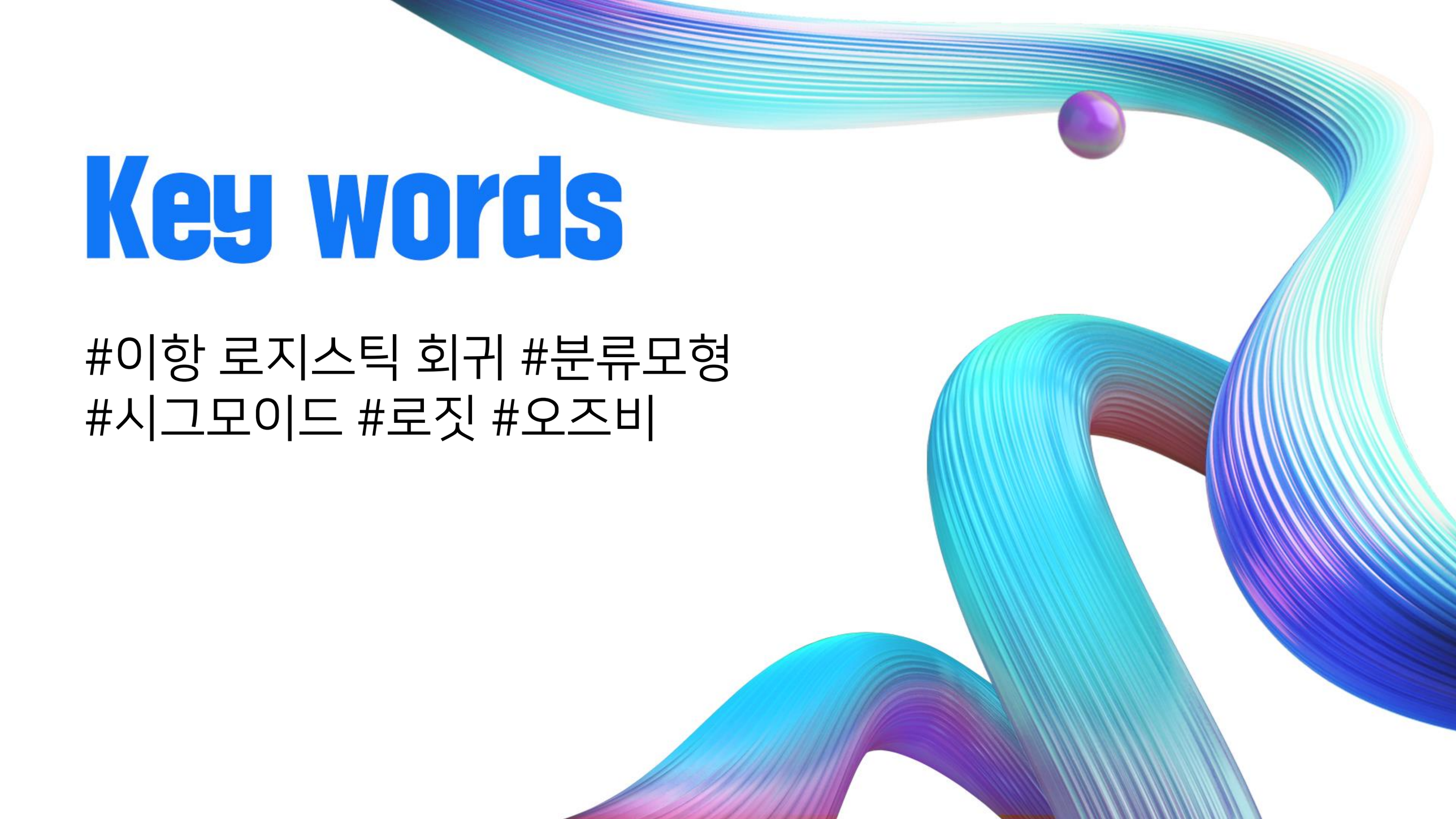
분류: 로지스틱 회귀 (Logistic Regression)

Key words

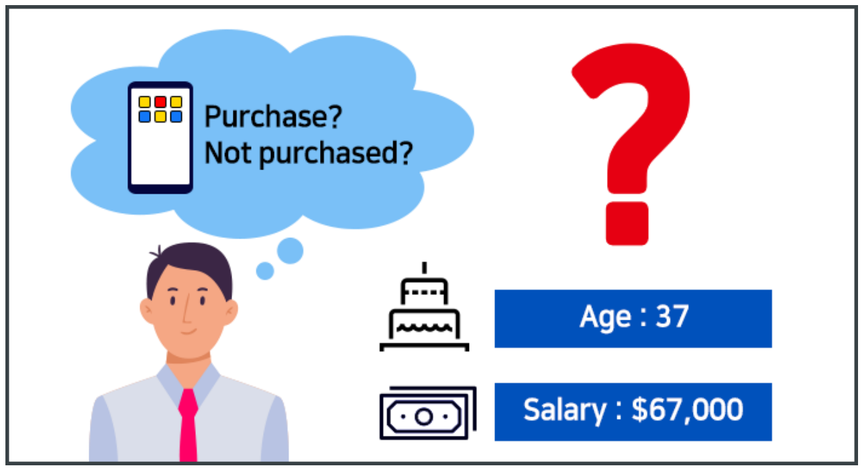
#이항 로지스틱 회귀 #분류모형
#시그모이드 #로짓 #오즈비



로지스틱 회귀모형 개요

로지스틱 회귀모형

- 로지스틱 회귀분석은 선형 회귀분석과 달리
  반응변수가 범주형 데이터인 경우에 사용되는 기법.
- 새로운 설명변수의 값이 주어질 때 반응변수의 각 범주에 속할
  확률이 얼마인지를 추정하고, 추정 확률을 분류기준값에 따라
  분류하는 목적으로 사용됨.



이항 로지스틱 회귀모형

이항 로지스틱 회귀모형

- 이항 로지스틱 회귀모형 : 이진(0/1)형 값을 가지는 반응변수를 여러 설명변수를 이용하여 회귀식의 형태로 예측하는 모형.
- 반응변수 Y는 1 또는 0의 값을 가지는 이진변수,
  설명변수는 $x_1,\ldots, x_k$ 로 k개인 경우에,
  $p = P(Y = 1|x_1, \ldots, x_k)$ 라고 하면,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

이항 로지스틱 회귀모형

이항 로지스틱 회귀모형

- 위 모형은 다음과 같이 p에 관한 식으로 표현할 수 있음.

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}$$ → 시그모이드 함수

이항 로지스틱 회귀모형

이항 로지스틱 회귀모형

- 범주형 반응변수의 범주가 두 개일 때, 관심 범주를 1, 다른 범주를 0으로 정의하면, 반응변수 Y는 관심범주에 속할 확률이 p인 베르누이 확률분포를 따르는 것으로 볼 수 있음.

$$\Pr(Y = 1) = p \ , \qquad \Pr(Y = 0) = 1 - p$$

- 여기서 확률 p를 독립변수의 함수로 설명하고자 함.
- 확률 p은 0과 1사이의 값이므로, $(-\infty, \infty)$의 범위를 가지는 독립변수의 선형함수로 나타낼 수 없음.

이항 로지스틱 회귀모형

로지스틱 모형식의 이해

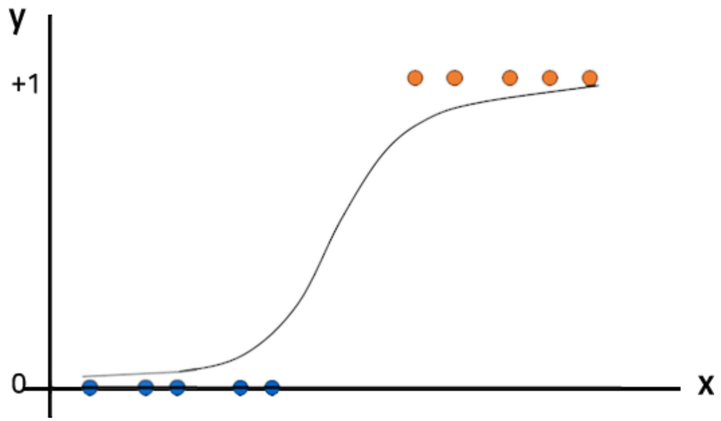
- 확률 p 대신 로그오즈($\log(\frac{p}{1-p})$)를 독립변수의 선형함수로 나타낸 것임.

$$logit(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$$

이항 로지스틱 회귀모형

로지스틱 모형식의 이해

- 설명변수가 1개인 경우 ($k = 1$), $p = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$ ($\beta_1 > 0$) 는 아래와 같은 비스듬한 S 곡선형태를 가짐.
  - p 는 언제나 0~1 사이의 값이 됨.
  - x 가 $-\infty$ 일 때, p 는 0.
  - x 가 ∞ 일 때, p 는 1.



이항 로지스틱 회귀모형

추정 및 예측

- (x_i, y_i) 의 표본 자료가 n개 주어지면 최대우도추정법,
  경사하강법 등을 이용하여 가장 적합한 곡선 함수 $(\widehat{\beta_0}, \widehat{\beta_1})$ 를 추정.
- 새로운 자료 x_{new} 가 주어졌을 때,

  $P(Y_{new} = 1)$는 $\hat{p}_{new} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{new})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{new})}$ 로 추정됨.

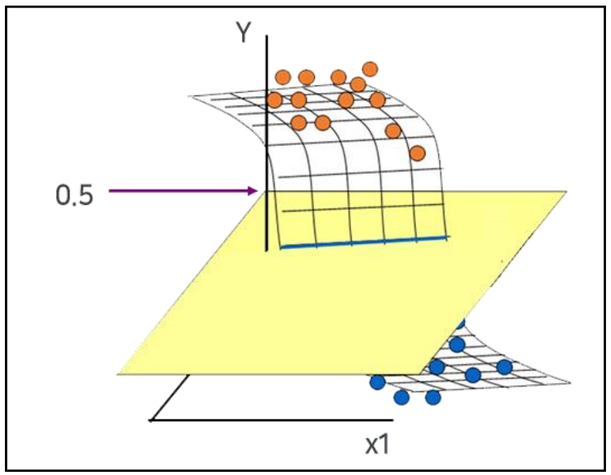
- $\hat{p}_{new} \geq threshold$ 면, $\hat{Y}_{new} = 1$
- $\hat{p}_{new} < threshold$ 면, $\hat{Y}_{new} = 0$

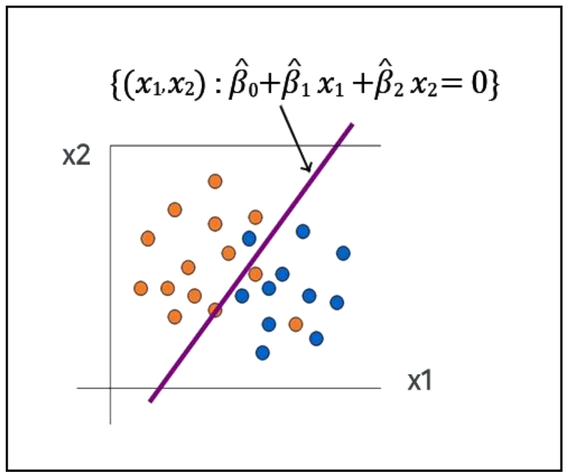
이항 로지스틱 회귀모형

로지스틱 회귀모형의 분리경계면

- 로지스틱 회귀모형은 선형의 결정경계를 가짐.
- 독립변수가 2개인 로지스틱 회귀모형과 threshold=0.5 일때의 초평면 (Hyperplane)

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$





이항 로지스틱 회귀모형

로지스틱 회귀와 오즈비(odds ratio)

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} \Leftrightarrow \quad \frac{p}{1-p}(\mathbf{odds}) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

- 나머지 변수는 모두 고정시킨 상태에서 한 변수 X_1 만 1만큼 증가시켰을 때, 변화하는 odds의 비율(오즈비)는 e^{β_1} 임을 알 수 있음.

$$\frac{odds(x_1+1, x_2)}{odds(x_1, x_2)} = \frac{e^{\beta_0+\beta_1(x_1+1)+\beta_2 x_2}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} = e^{\beta_1}$$

이항 로지스틱 회귀모형

로지스틱 회귀와 오즈비(odds ratio)

- x_1 만 1만큼 증가하면, 성공(관심범주, Y=1)에 대한 오즈가 $\exp(\beta_1)$배 변화함.

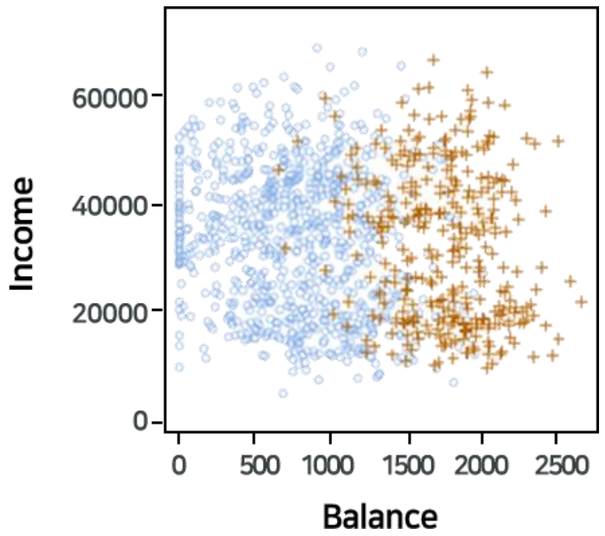
  - $\beta_1 > 0$: 관심범주에 속할 확률이 증가함.
    X_1 변수와 관심범주 간에는 양의 상관관계.

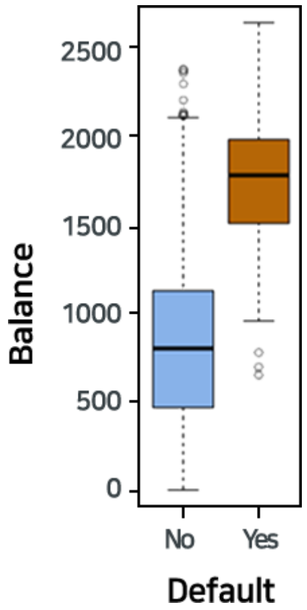
  - $\beta_1 < 0$: 관심범주에 속할 확률이 감소함.
    X_1 변수와 관심범주 간에는 음의 상관관계.

이항 로지스틱 회귀모형 사례

신용카드 연체에 관한 사례

- 어느 신용카드 회사의 고객의 잔액(balance)을 독립변수로 하여, 고객의 연체(Default) 가능성을 예측하고자 함.





이항 로지스틱 회귀모형 사례

신용카드 연체에 관한 사례

- 이항 로지스틱 회귀모형 추정 결과

	추정치	표준오차	Z 통계량	p-value
절편	-10.6513	0.3612	-29.5	<0.0001
잔액	0.0055	0.0002	24.9	<0.0001

- 추정된 모형식.

$$\hat{p} = \frac{\exp(-10.6513 + 0.0055x)}{1 + \exp(-10.6513 + 0.0055x)}$$

잔액이 $1 증가 시 연체가능성에 대한 오즈는 exp(0.0055)=1.0055배 증가함.

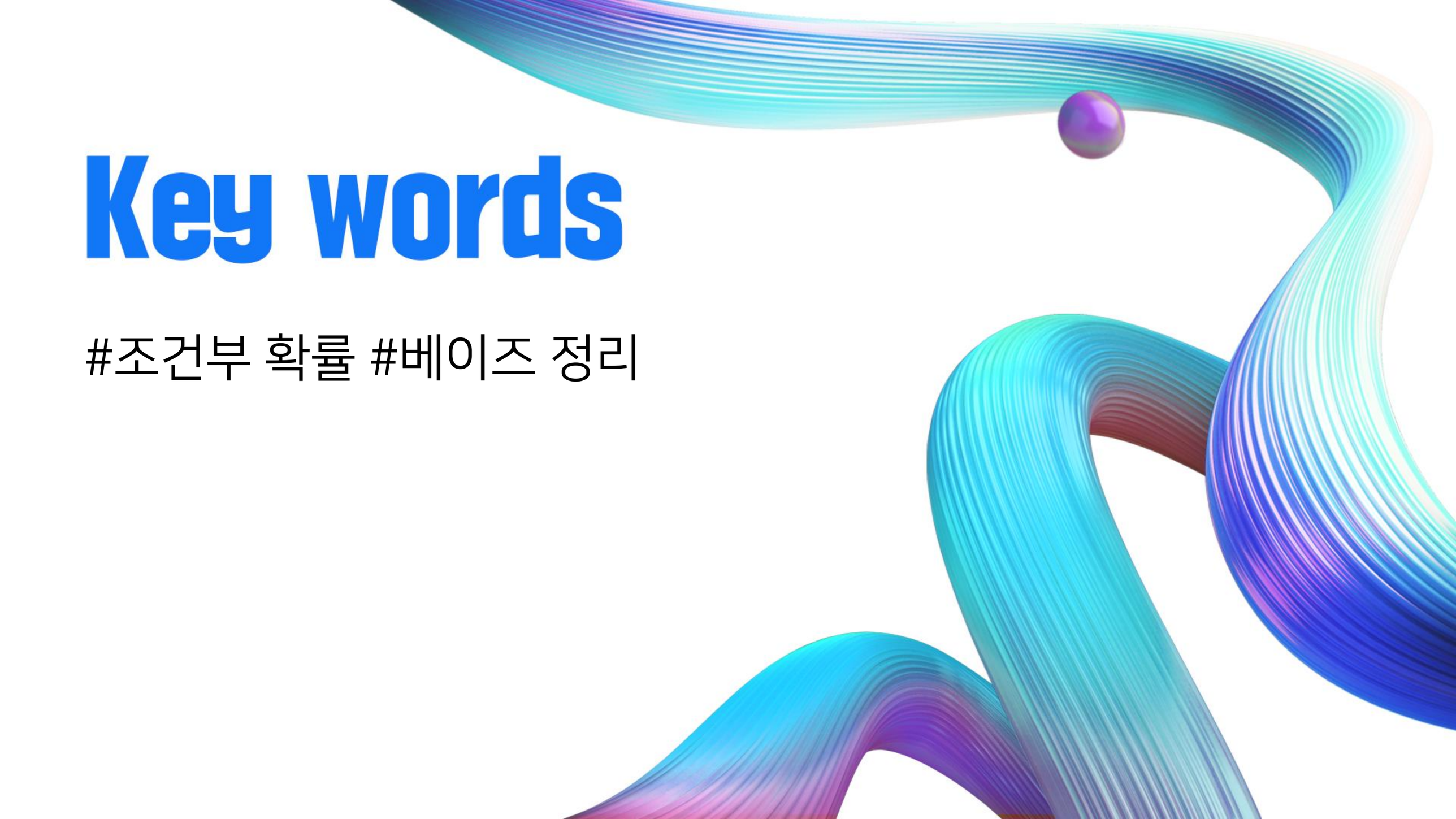
- 잔액이 $2,000인 경우 연체 가능성 예측.

$$\hat{p} = \frac{\exp(-10.6513 + 0.0055 \cdot 2000)}{1 + \exp(-10.6513 + 0.0055 \cdot 2000)} = 0.586$$

분류: 나이브베이즈
(Naïve Bayes)

Key words

#조건부 확률 #베이즈 정리



나이브 베이즈(Naïve Bayes) 분류기

나이브 베이즈 분류기 아이디어

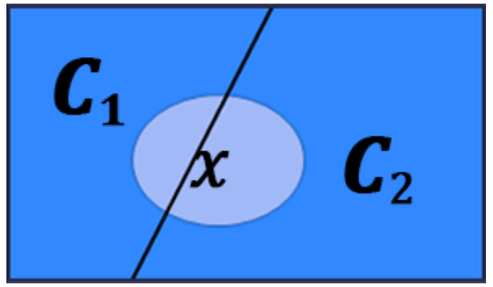
- 목표변수 Y가 2개의 범주 C_1, C_2 를 가진다고 할 때,
  특성변수 X의 값을 이용하여 Y의 범주를 예측하는 문제.
- $X = x$ 로 주어졌을 때 Y의 각 범주에 대한 조건부 확률을 비교하고자 함.
- $P[C_1|x] > P[C_2|x]$ 면 C_1으로 분류하고, 그렇지 않으면 C_2으로 분류함.
- $P[C_k|x]$ 는 훈련 자료에서 추정하기 어려움 $\Rightarrow$ 베이즈 정리를 이용.
- 나이브 베이즈 분류기는 생성(generative) 모델임.

나이브 베이즈(Naïve Bayes) 분류기

베이즈 정리의 활용

- 베이즈 정리에 의하면,

$$P[C_k|x] = \frac{P[x \cap C_k]}{P[x]} = \frac{P[x|C_k]\,P[C_k]}{P[x]}, \qquad k = 1,2$$



- $P[C_1|x] > P[C_2|x]$ 인지 여부.

$\Rightarrow \frac{P[x|C_1]P[C_1]}{P[x]} > \frac{P[x|C_2]P[C_2]}{P[x]}$ 인지 여부.

$\Rightarrow P[x|C_1]\,P[C_1] > P[x|C_2]\,P[C_2]$ 인지 여부로 판단하고자 함.

- $P[x|C_k]$와 $P[C_k]$ 는 훈련 데이터를 이용하여 쉽게 추정할 수 있음.

나이브 베이즈(Naïve Bayes) 분류기

n개의 특성변수를 가지는 분류 문제

- $P[C_k|x_1, \dots, x_n]$에 베이즈 정리를 적용.

$$P[C_k|x_1, \dots, x_n] = \frac{P[C_k]\, P[x_1|C_k]\, P[x_2|x_1, C_k] \ \dots\ P[x_n|x_1, \dots, x_{n-1}, C_k]}{P[x_1, \dots, x_n]}, \ k = 1,2$$

- 각 특성변수들이 모두 독립이라고 가정하면,

$$P[C_k]\, P[x_1|C_k]\, P[x_2|x_1, C_k] \ \dots\ P[x_n|x_1, \dots, x_{n-1}, C_k]$$
$$= P[C_k]\, P[x_1|C_k]\, P[x_2|C_k] \dots P[x_n|C_k]$$
$$= P[C_k]\, (\prod_{i=1}^{n} P[x_i|C_k]), \ k = 1,2$$

나이브 베이즈(Naïve Bayes) 분류기

n개의 특성변수를 가지는 분류 문제

- 예측

  $P[C_1]\,(\prod_{i=1}^{n} P[x_i|C_1]) \geq P[C_2]\,(\prod_{i=1}^{n} P[x_i|C_2])$면 범주 1로 분류.

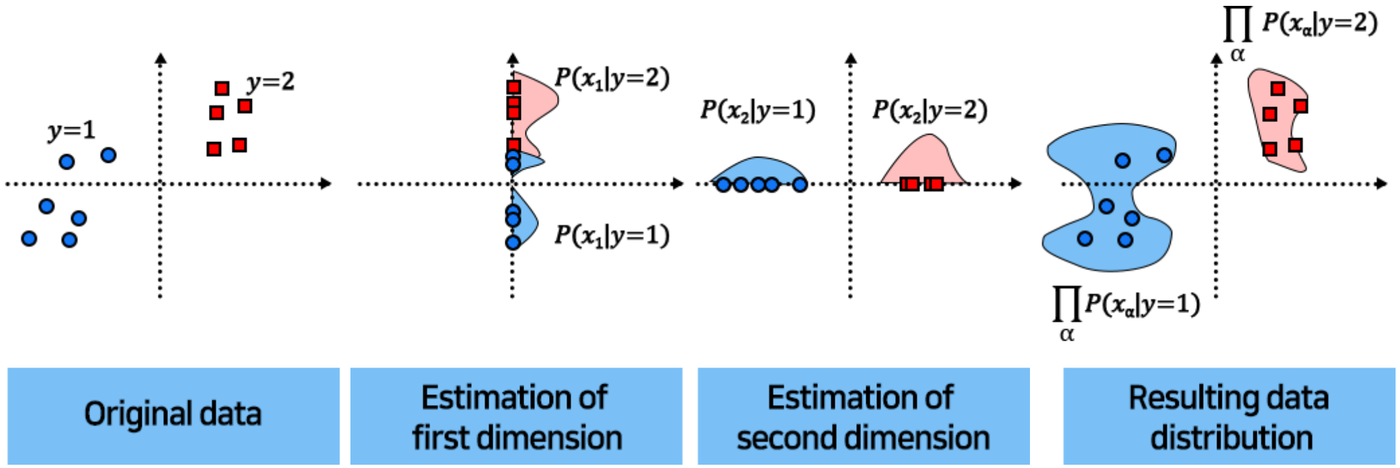
  $P[C_1]\,(\prod_{i=1}^{n} P[x_i|C_1]) < P[C_2]\,(\prod_{i=1}^{n} P[x_i|C_2])$면 범주 2로 분류.

  - $P[C_k]$
    - k번째 범주에 속할 확률.
  - $P[x_i|C_k]$
    - 목표변수가 k번째 범주일 때, 각 특성변수 x_i가 관찰될 확률.
    - x_i 의 자료형식(범주형/개수형/연속형)에 따라 적절한 확률분포를 가정하여 추정.

나이브 베이즈(Naïve Bayes) 분류기

n개의 특성변수를 가지는 분류 문제

- 예측



나이브 베이즈(Naïve Bayes) 분류기

나이브 베이즈 분류기 예시

- X1 = 'Sunny'이고 X2 = 'Normal'일 때, Y가 'Yes'인지 'No'인지를 분류하고자 함.

날씨(x1)	습도(x2)	테니스시합(Y)
Sunny	High	No
Sunny	High	No
Cloudy	High	Yes
Rain	High	Yes
Rain	Normal	Yes
Rain	Normal	No
Cloudy	Normal	Yes
Sunny	High	No
Sunny	Normal	Yes
Rain	Normal	Yes
Sunny	Normal	Yes
Cloudy	High	Yes
Cloudy	Normal	Yes
Rain	High	No

P[Y='Yes'] = 9/14
P[X1='Sunny'|Y='Yes']=2/9
P[X1='Normal'|Y='Yes']=6/9
P[Y='Yes' | X1='Sunny'이고 X2='Normal']
∝ 2/9 × 6/9 × 9/14=**0.095**

P[Y='No'] = 5/14
P[X1='Sunny'|Y='No']=3/9
P[X1='Normal'|Y='No']=1/9
P[Y='No' | X1='Sunny'이고 X2='Normal']
∝ 3/5 × 1/5 × 5/14= **0.043**

'Yes'로 최종 분류

나이브 베이즈(Naïve Bayes) 분류기

나이브 베이즈의 장단점

- 장점

  - 데이터의 크기가 커도 연산 속도가 빠름.
  - 학습에 필요한 데이터 양이 적어도 좋은 성능을 보이는 편.
  - 다양한 텍스트 분류나 추천 등에 활용됨.

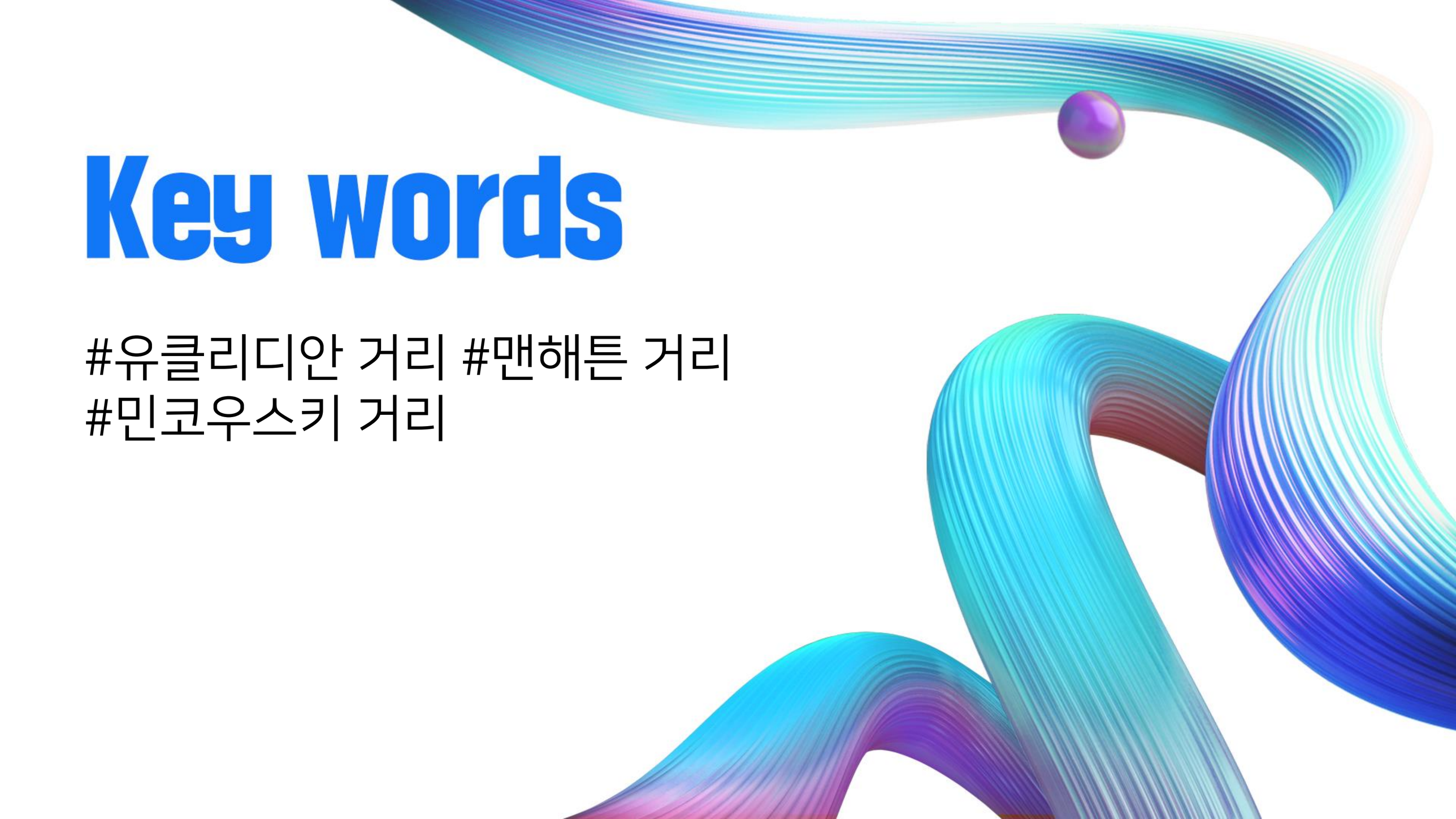
- 단점

  - Zero frequency 문제나 Underflow 문제가 있음.
  - 모든 독립변수가 독립이라는 가정이 너무 단순함.

분류: KNN(K-nearest Neighbor) Classifier

Key words

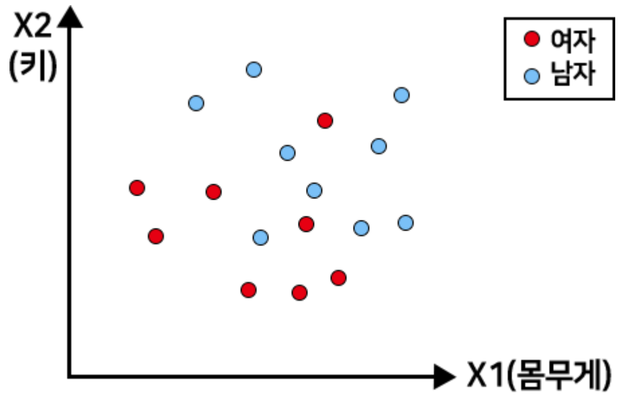
#유클리디안 거리 #맨해튼 거리
#민코우스키 거리



KNN(K-Nearest Neighbor)

KNN 알고리즘

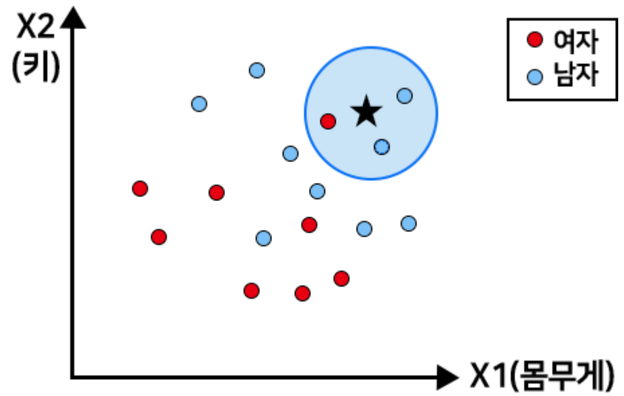
- 가장 간단한 지도학습 머신러닝 알고리즘.
- 훈련데이터를 저장해 두는 것이 모델을 만드는 과정의 전부임.



KNN(K-Nearest Neighbor)

KNN 알고리즘

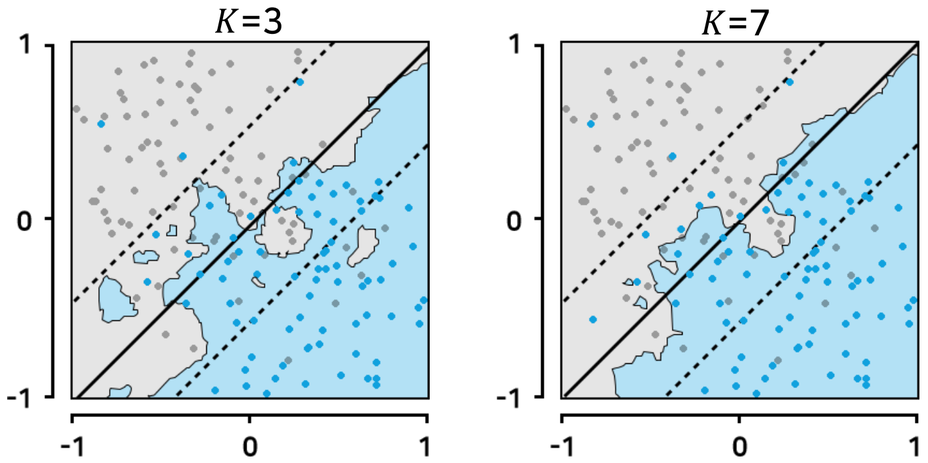
- 새로운 데이터가 입력되면 그 새로운 데이터 주변의 가장 가까운 K개의 훈련 데이터의 레이블을 확인한 뒤, 가장 많이 보이는 라벨로 분류하는 방법.
  - $K = 3$ 인 경우 새로운 데이터(★)에 대한 예측 : 남자



KNN(K-Nearest Neighbor)

K의 결정

- KNN에서 K의 결정은 매우 중요한 문제임.
- K가 작으면 이상점 등의 노이즈에 민감하게 반응하는 과적합의 문제.
- K가 크면 자료의 패턴을 잘 파악할 수 없어 예측 성능이 저하됨.
- 검증용(validation) 데이터를 이용하여 주어진 훈련 데이터에 가장 적절한 K를 찾아야 함.

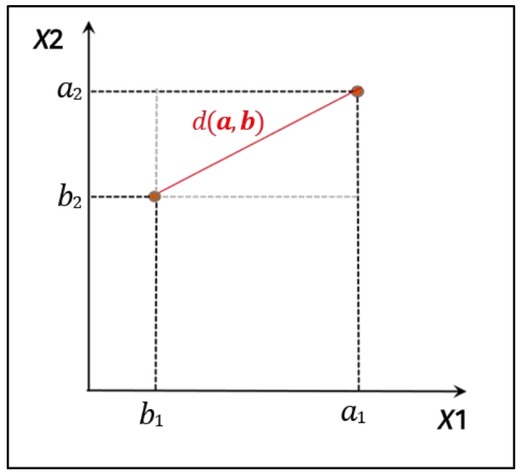


KNN(K-Nearest Neighbor)

거리의 측정

- n개의 특성변수를 가지는 자료에서 두 개의 관찰점 $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$와 $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$간의 거리를 측정하는 문제
- 유클리디안 거리

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$



KNN(K-Nearest Neighbor)

거리의 측정

- 맨해튼 거리

$$d(\boldsymbol{a}, \boldsymbol{b}) = |a_1 - b_1| + |a_2 - b_2| + \cdots + |a_n - b_n|$$

- 민코우스키 거리

$$d(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}$$

KNN(K-Nearest Neighbor)

거리의 측정

- 자료에 스케일에 차이가 있는 경우,
  스케일이 큰 특성변수에 의해 거리가 결정되어 버릴 수 있음.
  따라서 각 특성변수 별로 스케일이 유사해 지도록
  표준화 변환(Z score) 또는 min-max 변환으로
  스케일링을 해준 뒤 거리를 재는 것이 적절함.