
빅데이터 분석
및 처리 기술
Crayon

# 빅데이터 분류 분석

## 3 분류 심화 : 나이브 베이즈 기법

### 나이브 베이즈 기법

나이브(naïve) : 순진한, 전문 지식이 없는
베이즈 정리 : 두 확률 사이에 존재하는 관계를 설명하는 것

# 빅데이터 분류 분석

## 3 분류 심화 : 나이브 베이즈 기법

- 베이즈 정리
- 가우시안 나이브 베이즈 기법

## I 베이즈 정리(Bayes' theorem)

# 1. 개요

## "베이즈 분류 개념"

- 관측된 특징(features)이 주어졌을 때 레이블(label)의 확률을 계산 $P(L|features)$
- 베이즈 정리를 이용하여 다음과 같이 표현됨 $P(L|features) = \frac{P(features,\ L)}{P(features)} = \frac{P(features|L)P(L)}{P(features)}$
- 레이블이 $L\_1, L\_2$ 두 개인 경우 $P(L_1|features) = \frac{P(features|L_1)P(L_1)}{P(features)}, P(L_2|features) = \frac{P(features|L_2)P(L_2)}{P(features)}$
- 확률 값이 큰 레이블을 선택

$$P(L_1|features) > P(L_2|features) \text{ 이면, } L_1$$
$$P(L_1|features) \leq P(L_2|features) \text{ 이면, } L_2$$

II 나이브 베이즈 분류

# 1. 개요

## "생성 모델"

- 각 레이블에 대한 특징 데이터의 확률값을 계산할 수 있는 모델
- 생성 모델의 형태에 대한 가장 간단한 가정을 사용하여 단순하게 형성 가능

II 나이브 베이즈 분류

# 1. 개요

## "생성 모델"

- 레이블의 사후 확률 계산 방법

$$P(L_1|features) = \frac{P(features|L_1)P(L_1)}{P(features)},\ P(L_2|features) = \frac{P(features|L_2)P(L_2)}{P(features)}$$

- 생성 모델 가정

$$P(feartues|L) \sim N(\mu, \sigma^2)$$

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

"분석 환경 설정"

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
```

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "데이터 준비"

```
from sklearn.datasets import make_blobs
X, y = make_blobs(100, 2, centers=2, random_state=2, cluster_std=1.5)
plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='RdBu');
```

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

"데이터 준비"

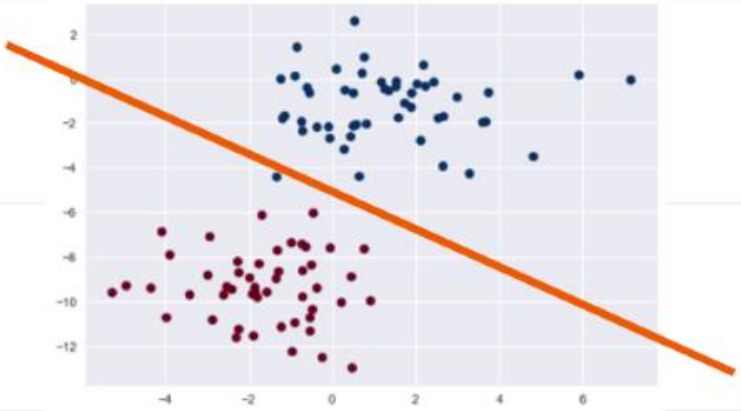
등방성 가우시안 정규분포를 이용하여 가상 데이터 생성

```
from sklearn.datasets import make_blobs
X, y = make_blobs(100, 2, centers=2, random_state=2, cluster_std=1.5)
plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='RdBu');
```

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

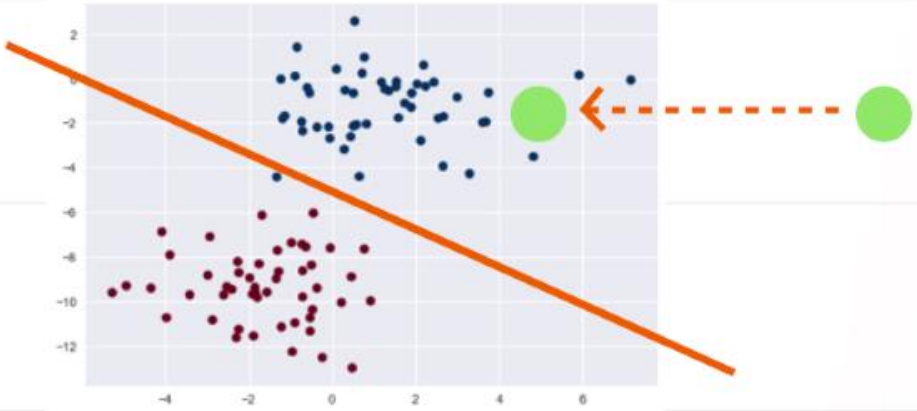
## "데이터 시각화"



II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "데이터 시각화"

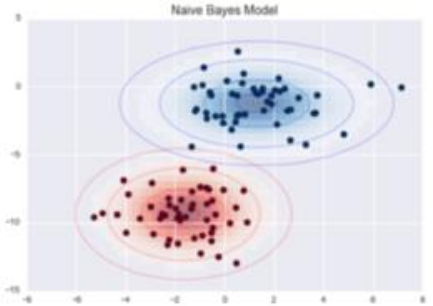


II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "가우시안 생성 모델"

- 전제 : 차원 사이에 공분산이 없는 가우스 분포를 따른다.
- 이 모델은 단순히 각 레이블 내 점의 평균과 표준 편차를 구하여 적합할 수 있다.
- 각 색상의 타원은 타원의 중심으로 갈수록 확률이 더 커지는 각 레이블에 대한 가우스 생성 모델을 나타낸다.



II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "모델 클래스 불러오기"

- from sklearn.naive_bayes import GaussianNB

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "모델 인스턴스 생성"

- model = GaussianNB()

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "모델 적합하기"

- model.fit(X, y)

## "레이블 예측"

- rng = np.random.RandomState(0)
  Xnew = [-6, -14] + [14, 18] * rng.rand(2000, 2)
  ynew = model.predict(Xnew)

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

"모델 적합하기"

- model.fit(X, y)

"레이블 예측"

- rng = np.random.RandomState(0)
  Xnew = [-6, -14] + [14, 18] * rng.rand(2000, 2)
  ynew = model.predict(Xnew)

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

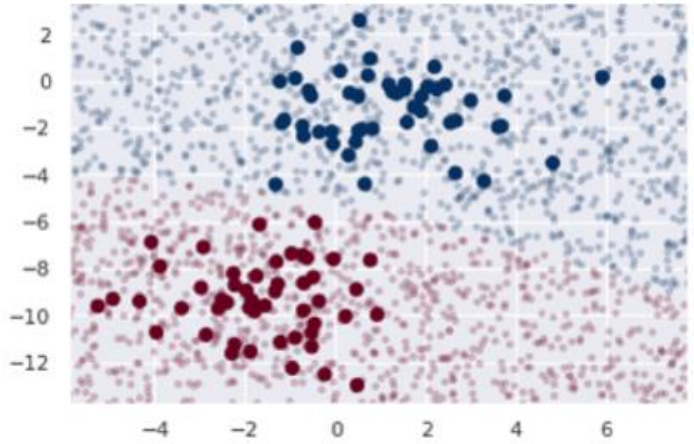
"예측값 시각화"

```
plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='RdBu')
lim = plt.axis()
plt.scatter(Xnew[:, 0], Xnew[:, 1], c=ynew, s=10, cmap='RdBu', alpha=0.2)
plt.axis(lim);
```

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

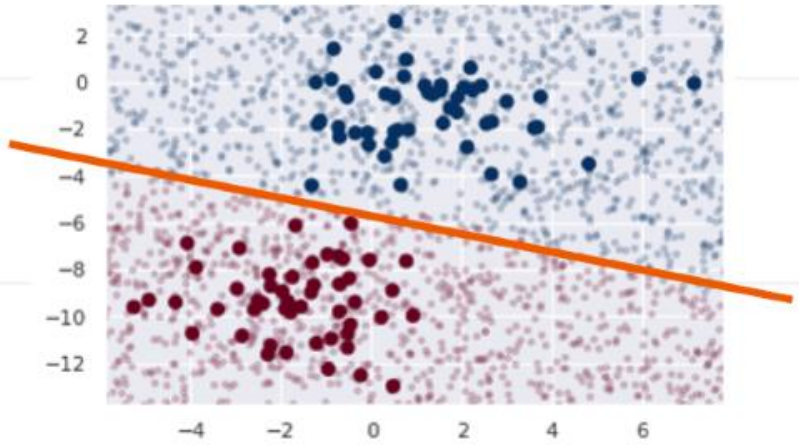
## “시각화 결과”



II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "시각화 결과"



II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "확률 표현의 장점"

- yprob = model.predict_proba(Xnew)
  yprob[-8:].round(2)

```
Array ( [ [ 0.89,  0.11],
          [ 1.  ,  0.  ],
          [ 1.  ,  0.  ],
          [ 1.  ,  0.  ],
          [ 1.  ,  0.  ],
          [ 1.  ,  0.  ],
          [ 0.  ,  1.  ],
          [ 0.15,  0.85 ] ] )
```

II 나이브 베이즈 분류

# 2. 가우스 나이브 베이즈

## "확률 표현의 장점"

- yprob = model.predict_proba(Xnew)
  yprob[-8:].round(2)

```
Array ( [ [ 0.89, 0.11],
          [ 1. , 0. ],
          [ 1. , 0. ],
          [ 1. , 0. ],
          [ 1. , 0. ],
          [ 1. , 0. ],
          [ 0. , 1. ],
          [ 0.15, 0.85 ] ] )
```

분류 결과가 명확한 점과 애매한 점을 구분할 수 있다.

# 이번 시간에는

3 분류 심화 : 나이브 베이즈 기법

베이즈 정리

가우시안 나이브 베이즈 기법

# 이번 시간에는

**실습 참고 자료**

Colab 노트북 파일
Scikit-Learn 공식 사이트 자료
→ https://scikit-learn.org/stable/user_guide.html

# 이번 시간에는

## 과제 안내

과　　제 : 퀴즈
제출 방법 : 과제 게시판 제출 방법 안내 참조

## 질의 응답 게시판

학습 내용, 퀴즈, 과제 등에 대한 질의응답 게시판을 통한 질의응답

# 다음 시간에는

8 빅데이터 군집 분석

- 비지도 학습의 개념
- 심화 : k-평균 군집 분석
- 심화 : 가우스 혼합 모델