교육부 Ministry of Education

NRF 한국연구재단 National Research Foundation of Korea

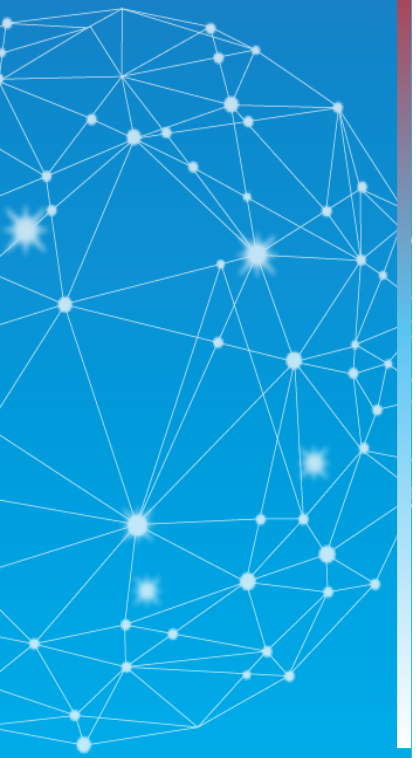
6장. 분류_로지스틱 회귀, 최근접이웃, 나이브베이즈

- 참고문헌: 공돌이의 수학정리노트, 나이브베이즈

  https://bkshin.tistory.com, 머신러닝 나이브베이즈분류

  StatQuest with Josh Starmer, Naïve Bayes

Contents



- 나이브 베이즈

6-3. 나이브 베이즈

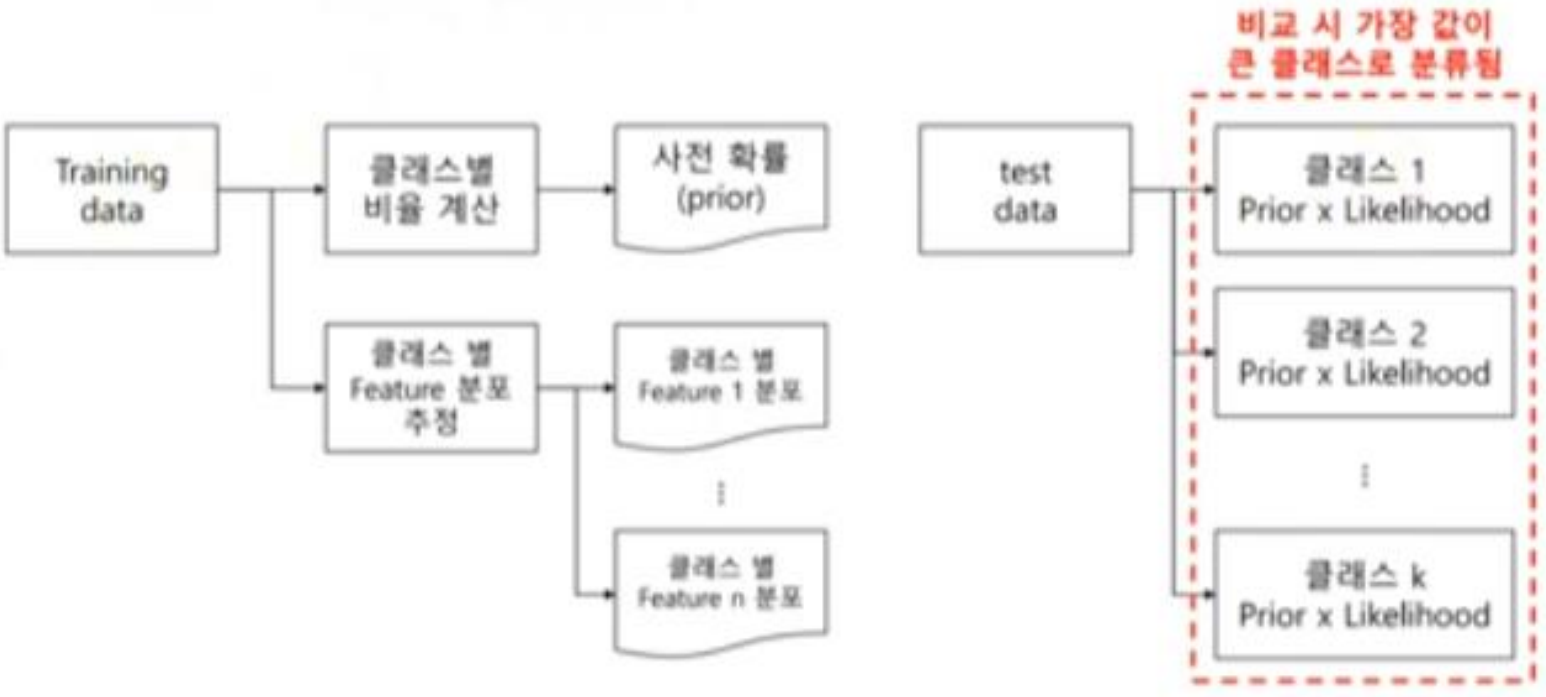
❖ 나이브 베이즈 분류

- 베이즈 정리에 기반한 통계적 분류 기법
- 가장 단순한 지도 학습 (supervised learning) 중 하나
- 모든 Feature가 서로 독립(independent)이어야 한다는 가정이 필요
- 각 클래스에 대한 가능도(likelihood) 비교를 통한 분류
- 최대우도법과 베이즈 정리의 이해 필요
- 분류 판단근거 = 사전지식*추가정보
- 스팸 메일 필터, 텍스트 분류, 감정 분석, 추천 시스템, 제품평가 등에 광범위하게 활용되는 분류 기법

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 작동원리



6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 장점

- ✓ 구현하기 쉬움
- ✓ 적은 양의 학습데이터로도 학습이 가능
- ✓ 대부분의 경우에 비교적 좋은 결과를 얻음
- ✓ 연속형보다 이산형 데이터에서 성능이 좋음

➢ 단점

- ✓ 실제 문제에 Conditional independence 가정의 부적합으로 인해 정확도에 손실이 있음
- ✓ 입력 특징변수 사이의 연관성을 학습할 수 없음

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 조건부 확률

✓ 어떤 사건이 발생했을 때 다른 사건이 일어나는데 영향을 줄 수 있음

✓ 한 사건이 다른 사건에 영향을 받아 발생할 확률이 달라짐을 의미

✓ 사건 B가 일어났을 때 사건 A가 일어날 확률

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

- $P(A)$: 사전확률 (prior probability)
- $P(A|B)$: 사후확률 (posterior probability)

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 결합 확률

✓ 확률 모델을 구현하는데 가장 중요한 요소

✓ 랜덤 변수의 집합 정의

➢ 랜덤 변수들 사이의 결합 확률

✓ 의미

➢ 여러 사건이 함께 일어날 가능성에 대한 통계적 척도

➢ 두 사건이 독립인 경우

$$P(A \cap B) = P(A)P(B)$$

➢ 두 사건이 독립이 아닌 경우

$$P(A \cap B) = P(B|A)P(A)$$

✓ 랜덤 변수들 사이에 독립이 아닌 경우 결합확률을 구해야 변수들 사이의 조건부 확률을 알아낼 수 있음

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 베이지안 정리

✓ 일어나지 않았거나 불확실한 사건에 대한 확률로

(1) 주관적인 가설의 사전 확률을 정하고

(2) 관찰된 데이터를 기반으로 가능도를 계산해서

(3) 처음 설정한 주관적 확률을 보정

✓ 정보를 업데이트하면서 사후 확률 P(A|B)를 구하는 것

$$P(A \mid B) = \frac{P(B \cap A)P(A)}{P(B)}$$

$$P(B) = P(B \cap A) + P(B \cap A')$$

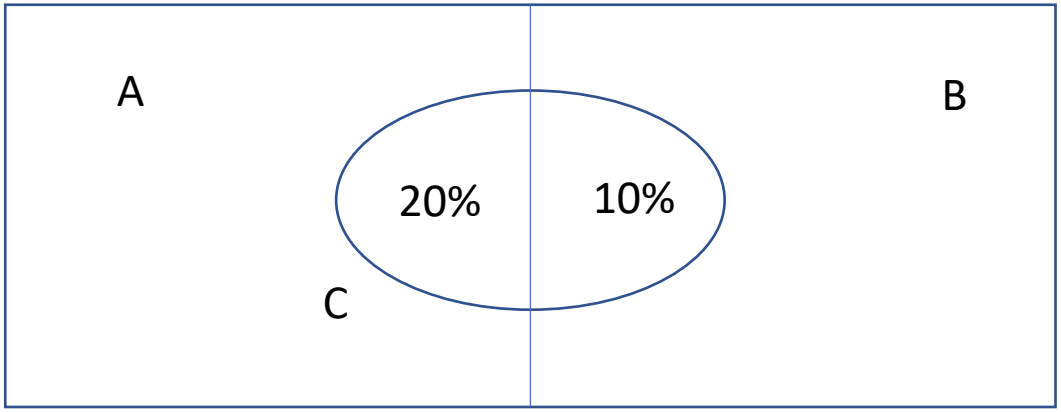
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')(P(A'))}$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 베이지안 정리

✓ 기계 A, B 있고, 기계 A에서 불량제품을 생산할 확률이 20%, 기계 B에서 불량제품을 생산할 확률이 10% 라고 할 때, 제품 검사결과 불량품이 나왔을 때 이 불량품이 기계 A에서 생산되었을 확률?



$$P(A|C) = \frac{0.2 * 0.5}{0.2 * 0.5 + 0.1 * 0.5}$$

6-3. 나이브 베이즈

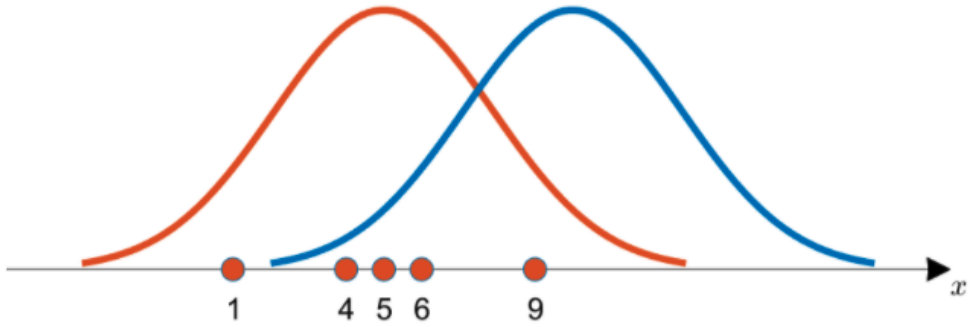
❖ 나이브 베이즈 분류

➢ 최대우도법

✓ 우도(likelihood, 가능도)를 최대화하는 지점을 찾는 것

✓ 우도

- 모델과 추정치가 데이터와 잘 맞는 정도를 확률로 표현한 것
- 데이터가 주어졌을 때 분포의 likehood를 의미
- 데이터가 이 분포로부터 나왔을 가능도



❖ 나이브 베이즈 분류

➢ 최대우도법(Maximum Likelihood Estimation)

- ✓ 데이터의 밀도를 추정하는 한 방법으로 파라미터로 구성된 어떤 확률 밀도 함수에서 관측된 표본 데이터 집합이 있고, 이 표본에서 파라미터를 추정하는 방법
- ✓ 모델 파라미터를 관측 값에만 의존하여 예측하는 방법으로 주어진 파라미터를 기반으로 likelihood를 최대화

Likelihood function

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

Log-Likelihood function

$$L(\theta|x) = logP(x|\theta) = \sum_{i=1}^{n} logP(x_i|\theta)$$

Log-Likelihood function 편미분

$$\frac{\partial}{\partial\theta} L(\theta|x) = \frac{\partial}{\partial\theta} logP(x|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} logP(x_i|\theta) = 0$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 최대우도법(Maximum Likelihood Estimation)

✓ MAP(Maximum a Posteriori Estimiation) 최대 사후 확률 추정법

- 주어진 관측 결과와 사전 확률을 결합해 최적의 모수를 찾아내는 방법
- MLE가 $f(x|\theta)$ 라면 MAP는 $f(\theta|x)$
- 데이터와 제일 잘 맞는 추정치를 찾고 주어진 데이터를 기반으로 최대 확률을 갖는 파라미터를 찾음

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 조건부 독립의 정의

X 의 확률 분포가 Z 의 값이 주어지는 경우 Y 에 대해 독립

즉, $P(X=x \mid Y=y, Z=z) = P(X=x \mid Z=z)$ 이면 X 는 Y 에 대해 conditional independence

$$P(X|Y,Z) = P(X|Z)$$

$$P(X,Y|Z) = P(X|Y,Z)P(Y|Z) = P(X|Z)P(Y|Z)$$

$$P(X_1, X_2, \ldots, X_n|Y) = \prod_{k=1}^{n} P(X_k|Y)$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

- 나이브 베이즈는 조건부 확률 모델
- 분류될 인스턴스들은 N 개의 특성 (독립변수)을 나타내는 벡터 X로 표현되며, 나이브 베이즈 분류기는 이 벡터를 이용하여 k개의 가능한 확률적 결과들 (클래스)을 할당

$$P(C_k|x_1, \cdots, x_n)$$

- 베이즈 정리와 조건부확률 이용

$$P(C_k|x_1, \cdots, x_n) = \frac{P(C_k)P(x_1, \cdots, x_n|C_k)}{P(x_1, \cdots, x_n)}$$

- 다음과 같이 표현가능

$$posterior = \frac{prior \ * likelihood}{evidence}$$

관찰값

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 분자는 결합확률 모델

$$P(C_k, x_1, \cdots, x_n) = P(C_k)P(x_1, \cdots, x_n|C_k)$$

$$= P(C_k)P(x_1|C_k)\, P(x_2, \cdots, x_n|C_k, x_1)$$

$$= P(C_k)P(x_1|C_k)\, P(x_2|C_k, x_1)\, P(x_3, \cdots, x_n|C_k, x_1, x_2)$$

$$= P(C_k)P(x_1|C_k)\, P(x_2|C_k, x_1) \cdots P(x_n|C_k, x_1, x_2, x_3, \cdots, x_{n-1})$$

조건부 독립성

$$= P(C_k)\, P(x_1|C_k)\, P(x_2|C_k) \cdots$$

$$= P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

$$P(C_k|x_1, \cdots, x_n) = \frac{P(C_k)P(x_1, \cdots, x_n|C_k)}{P(x_1, \cdots, x_n)} = \frac{P(C_k) \prod_{i=1}^{n} P(x_i|C_k)}{P(x_1, \cdots, x_n)}$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Bayesian 분류기 학습

✓ 학습데이터 셋 D에 대해 각 데이터는 n개의 특징 변수로 이루어짐

$$X = (x_1, x_2, \dots, x_n)$$

✓ m 개의 클래스 존재 $C_1, C_2, \dots, C_m$

✓ If $P(C_i|X) > P(C_j|X)$ $(1 \le j \le m, j \ne i)$, X 를 클래스 i 로 예측

✓ Bayesian 분류기의 학습 방법

- Maximum posteriori probability 방법 적용

$$P(C_k|x_1, \cdots, x_n) = \frac{P(C_k)P(x_1, \cdots, x_n|C_k)}{P(x_1, \cdots, x_n)} = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x_1, \cdots, x_n)}$$

- $P(X)$ 가 상수이므로 결국 $P(X|C_i)P(C_i)$ 를 최대로 하는 것과 동일

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Naïve 가정에 따른 $P(x_k \mid C_i)$ 계산

✓ Discrete 랜덤 변수인 경우

$$P(x_k|C_i) = \frac{x_k\text{를 포함하는 학습 데이터 가운데 클래스}C_i\text{에 속하는 개수}}{\text{클래스 }C_i\text{에 속하는 학습 데이터의 개수}}$$

✓ Continuous 랜덤 변수인 경우

- 일반적으로 평균 μ과 표준편차 σ를 가지는 Gaussian 분포로 가정

$$P(x_k|C_i) = \mathcal{N}(x_k, \mu_{C_i}, \sigma_{C_i})$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Training 과정

✓ X_i 가 discrete인 경우

✓ 각 y_k 에 대해 π_k 추정 $\quad \widehat{\pi_k} = P(Y = y_k)$

✓ 특징변수 X_i 의 값 x_{ij} 에 대해 $\theta_{i,j,k}$ 추정 : MLE (Maximum Likelihood estimation)

$$\widehat{\theta_{i,j,k}} = P(X_i = x_{ij} | Y = y_k)$$

➢ Testing 과정

✓ 새로운 입력 $X^{new} = (X_1^{new}, X_2^{new}, \dots, X_n^{new})$ 분류 : MAP (Maximum a Posteriori estimation)

$$Y^{new} = arg \max_{y_k} \widehat{\pi_k} \prod_i \widehat{\theta_{i,new,k}}$$

❖ 나이브 베이즈 분류

➢ "Zero" problem

✓ 학습데이터에 $X\_1=a$ 이고 Y=b 인 경우가 없었다면

$$\hat{P}(X_1 = a|Y = b) = 0$$

✓ $X_2, X_3, \dots, X_n$ 의 값에 무관하게

$$\hat{P}(Y = b|X_1 = a, X_2, \dots, X_n) = 0$$

$$\hat{P}(X_1 = a, X_2, \dots, X_n|Y) = \hat{P}(X_1 = a\,|Y) \prod_{i=1}^{n} \hat{P}(X_i|Y)$$

✓ Add-one 으로 해결

$$\hat{P}(X_1 = a|Y = b) = \frac{T_{ba} + 1}{\sum_{a'}(T_{ba'} + 1)}$$

- T_ba : 클래스 b 에서 특징변수 a 의 빈도수

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Numerical underflow 문제

- ✓ 입력 특징변수의 개수가 많아 지면 $P(x|Y)$ 값은 매우 작을 수 있음
- ✓ 이로 인해 특정 고차원 입력 특징벡터를 관찰할 확률이 작게 됨
- ✓ 이것은 underflow 를 일으킬 가능성이 있음
- ✓ Log 적용하여 해결

$$Y^{new} = arg \max_{y_k} \widehat{\pi_k} \prod_i \widehat{\theta_{i,new,k}}$$

$$Y^{new} = arg \max_{y_k} \left[\log \widehat{\pi_k} + \sum_i log \, \widehat{\theta_{i,new,k}} \right]$$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 스팸메일 분류

총 17개

정상

Dear: 8

Friend: 5

Lunch: 3

Money: 1

스팸

Dear: 2

Friend: 1

Lunch: 0

Money: 4

총 7개

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

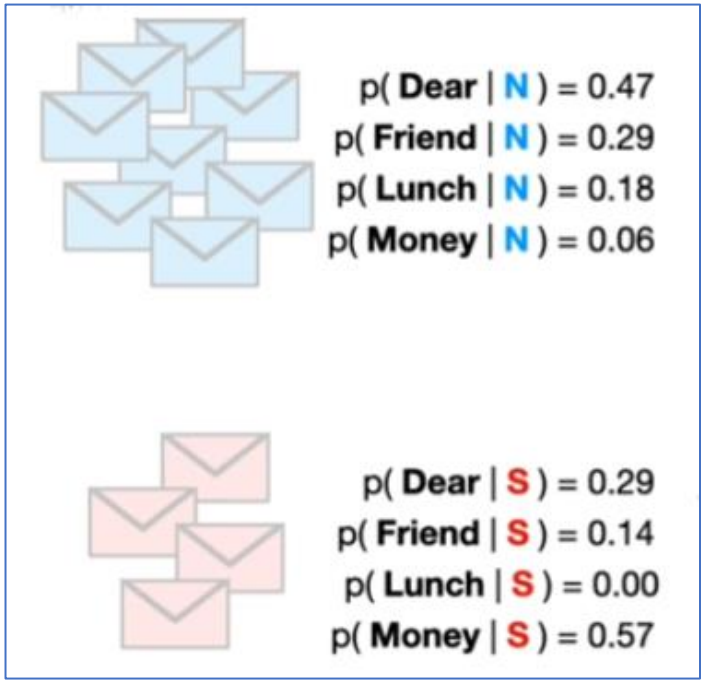
➢ 스팸메일 분류

새로운 메일

Dear Friend

p(N)=0.67

정상



스팸

p(S)=0.33

$p(N)*p(Dear|N)*p(Friend|N)$
$= 0.67*0.47*0.29 = 0.09$

$p(N|Dear\ Friend) \propto 0.09$ 정상메일

$p(S|Dear\ Friend) \propto 0.01$

$p(S)*p(Dear|S)*p(Friend|S)$
$= 0.33*0.29*0.14 = 0.01$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

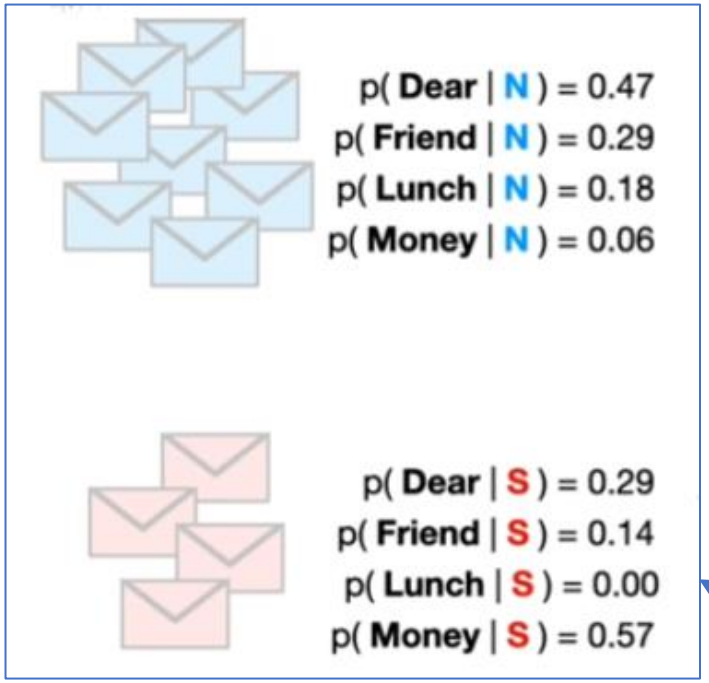
➢ 스팸메일 분류

새로운 메일

Lunch Money Money Money Money

p(N)=0.67

정상



스팸

p(S)=0.33

$p(N)*p(Lunch|N)*p(Money|N)^4$
$= 0.67*0.18*0.06^4 = 0.000002$

∨ 정상메일

$p(S)*p(Lunch|N)*p(Money|N)^4$
$= 0.33*0*0.57^4 = 0$

잘못된 결정!!

이유

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 스팸메일 분류

새로운 메일

Lunch Money Money Money Money

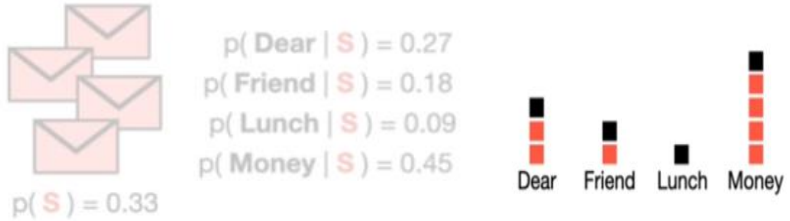
p(N)=0.67

정상

$p(N)*p(Lunch|N)*p(Money|N)^4$
$= 0.67*0.19*0.1^4 = 0.00001$

∧ 스팸메일

스팸



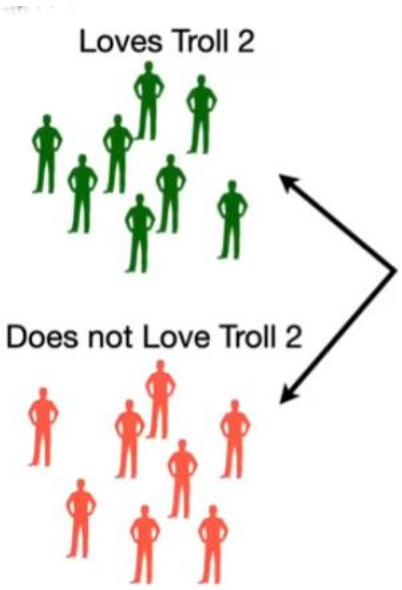
$p(S)*p(Lunch|N)*p(Money|N)^4$
$= 0.33*0.09*0.45^4 = 0.00122$

p(S)=0.34

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

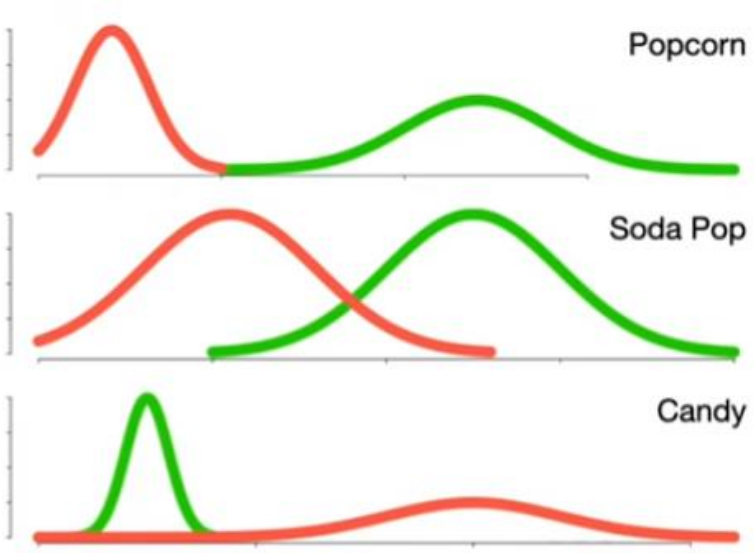
➢ 영화 Troll 2 선호도 분류



Popcorn (grams)	Soda Pop (ml)	Candy (grams)
24.3	750.7	0.2
28.2	533.2	50.5
etc.	etc.	etc.

Popcorn (grams)	Soda Pop (ml)	Candy (grams)
2.1	120.5	90.7
4.8	110.9	102.3
etc.	etc.	etc.

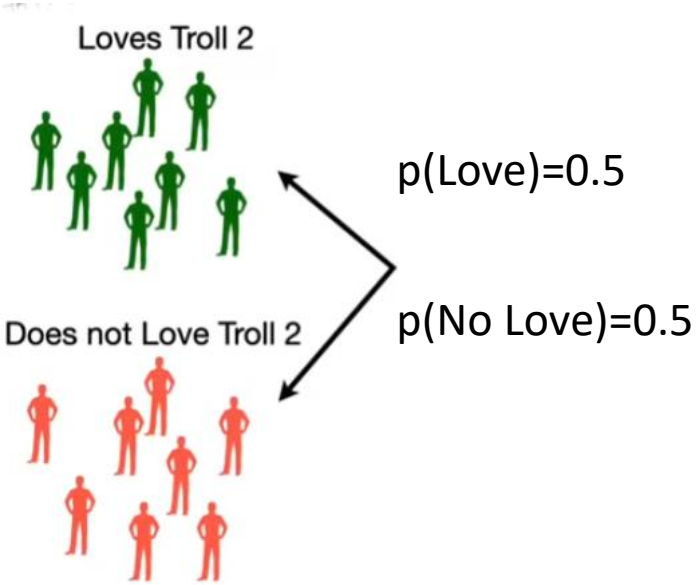
likelihood



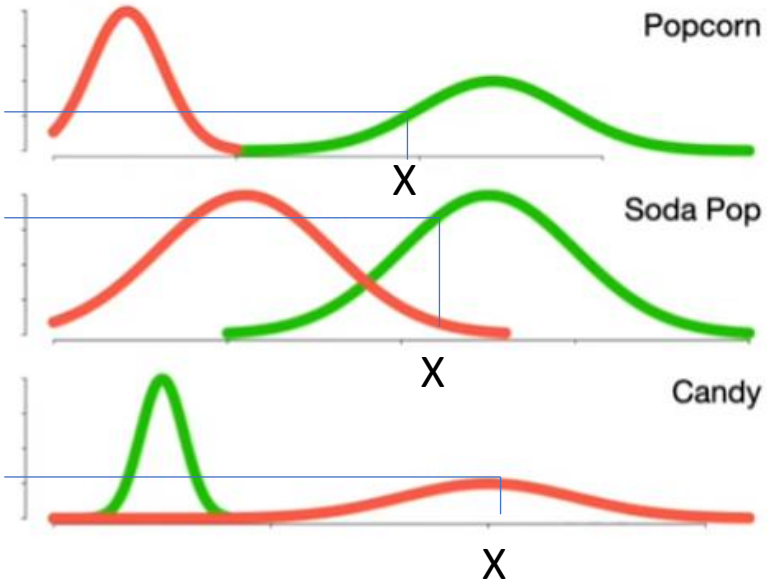
6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 영화 Troll 2 선호도 분류



likelihood



p(Love)*L(popcorn=20|Love)*L(soda=500|Love)*L(candy=25|Love)

p(No Love)*L(popcorn=20|No Love)*L(soda=500|No Love)*L(candy=25|No Love)

로그 변환으로 계산!

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ 날씨와 기온에 따른 축구 경기 여부 분류

날씨 (x1)	기온(x2)	경기여부(c1=yes /c2=no)
Sunny	Hot	C2
Sunny	Hot	C2
Overcast	Hot	C1
Rainy	Mild	C1
Rainy	Cool	C1
Rainy	Cool	C2
Overcast	Cool	C1
Sunny	Mild	C2
Sunny	Cool	C1
Rainy	Mild	C1
Sunny	Mild	C1
Overcast	Mild	C1
Overcast	Hot	C1
Rainy	Mild	C2

❖ 나이브 베이즈 분류

➢ 날씨와 기온에 따른 축구 경기 여부 분류 (학습과정)

- ✓ 전체 학습 데이터 개수 = 14
- ✓ $P(C_1)$ = 9/14 = 0.64
- ✓ $P(C_2)$ = 5/14 = 0.36
- ✓ 클래스별 likelihood 계산

x_1	$P(x_1 = v\|C_1)$	$P(x_1 = v\|C_2)$
Sunny	$\frac{2}{9}$	$\frac{3}{5}$
Rainy	$\frac{3}{9}$	$\frac{2}{5}$
Overcast	$\frac{4}{9}$	0

x_2	$P(x_2 = v\|C_1)$	$P(x_2 = v\|C_2)$
Hot	$\frac{2}{9}$	$\frac{2}{5}$
Mild	$\frac{4}{9}$	$\frac{2}{5}$
Cool	$\frac{3}{9}$	$\frac{1}{5}$

❖ 나이브 베이즈 분류

➢ 날씨와 기온에 따른 축구 경기 여부 분류 (예측과정)

✓ X=[$Overcast, Mild$] 에 대한 클래스 예측

- C_1 에 대한 확률 예측

$$P(C_1|X) \propto P(X|C_1)P(C_1) = P(x_1 = overcast|C_1)P(x_2 = mild|C_1)P(C_1)$$

$$= \frac{4}{9} * \frac{4}{9} * 0.64 = 0.126$$

- C_2에 대한 확률 예측

$$P(C_2|X) \propto P(X|C_2)P(C_2) = P(x_1 = overcast|C_2)P(x_2 = mild|C_2)P(C_2)$$

$$= 0 * \frac{2}{5} * 0.36 = 0.0$$

✓ 따라서, $X = [Overcast, Mild] \in C_1$

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Scikit-learn을 활용한 나이브 베이즈 분류

✓ 날씨, 기온에 따른 축구 여부 분류

- 2개의 Feature (Weather, Temp)와 1개의 Label (Play)로 구성된 dataset

```
# Assigning features and label variables
weather=['Sunny', 'Sunny', 'Overcast', 'Rainy', 'Rainy', 'Rainy', 'Overcast', 'Sunny',
'Sunny', 'Rainy', 'Sunny', 'Overcast', 'Overcast', 'Rainy']
temp=['Hot', 'Hot', 'Hot', 'Mild', 'Cool', 'Cool', 'Cool', 'Mild', 'Cool', 'Mild', 'Mild',
'Mild', 'Hot', 'Mild']

play=['No', 'No', 'Yes', 'Yes', 'Yes', 'No', 'Yes', 'No', 'Yes', 'Yes', 'Yes', 'Yes', 'Yes', 'No']
```

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Scikit-learn을 활용한 나이브 베이즈 분류

✓ 날씨, 기온에 따른 축구 여부 분류

- string을 int로 바꾸어 주는 Feature Encoding

```
# Import LabelEncoder
from sklearn import preprocessing

#creating labelEncoder
le = preprocessing.LabelEncoder()
# Converting string labels into numbers.
weather_encoded=le.fit_transform(weather)
temp_encoded=le.fit_transform(temp)
label=le.fit_transform(play)
```

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Scikit-learn을 활용한 나이브 베이즈 분류

✓ 날씨, 기온에 따른 축구 여부 분류

- 인코딩 된 두 feature를 결합

```
#Combinig weather and temp into single listof tuples
features = zip(weather_encoded, temp_encoded)
features = list(features)

#Import Gaussian Naive Bayes model
from sklearn.naive_bayes import GaussianNB

model = GaussianNB()
model.fit(features,label)
predicted= model.predict([[0,2]]) # 0:Overcast, 2:Mild
```

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Scikit-learn을 활용한 나이브 베이즈 분류

✓ Label이 여러개인 나이브 베이즈

```
from sklearn import datasets

wine = datasets.load_wine()

# print the names of the 13 features
print("Features: ", wine.feature_names)

# print the label type of wine(class_0, class_1, class_2)
print("Labels: ", wine.target_names)
```

```
Features:  ['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium', 'total_phenols',
'flavanoids', 'nonflavanoid_phenols', 'proanthocyanins', 'color_intensity', 'hue',
'od280/od315_of_diluted_wines', 'proline']
Labels:  ['class_0' 'class_1' 'class_2']
```

6-3. 나이브 베이즈

❖ 나이브 베이즈 분류

➢ Scikit-learn을 활용한 나이브 베이즈 분류

✓ Label이 여러개인 나이브 베이즈

```
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn import metrics

X_train, X_test, y_train, y_test = train_test_split(wine.data, wine.target,
test_size=0.3, random_state=109)

gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred = gnb.predict(X_test)

print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

Accuracy: 0.9074074074074074

교육부 Ministry of Education

NRF 한국연구재단 National Research Foundation of Korea

thank you

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된
디지털신기술인재양성 혁신공유대학사업의 연구결과입니다.

Convergence and Open Sharing System COSS 인공지능

전남대학교 CHONNAM NATIONAL UNIVERSITY

성균관대학교 SUNGKYUNKWAN UNIVERSITY

KNU 경북대학교 KYUNGPOOK NATIONAL UNIVERSITY

전주대학교 JEONJU UNIVERSITY

서울시립대학교 UNIVERSITY OF SEOUL

서울과학기술대학교 SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

영진전문대학교 YEUNGJIN UNIVERSITY