

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curve
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

# Gaussian Naïve Bayes

- Gaussian : 모든 feature가 정규분포를 따르는 확률변수
- Naïve : feature는 같은 class 내에서 조건부독립
- Bayes : Bayes' rule을 이용하여 추정모형을 단순화
- 모든 feature가 연속변수라는 것을 제외하면 discrete Naïve Bayes와 대부분의 장단점을 공유한다.
- 신용평가, 질병분류, 안면인식 등에 사용한다.

## 1 Gaussian Bayes

### 1.1 Gaussian Naïve Bayes

- Target  $Y$ 가 이산확률변수, feature  $X$ 는 정규분포를 따르는 연속확률변수일 때 사용하는 분류 모형
- Bayesian의 posterior는 다음과 같다.

$$\Pr(Y = c|X_1, \dots, X_K) = \frac{\Pr(Y = c)\Pr(X_1, \dots, X_K|Y = c)}{\sum_{c=1}^C \Pr(Y = c)\Pr(X_1, \dots, X_K|Y = c)}$$

- $(X_k|Y = c)$ 가 조건부독립이라면 posterior는 다음과 같이 쓸 수 있다.

$$\Pr(Y = c|X_1, \dots, X_K) = \frac{\Pr(Y = c)\prod_{k=1}^K \Pr(X_k|Y = c)}{\sum_c \Pr(Y = c)\prod_{k=1}^K \Pr(X_k|Y = c)}$$

- Gaussian Bayes 모형:  $\Pr(X|Y)$ 가 multivariate Gaussian distribution
- Gaussian Naïve Bayes:  $X_1, \dots, X_K$ 가 조건부독립인 정규분포

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 1.2 연속확률변수의 Bayes' theorem

- Bayes' Theorem의 각 항은 확률을 의미한다. 따라서 연속함수의 경우 확률밀도함수의 값을 직접 적용할 수 없지만, 결과적으로 밀도함수를 확률로 사용한 형태가 된다.

- $g(x, y)$ 를 두 확률변수의 결합확률분포라고 하고  $f(x)$ 를  $X$ 의 밀도함수,  $\Pr(y)$ 를  $Y$ 의 질량함수라고 하자. 각 확률함수는 다음 조건을 만족한다.

$$\int g(x, y)dx = \Pr(y), \quad g(x, 0) + g(x, 1) = f(x)$$

- 여기서  $X$ 에 대한  $Y$ 의 조건부확률은 확률을 확률밀도로 근사한 후 극한을 취해 사용하는 것으로 이해하면 된다. 조건부확률은 L'Hôpital's rule과 Leibniz rule을 이용해 다음과 같이 정리할 수 있다.

$$g_{Y|X}(y) = \lim_{\Delta x \rightarrow 0} \Pr(y|x < X < x + \Delta x) = \lim_{\Delta x \rightarrow 0} \frac{\int_x^{x+\Delta x} g(x, y)dx}{\int_x^{x+\Delta x} f(x)dx} = \lim_{\Delta x \rightarrow 0} \frac{g(x + \Delta x, y)}{f(x + \Delta x)} = \frac{g(x, y)}{f(x)}$$

- 다음 식을  $g(x, y)$ 에 대해 정리해서 위 식에 대입하면 질량함수를 밀도함수로 바꾸어도 Bayes' theorem은 성립한다는 것을 확인할 수 있다.

$$g_{X|Y}(x) = \frac{g(x, y)}{\Pr(y)}$$

## 2 예제: 몸무게와 키를 이용한 성별예측

- 키와 몸무게를 이용하여 남녀 학생을 구분해보자.

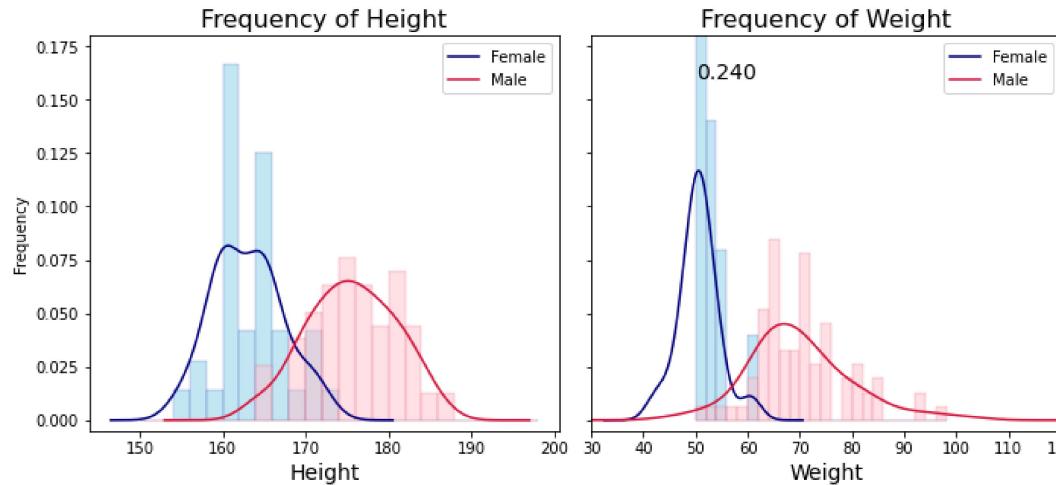
$$Y \in \{\text{female}, \text{male}\}, \quad X = (X_{\text{height}}, X_{\text{weight}})$$

- 키와 몸무게는 성별에 영향을 미치지 않으므로 이 분석은 자료 사이의 단순한 상관관계를 살피려는 목적
- 각 feature는 정규분포를 가정한다. 최소한 unimodal 조건은 만족해야 하며 골이 깊은 bimodal일 경우 주의해서 사용한다.

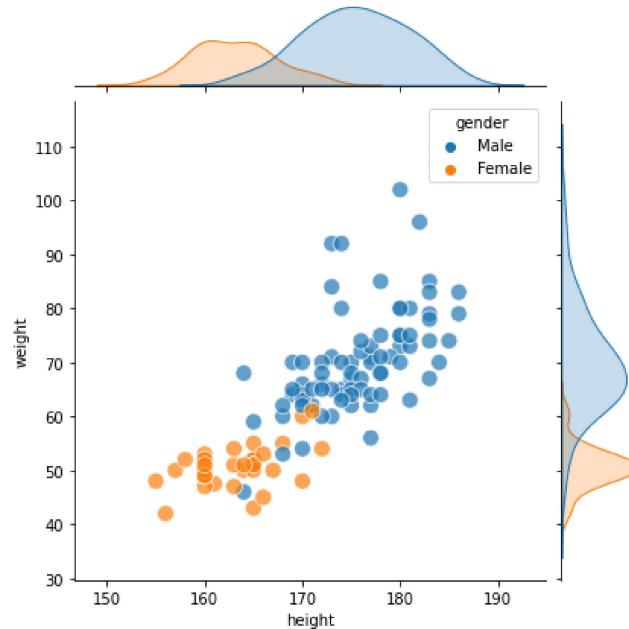
## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

Figure 1. Height and Weight



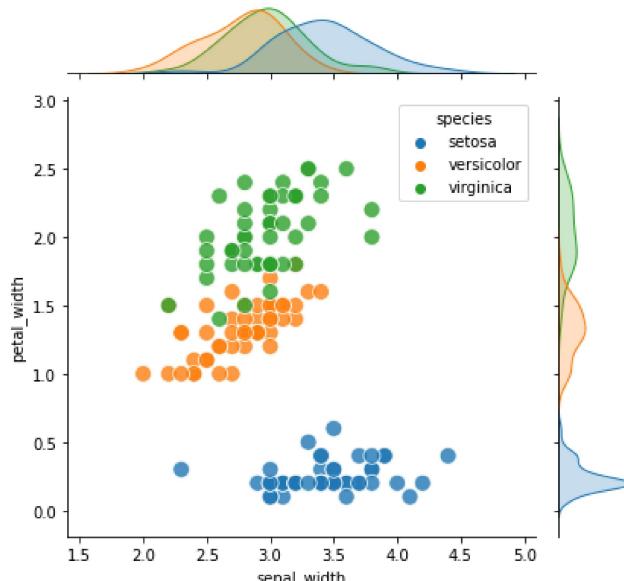
- 수평 혹은 수직분할 중 하나가 남녀 구분에 핵심적인 역할을 한다.
- 다른 분할의 추가하여 임의의 직선을 decision boundary로 사용해도 수평분할보다 별반 나아질 것 같지 않다.



## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

Figure 3. Joint density of sepal and petal widths



## Gaussian Distribution

- 일변수정규분포

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- $\Pr(X|y)$ 가 다변수정규분포 multivariate normal distribution 를 따른다면 확률밀도함수는 다음과 같다.

$$f(X = \mathbf{x}|Y = c) = \frac{1}{(2\pi)^{K/2} |\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

여기서  $\mathbf{x}, \boldsymbol{\mu}_c \in \mathbb{R}^K$ 이고  $\Sigma_c$ 는 크기가  $K \times K$ 인  $y_c$ 에 대한  $X$ 의 조건부 공분산이다.

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산이 다른 경우
  - 4.2 Feature 가 하나이면서 조건부분산이 같은 경우
  - 4.3 Feature 가 여러 개이면서 조건부분산이 다른 경우
  - 4.4 Feature 가 여러 개이면서 조건부분산이 같은 경우
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 3 Gaussian Bayes classifier

- 모형의 각 class에 대한 조건부 공분산의 형태에 따라 구분한다.

### 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes

- 일반적인 형태로 각 feature 사이의 공분산이 0이 아니다.

1.  $X|c$ 의 분산  $\Sigma_c$ 는 모든 class에서 같지 않다.
2.  $X|c$ 의 분산은 모든  $c$ 에 대해  $\Sigma$ 로 동일하다.

$$\Sigma_c = \begin{bmatrix} \sigma_{1|c}^2 & \sigma_{12|c} & \cdots & \sigma_{1K|c} \\ \sigma_{21|c} & \sigma_{2|c}^2 & \cdots & \sigma_{2K|c} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1|c} & \sigma_{K2|c} & \cdots & \sigma_{K|c}^2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_K^2 \end{bmatrix}$$

- $Y$ 는  $C$ 개의 범주를 갖는 확률변수이고  $X$ 는 다변수 Gaussian 분포를 따르는 연속확률 변수일 경우 posterior:

$$\Pr(y_c|\mathbf{x}, \boldsymbol{\mu}, \Sigma) \propto \Pr(Y=c) \times \frac{1}{(2\pi)^{K/2} |\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

- 각  $c$ 에 대해 prior  $\pi_c \equiv \Pr(Y=c)$ 와 likelihood  $\Pr(X_1, \dots, X_K|Y=c)$ 을 추정

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 3.2 Gaussian Naïve Bayes

- 모든  $X_k|y_c$ 가 조건부독립이면 공분산은 0이고 따라서  $\Sigma_c$ 는 대각행렬이다.

1.  $X|c$ 의 공분산행렬  $\Sigma_c = \text{diag}(\sigma_{k|c}^2)$ 은 다르다.
2. 모든 class에서 공분산행렬  $\Sigma_c = \text{diag}(\sigma_c^2) = \Sigma$ 은 동일하다.

$$\Sigma_c = \begin{bmatrix} \sigma_{1|c}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2|c}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{K|c}^2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_K^2 \end{bmatrix},$$

- posterior는 다음과 같다.

$$\Pr(y_c|\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \propto \Pr(Y=c) \times \prod_{k=1}^K \frac{1}{(2\pi)^{K/2} \sigma_{k|c}} \exp\left(-\frac{1}{2} \left(\frac{x_k - \mu_{k|c}}{\sigma_{k|c}}\right)^2\right)$$

### 3.2.1 Posterior 계산에 필요한 정보

- prior  $\pi_c \equiv \Pr(Y=c)$
- likelihood  $\Pr(X_1, \dots, X_K|Y=c)$ 
  - 각 class에 해당하는 조건부 평균  $\mu_{k|c}$
  - 각 class에 해당하는 조건부 분산  $\sigma_{k|c}^2$

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

### 3.2.2 Class마다 공분산이 다를 경우 추정할 모수의 수

- Prior:
  - Class의 수가  $C$ 이고 그 합은 1이므로  $C - 1$ 개의 확률값
- Likelihood:
  - 각 class에 해당하는 조건부 평균  $\mu_{k|c}$ 에  $C \times K$ 개의 모수
  - 각  $\Sigma_c$ 에는  $\frac{K(K+1)}{2}$  개의 모수
- 추정해야 하는 모수의 수는

$$CK + \textcolor{red}{C} \times \frac{K(K+1)}{2} + C - 1$$

### 3.2.3 모든 class의 공분산이 같을 경우 추정할 모수의 수

- Prior와 조건부 평균의 추정량 갯수는 동일하고 조건부 분산

$$CK + \frac{K(K+1)}{2} + C - 1$$

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

### 3.3 Gaussian Naïve Bayes의 추정

- 주어진 자료를 이용하여 평균과 분산, 두 모수를 구하는 문제이다.

$$\Pr(\boldsymbol{\mu}, \Sigma | X, Y) = \frac{\Pr(X|\boldsymbol{\mu}, \Sigma, Y) \times \Pr(\boldsymbol{\mu}, \Sigma)}{\Pr(X, Y)} = \frac{\Pr(X|\boldsymbol{\mu}, \Sigma) \times \Pr(Y) \times \Pr(\boldsymbol{\mu}, \Sigma)}{\Pr(X, Y)}$$

- 모수 추정에는 regularization 없이 likelihood를 사용한다.
- 모수의 추정값은 각 class별로 표본을 구분하고 해당  $(x, y)$ 로 MAP를 이용하여 계산한다.

$$\max_{\pi_c, \mu_{k|c}, \sigma_{k|c}} \prod_{i \in \text{class } c} \pi_c \prod_{k=1}^K \frac{1}{(2\pi)^{K/2} \sigma_{k|c}} \exp\left(-\frac{1}{2} \left(\frac{x_{ik|c} - \mu_{k|c}}{\sigma_{k|c}}\right)^2\right)$$

$$\text{subject to } \sum_{c=1}^C \pi_c = 1$$

$$\max_{\pi_c, \mu_{k|c}, \sigma_{k|c}} n_c \ln \pi_c - \frac{n_c K^2}{2} \ln 2\pi - n_c \left( \sum_{k=1}^K \ln \sigma_{k|c} \right) - \frac{1}{2} \sum_{i \in \text{class } c} \sum_{k=1}^K \left( \frac{x_{ik|c} - \mu_{k|c}}{\sigma_{k|c}} \right)^2$$
$$\text{subject to } \sum_{c=1}^C \pi_c = 1$$

- 일계조건을 이용하여 해를 구한다.

$$\mu_{k|c} = \frac{\sum_{i \in \text{class } c} x_{ik|c}}{n_c}, \quad \sigma_{k|c}^2 = \frac{\sum_{i \in \text{class } c} (x_{ik|c} - \mu_{k|c})^2}{n_c}, \quad \pi_c = \frac{n_c}{n}$$
$$\mu_{k|c} = \frac{\sum_{i=1}^n \mathbb{1}(y_i = c) x_{ik}}{\sum_{i=1}^n \mathbb{1}(y_i = c)}, \quad \sigma_{k|c}^2 = \frac{\sum_{i=1}^n \mathbb{1}(y_i = c) (x_{ik} - \mu_{k|c})^2}{\sum_{i=1}^n \mathbb{1}(y_i = c)}, \quad \pi_c = \frac{\sum_{i=1}^n \mathbb{1}(y_i = c)}{n}$$

- Class 분포에 대한 prior는 표본에서 구하거나 임의의 확률을 지정하여 사용하기도 한다.

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature가 하나이면서 조건부분산
  - 4.2 Feature가 하나이면서 조건부분산
  - 4.3 Feature가 여러 개이면서 조건부분산
  - 4.4 Feature가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 3.4 Gaussian Naïve Bayes Classifier

- $X$ 에 대한 예측  $y^{\text{pred}}$ 는 앞에서 구한 posterior 극대화 문제의 해이다.

$$y^{\text{pred}} = \arg \max_c \Pr(Y = c) \prod_{k=1}^K \Pr(X_k | Y = c)$$

- 즉, 모든  $a$ 에 대해 다음 조건을 만족하면  $x$ 에 대한 '최적' 예측은  $c$ 가 된다.

$$\Pr(Y = c | X) \geq \Pr(Y = a | X)$$

- Gauss naïve Bayes 분류기의 장점은 분류기를 간단히 계산할 수 있다는 점이다.
- Gaussian naïve Bayes 모형에서 feature의 조건부 분산은 서로 다를 수도 같을 수도 있다.  
분산에 대한 조건에 따라 분류기의 형태가 선형 혹은 비선형이 된다.

## 4 Binary classifier와 decision boundary

- 이 section에서 살펴볼 decision boundary의 성질은 conditional independence와는 무관하게 성립한다.
- $y \in \{0, 1\}$ 인 binary 분류 모형에서  $y^{\text{pred}} = 1$ 인 decision region은 다음과 같다.

$$\frac{\Pr(Y = 1 | x)}{\Pr(Y = 0 | x)} = \frac{\pi_1 \Pr(x | Y = 1)}{\pi_0 \Pr(x | Y = 0)} \geq 1$$

- 위 식에 로그변환을 취해주면

$$\ln \frac{\Pr(Y = 1 | x)}{\Pr(Y = 0 | x)} = \ln \pi_1 + \ln \Pr(x | Y = 1) - \ln \pi_0 - \ln \Pr(x | Y = 0) \geq 0$$

- $\Pr(x | Y)$ 의 선택에 따라 decision boundary의 형태가 결정된다.

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 4.1 Feature 가 하나이면서 조건부분산이 동일한 경우

- $K = 1$ 인 경우 decision region의 형태를 살펴보자.

$$\begin{aligned}\ln \frac{\Pr(Y=1|x)}{\Pr(Y=0|x)} &= \ln \pi_1 + \ln \Pr(x|Y=1) - \ln \pi_0 - \ln \Pr(x|Y=0) \\ &= \ln \pi_1 + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \left( \frac{x - \mu_1}{\sigma} \right)^2 \\ &\quad - \ln \pi_0 - \ln \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2} \left( \frac{x - \mu_0}{\sigma} \right)^2 \\ &= \ln \pi_1 - \ln \pi_0 + \frac{\mu_1 - \mu_0}{\sigma^2} \cancel{x} - \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2} \geq 0\end{aligned}$$

- 정규성 가정이 만족하면 decision surface는 한 점이고, 이 점을 기준으로 feature 공간을 둘로 나눈다.

$$x > \frac{2\sigma^2(-\ln \pi_1 + \ln \pi_0) + (\mu_1^2 - \mu_0^2)}{2(\mu_1 - \mu_0)} \Leftrightarrow \text{predict } y^{\text{pred}} = 1$$

Under the Gaussian naive Bayes assumptions, the cutoff is  $x*=167.8$ .

Female	4.17
Male	5.31
Pooled standard deviation	4.94

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

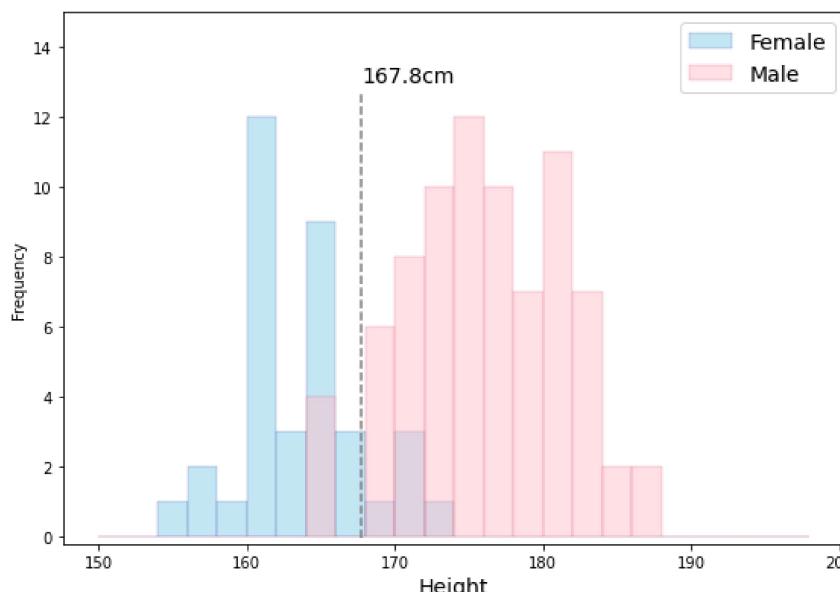
- Sample standard deviations

Class	standard deviation
Female	4.11
Male	5.28
Pooled standard deviation	4.94

```
from scipy.stats import levene  
levene(height_male, height_female, center='mean')
```

- Levene 검증의 p-value는 0.10으로 분산이 동일하다는 귀무가설을 기각하지 못한다.

Figure 6. Decision Boundary



## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 4.2 Feature 가 하나이면서 조건부분산이 다른 경우

- $K = 1$ 인 경우 decision surface의 형태를 살펴보자.

$$\begin{aligned} \ln \frac{\Pr(Y=1|x)}{\Pr(Y=0|x)} &= \ln \pi_1 + \ln \Pr(x|Y=1) - \ln \pi_0 - \ln \Pr(x|Y=0) \\ &= \ln \pi_1 + \ln \frac{1}{\sqrt{2\pi\sigma_1^2}} - \frac{1}{2} \left( \frac{x - \mu_1}{\sigma_1} \right)^2 \\ &\quad - \ln \pi_0 - \ln \frac{1}{\sqrt{2\pi\sigma_0^2}} + \frac{1}{2} \left( \frac{x - \mu_0}{\sigma_0} \right)^2 \end{aligned}$$

- $y^{\text{pred}} = 1$ 인 decision region은 다음 조건을 만족시키는  $x$ 이다  $\$ \$ \ln \pi_1 - \ln \pi_0 - \frac{1}{2}(\ln 2\pi\sigma_1^2 + \ln 2\pi\sigma_0^2) + \frac{1}{2}(\ln(\sigma_1^2 - \sigma_0^2)) \leq 0 \$ \$$
- $\ln \pi_1 - \ln \pi_0 - (\ln \sigma_1 - \ln \sigma_0) - \frac{1}{2} \left[ \left( \frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) x^2 - 2 \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2} \right) x + \left( \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2} \right) \right] \geq 0$

- Decision boundary는  $x$ 의 이차함수이다.
- 정규분포가 아니거나 조건부 분산이 다르다면  $\Pr(Y=1|x^*) = \Pr(Y=0|x^*)$  을 만족시키는  $x^*$ 는 일반적으로 유일하게 결정되지 않으며, 따라서 decision region을 두 개의 볼록집합으로 표시하지 못할 수 있다.

'Solutions to the condition of decision boundary are  $x^*=167.6, 118.2.$ '

## Gaussian Naïve Bayes 모형과 비교

```
print(confusion_matrix(y, x.height <167.6))

gnb = GaussianNB(priors=1-prior_sample)
gnb.fit(x, y)
y_hat = gnb.predict(x)
print(confusion_matrix(y, y_hat))
```

## Contents

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

One feature: Height Only

Confusion Matrix by hand calculation :  
[[31 5]  
[ 4 75]]

Confusion Matrix by GaussianNB with uniform prior, logistic regression:  
[[32 4]  
[ 7 72]]

Confusion Matrix by GaussianNB with fitted prior:  
[[31 5]  
[ 4 75]]

	precision	recall	f1-score	support
Female	0.89	0.86	0.87	36
Male	0.94	0.95	0.94	79
accuracy			0.92	115
macro avg	0.91	0.91	0.91	115
weighted avg	0.92	0.92	0.92	115

### 4.3 Feature 가 여러 개이면서 조건부분산이 다른 경우

$$f(X = \mathbf{x}|Y = c) = \frac{1}{(2\pi)^{N/2} |\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\right), \quad c = 0, 1$$

- $y^{\text{pred}} = 1$ 인 decision region은  $\Pr(Y = 1|\mathbf{x}) \geq \Pr(Y = 0|\mathbf{x})$  을 만족하는  $\mathbf{x}$ 이므로

$$\pi_1 \Pr(\mathbf{x}|Y = 1) - \pi_0 \Pr(\mathbf{x}|Y = 0) \geq 0$$

$$\ln \pi_1 - \frac{1}{2} \ln \det(\Sigma_1) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \ln \pi_0 + \frac{1}{2} \ln \det(\Sigma_0) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) \geq 0$$

$$-\mathbf{x}^T (\Sigma_1^{-1} - \Sigma_0^{-1}) \mathbf{x} + 2(\boldsymbol{\mu}_1^T \Sigma_1^{-1} - \boldsymbol{\mu}_0^T \Sigma_0^{-1}) \mathbf{x} - (\boldsymbol{\mu}_1^T \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \Sigma_0^{-1} \boldsymbol{\mu}_0) + 2 \ln \frac{\pi_1}{\pi_0} - \ln \frac{\det(\Sigma_1)}{\det(\Sigma_0)} \geq 0$$

$$\mathbf{x}^T Q \mathbf{x} + 2\mathbf{w}^T \mathbf{x} + b \leq 0 \Leftrightarrow \text{predict } y^{\text{pred}} = 1$$

- Decision boundary는 quadratic function이며, 2차원 평면에선 타원, 포물선, 쌍곡선의 형태이다.
- Standard deviations of each feature per class, (n\_classes, n\_features)

```
np.sqrt(gnb.var_) # height, weight
array([[5.2766442, 9.74139254],
       [4.11289378, 3.73636928]])
```

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature가 하나이면서 조건부분산
  - 4.2 Feature가 하나이면서 조건부분산
  - 4.3 Feature가 여러 개이면서 조건부분산
  - 4.4 Feature가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시작화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

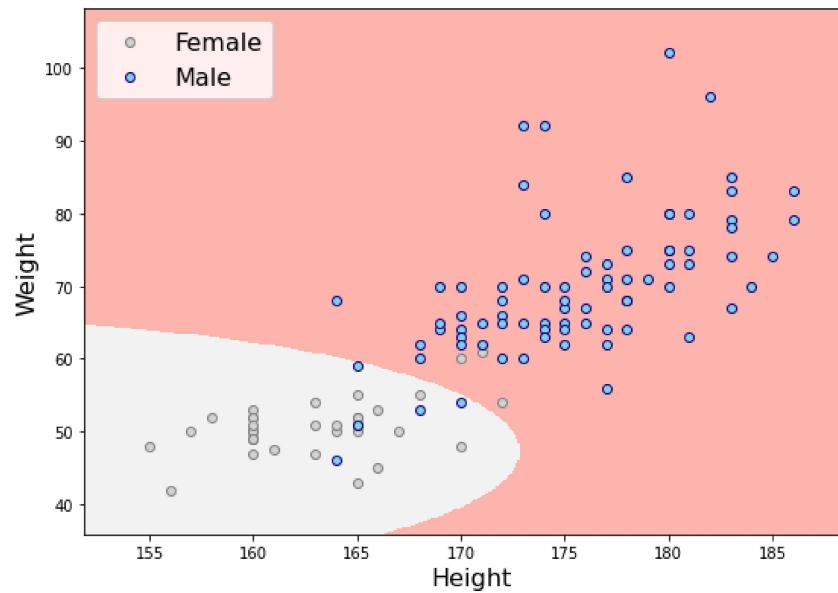
Confusion Matrix by GaussianNB with two features, height and weight:

$\begin{bmatrix} 33 & 3 \\ 5 & 74 \end{bmatrix}$

	precision	recall	f1-score	support
Male	0.96	0.94	0.95	79
Female	0.87	0.92	0.89	36
accuracy			0.93	115
macro avg	0.91	0.93	0.92	115
weighted avg	0.93	0.93	0.93	115

C:\Users\K5\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but GaussianNB was fitted with feature names  
warnings.warn(

Figure 7. 2D Decision Region of GaussianNB



## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 4.4 Feature 가 여러 개이면서 조건부분산이 같은 경우

- 조건부 분산이 다른 경우에서 구한 조건을 이용하자.

$$\ln \frac{\Pr(Y = 1|x)}{\Pr(Y = 0|x)} = \ln \pi_1 + \ln \Pr(x|Y = 1) - \ln \pi_0 - \ln \Pr(x|Y = 0) \geq 0$$

\$\$\mathbf{x}^\top \Sigma^{-1} (\mu\_1 - \mu\_0) \geq 0  
따라서  $\mathbf{x}^\top \Sigma^{-1} (\mu_1 - \mu_0) \geq 0$   
 $\mathbf{w}^\top \mathbf{x} + b \leq 0 \Leftrightarrow \text{predict } y^{\text{pred}} = 1$  \$\$

## 4.5 Logistic Regression과의 관계

- Logistic regression은 조건부분산이 모두 같다고 가정한다.

$$\begin{aligned}\ln \Pr(Y = 0|\mathbf{x}) &= \frac{\Pr(\mathbf{x}, Y = 0)}{\Pr(\mathbf{x}, Y = 0) + \Pr(\mathbf{x}, Y = 1)} \\&= \frac{\pi_0 \Pr(\mathbf{x}|Y = 0)}{\pi_0 \Pr(\mathbf{x}|Y = 0) + \pi_1 \Pr(\mathbf{x}|Y = 1)} \\&= \left\{ 1 + \frac{\pi_1}{\pi_0} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu_0)^\top \Sigma^{-1} (\mathbf{x} - \mu_0) + \frac{1}{2} (\mathbf{x} - \mu_1)^\top \Sigma^{-1} (\mathbf{x} - \mu_1) \right] \right\}^{-1} \\&= \left\{ 1 + \exp \left[ \ln \frac{\pi_1}{\pi_0} + (\mu_1 - \mu_0)^\top \Sigma^{-1} \mathbf{x} + \frac{1}{2} (\mu_1^\top \Sigma^{-1} \mu_1 - \mu_0^\top \Sigma^{-1} \mu_0) \right] \right\}^{-1} \\&= \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x} - b)}\end{aligned}$$

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산이 동일한 경우
  - 4.2 Feature 가 하나이면서 조건부분산이 다른 경우
  - 4.3 Feature 가 여러 개이면서 조건부분산이 동일한 경우
  - 4.4 Feature 가 여러 개이면서 조건부분산이 다른 경우
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

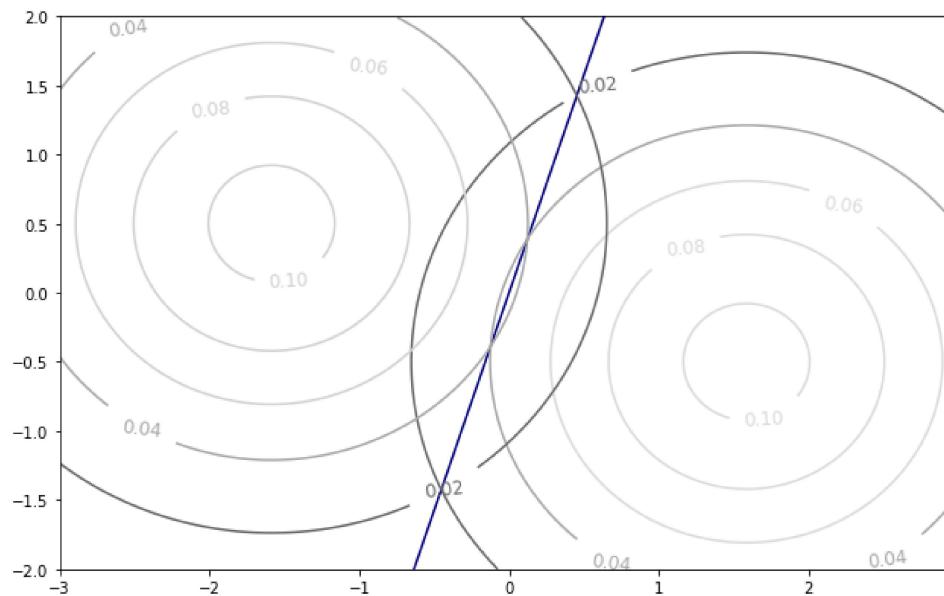
## 5 Gaussian Bayes - 시각화

- 앞의 결과를 decision boundary의 형태를 기준으로 정리하면
    - 조건부 독립성과 무관하게 조건부 공분산이 같으면 decision boundary는 초평면
    - 조건부 독립성과 무관하게 조건부 공분산이 다르면 quadratic form
  - 조건부 독립성 조건은 추정해야 할 모수의 수를 획기적으로 줄여주는 역할
  - 가정이 성립하지 않으면 Naïve Bayes 모형의 예측결과는 이를 감안하여 해석
  - 다음의 상황에서 decision boundary가 어떻게 결정되는지 등고선 그림을 이용해 알아본다.
1. 조건부공분산이 동일하며 조건부독립은 성립하거나 성립하지 않는 경우
  2. 조건부공분산이 다르며 조건부독립은 성립하거나 성립하지 않는 경우

### 5.1 조건부공분산이 동일하며 조건부독립은 성립하거나 성립하지 않는 경우

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_K^2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_K^2 \end{bmatrix}$$

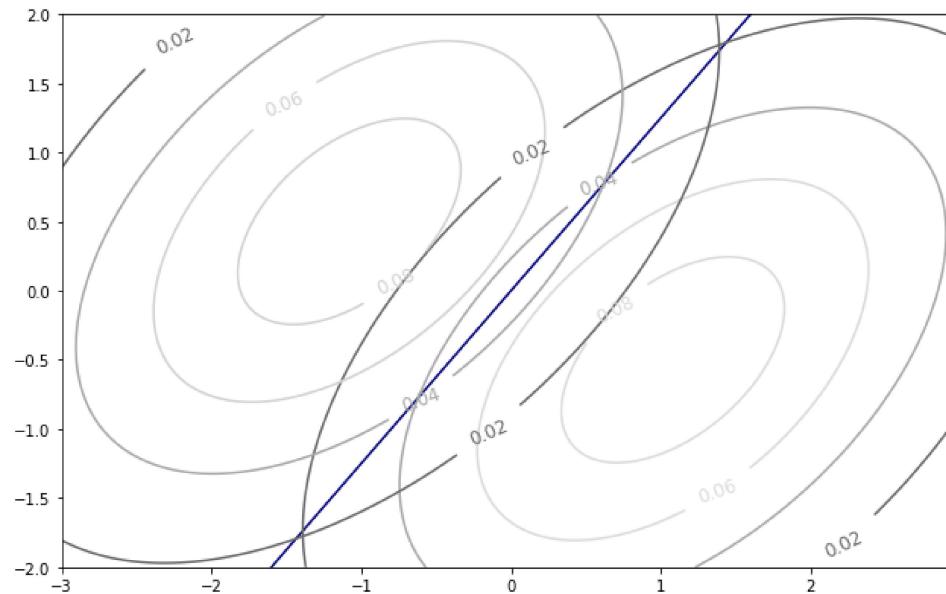
Figure 8-1. Common Diagonal Covariance Matrix



## Contents

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature가 하나이면서 조건부분산이 다른 경우
  - 4.2 Feature가 하나이면서 조건부분산이 같은 경우
  - 4.3 Feature가 여러 개이면서 조건부분산이 다른 경우
  - 4.4 Feature가 여러 개이면서 조건부분산이 같은 경우
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

Figure 8-2. Common Covariance Matrix and Correlated Features



### 5.2 조건부공분산이 다르며 조건부독립성은 성립하거나 성립하지 않는 경우

$$\Sigma_c = \begin{bmatrix} \sigma_{1|c}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2|c}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{K|c}^2 \end{bmatrix}, \quad \Sigma_e = \begin{bmatrix} \sigma_{12|c}^2 & \sigma_{12|c} & \cdots & \sigma_{1K|c} \\ \sigma_{21|c} & \sigma_{21|c}^2 & \cdots & \sigma_{2K|c} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1|c} & \sigma_{K2|c} & \cdots & \sigma_{KK|c}^2 \end{bmatrix}$$

- Sklearn의 Naïve Bayesian 모형들은 조건부독립성은 성립하지만 조건부공분산은 서로 다를 수 있다고 가정한다.

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naï
  - ▼ 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bay
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning cur
  - 9.3 Gaussian Naïve Bayes vs. Logistic
  - 9.4 update and maintenance

Figure 8-3. Class-specific Diagonal Covariance Matrix: GaussianNB

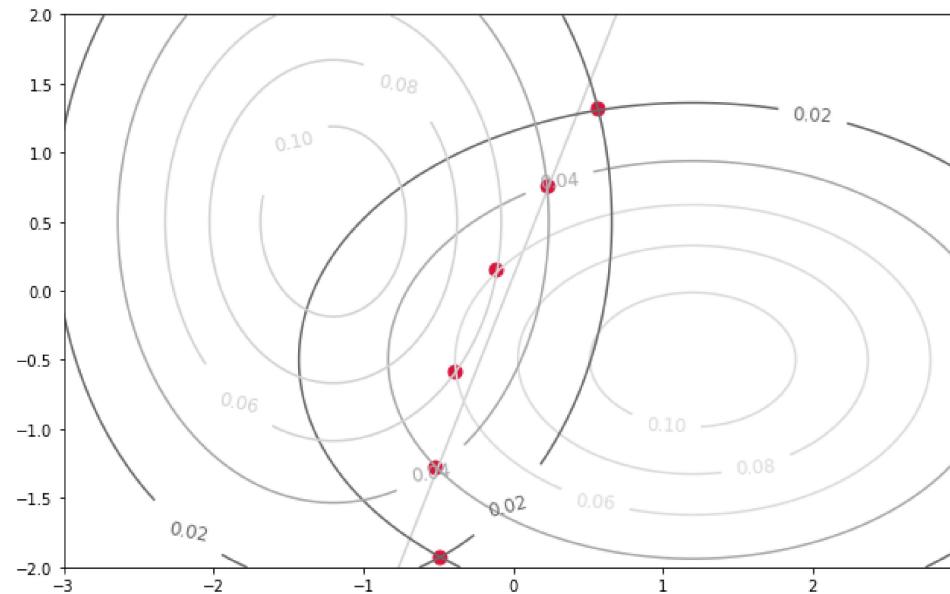
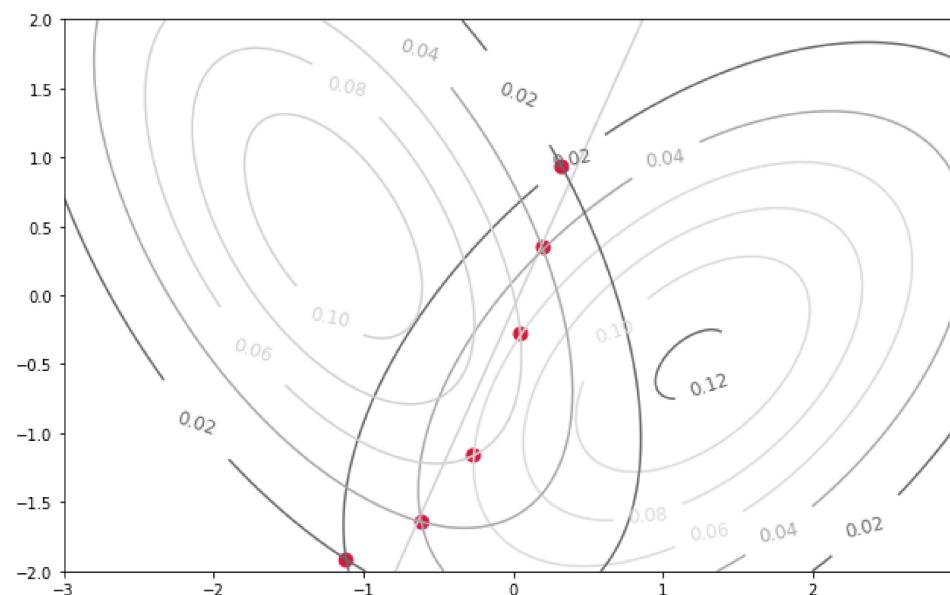


Figure 8-4. Covariance Matrix with no restriction



## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 5.3 Multiclass Gaussian Naïve Bayes

- $X_k$ 은 조건부 독립이고 모든 class에서 동일한 분산을 갖는다고 가정하자.
- Posterior는 선형함수의 지수변환으로 쓸 수 있다.

$$\Pr(Y = c|X) = \frac{\Pr(Y = c)\Pr(X|Y = c)}{\sum_{j=1}^C \Pr(Y = j)\Pr(X|Y = j)} = \frac{\exp(\mathbf{w}_c^\top \mathbf{x} + b_c)}{\sum_{j=1}^C \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)}$$

- Class의 수와 관계없이 decision boundary는 선형이다.

$$\ln \frac{\Pr(Y = c|X)}{\Pr(Y = j|X)} = \ln \frac{\exp(\mathbf{w}_c^\top \mathbf{x} + b_c)}{\exp(\mathbf{w}_j^\top \mathbf{x} + b_j)} = (\mathbf{w}_c^\top - \mathbf{w}_j^\top) \mathbf{x} + (b_c - b_j)$$

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## Posterior 유도

$$\begin{aligned}
 \Pr(Y = c|X) &= \frac{\Pr(Y = c)\Pr(X|Y = c)}{\sum_{j=1}^C (\Pr(Y = j)\Pr(X|Y = j))} \\
 &= \frac{\pi_c \prod_{k=1}^K \frac{1}{(2\pi)^{N/2}\sigma_k} \exp\left(-\frac{1}{2}\left(\frac{x_k - \mu_{k|c}}{\sigma_k}\right)^2\right)}{\sum_{j=1}^C \pi_j \prod_{k=1}^K \frac{1}{(2\pi)^{N/2}\sigma_k} \exp\left(-\frac{1}{2}\left(\frac{x_k - \mu_{k|j}}{\sigma_k}\right)^2\right)} \\
 &= \frac{\exp\left(\ln \pi_c - \frac{1}{2} \sum_{k=1}^K \left(\frac{x_k - \mu_{k|c}}{\sigma_k}\right)^2\right)}{\sum_{j=1}^C \exp\left(\ln \pi_j - \frac{1}{2} \sum_{k=1}^K \left(\frac{x_k - \mu_{k|j}}{\sigma_k}\right)^2\right)} \\
 &= \frac{1}{\sum_{j=1}^C \exp\left(\ln \pi_j - \ln \pi_c + \frac{1}{2} \sum_{k=1}^K \left(\frac{x_k - \mu_{k|j}}{\sigma_k}\right)^2 - \frac{1}{2} \sum_{k=1}^K \left(\frac{x_k - \mu_{k|c}}{\sigma_k}\right)^2\right)} \\
 &= \frac{1}{\sum_{j=1}^C \exp\left(\ln \pi_j - \ln \pi_c + \sum_{k=1}^K \frac{2x_k(\mu_{k|j} - \mu_{k|c}) + (\mu_{k|j}^2 - \mu_{k|c}^2)}{2\sigma_k}\right)} \\
 &= \frac{\exp\left(\ln \pi_c + \sum_{k=1}^K \frac{2x_k\mu_{k|c} + \mu_{k|c}^2}{2\sigma_k}\right)}{\sum_{j=1}^C \exp\left(\ln \pi_j + \sum_{k=1}^K \frac{2x_k\mu_{k|j} + \mu_{k|j}^2}{2\sigma_k}\right)} \\
 &= \frac{\exp(\mathbf{w}_c^\top \mathbf{x} + b_c)}{\sum_{j=1}^C \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)}
 \end{aligned}$$

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 6 연속확률변수의 독립성 검증

- 설문조사의 키와 몸무게가 조건부 독립이라고 볼 수 있을까?
- 각 성별 내에서 키와 몸무게가 서로 독립적인가?

### 6.1 Chi-Square test 복습

- 범주형 자료의 독립성은 Chi-Square test로 검증이 가능하다. 하지만 아직까지 일반적인 상황에서 연속변수에 적용할 수 있는 검증방법은 알려져 있지 않다.

	$y_1$	$y_2$
$x_1$	$p_{11}$	$p_{12}$
$x_2$	$p_{21}$	$p_{22}$
	$p_{y1}$	$p_{y2}$

- $X$ 와  $Y$ 가 독립적이라면 관찰값(observed value)  $p_{ij}$ 와 기대값(expected value)  $E_{ij} = p_{x_i} \times p_{y_j}$ 는 서로 다르지 않아야 한다.
- $df = (C_X - 1)(C_Y - 1)$ ,  $C_X$ 와  $C_Y$ 를 두 확률변수의 범주수라고 하면, 검증통계량은 다음과 같다.

$$\sum_{i=1}^{C_X} \sum_{j=1}^{C_Y} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{df}^2$$

### 6.2 Chi-square test를 연속확률변수에 적용

- Chi-square test의 통계량은 계급구간을 어떻게 정하느냐에 따라서 검증통계량이 달라지므로 연속변수를 이산변수로 만들어  $\chi^2$  검증을 할 수 없다.
- Bin size를 바꾸어 보면 p-value가 따라서 '크게' 변하는 것을 확인할 수 있다.

the p-values for conditional independence of discretized height and weight are:

Female: 0.022, Male: 0.364

## Contents ⚙

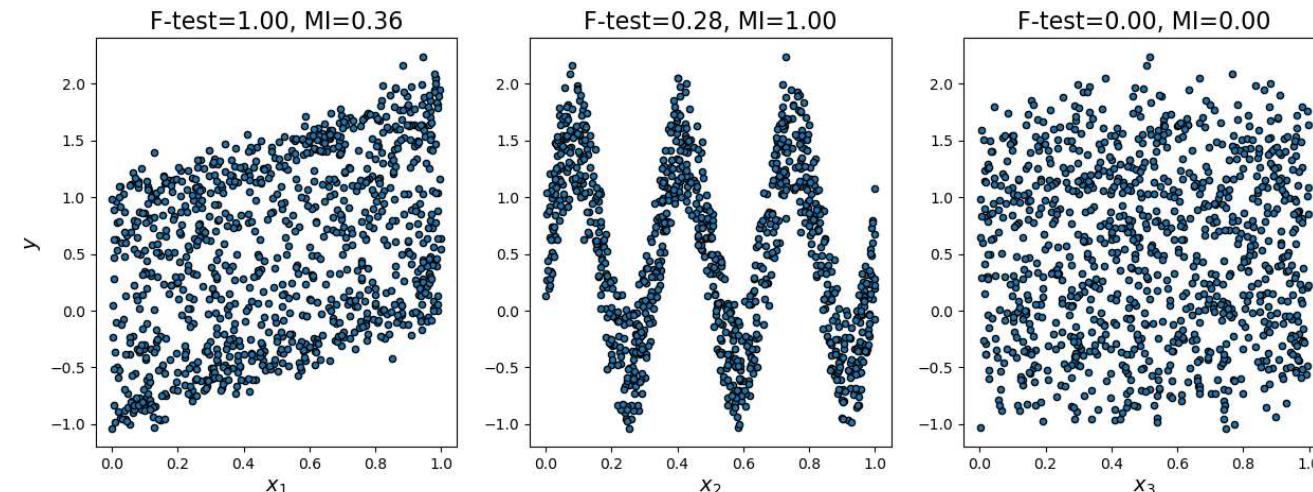
- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 6.3 mutual information

- Scikit-learn에선 information entropy 를 이용하여 두 변수 사이의 정보량을 비교하는 함수를 제공한다.
  - 다중공선성 문제를 해결하기 위해 상관관계가 높은 독립변수를 골라 제거하기 위해 사용하는 방법으로 독립성 검증에 사용할 수 있다. 하지만 Naïve Bayes가 요구하는 독립성과는 다르므로 참고 목적으로 사용
  - 비선형 관계도 어느정도 설명할 수 있다는 장점이 있지만 해당 통계량의 분포는 알려져 있지 않다.  
따라서 귀무가설 하에서 통계량의 분포를 만들어 검증통계량과 비교해야 한다.
- ```
sklearn.feature_selection.mutual_info_regression(X, y, discrete_features='auto', n_neighbors=3, copy=True, random_state=None)
```
- X: array-like or sparse matrix, shape (n\_samples, n\_features), Feature matrix.
  - y: array-like of shape (n\_samples,), Continuous target vector.
  - discrete\_features: {'auto', bool, array-like}, default='auto',  
If bool, then determines whether to consider all features discrete or continuous. If array, then it should be either a boolean mask with shape (n\_features,) or array with indices of discrete features. If 'auto', it is assigned to False for dense X and to True for sparse X.

보통은 그냥 사용한다. 참고만 하자.

Figure 4. Mutual Information



## Contents

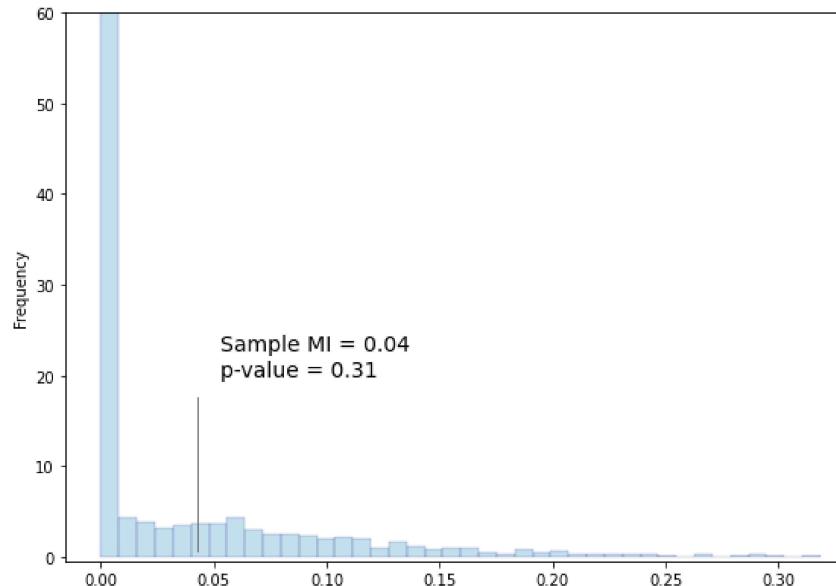
- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature가 하나이면서 조건부분산
  - 4.2 Feature가 하나이면서 조건부분산
  - 4.3 Feature가 여러 개이면서 조건부분산
  - 4.4 Feature가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

p-value for conditional independence 0.250 for female  
p-value for conditional independence 0.050 for male

[Toggle show/hide](#)

[Toggle show/hide](#)

Figure 5. Distribution of Mutual Information under the Null



- 검증은 두 변수가 독립적이 아니라고 보기 어렵다고 결론을 내리기 위한 것이 아니라
- 독립성을 가정한 분석방법에 대한 정당성을 부여하기 위한 한가지 방법

## 7 Palmer Penguins

- Iris dataset의 경우 Setosa는 완전분리가 가능하기 때문에 multinomial regression 모형을 적용하기 적합하지 않다.
- 최근 자주 볼 수 있는 남극 penguin 자료는 multinomial regression 모형을 연습하기 좋은 조건들을 갖추고 있다.

Figure 9. Palmer Penguins



source: @allison\_horst <https://github.com/allisonhorst/penguins> (<https://github.com/allisonhorst/penguins>)

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 7.1 Palmer Penguins dataset

- 이 자료는 2007 - 2009 사이에 Dr. Kristen Gorman가 수집한 것으로 세 종류의 남극펭귄에 대한 344개 관찰값이다.
  - species: penguin species (Chinstrap, Adélie, or Gentoo)
  - culmen\_length\_mm: culmen length (mm)
  - culmen\_depth\_mm: culmen depth (mm)
  - flipper\_length\_mm: flipper length (mm)
  - body\_mass\_g: body mass (g)
  - island: island name (Dream, Torgersen, or Biscoë) in the Palmer Archipelago (Antarctica)
  - sex: penguin sex
- Chinstrap의 경우 단일 feature로는 거의 구분이 불가능하지만, 두 feature를 동시에 고려하면 상당한 수준까지 분리가 가능하다.

## 7.2 Sample distribution

- 총 344개의 관찰값을 포함하고 있으며 3,339 두 개의 표본은 종과 지역을 제외하곤 모두 missing value이다.
- 위 두 표본을 포함하여 총 11개의 표본에서 암/수에 대한 정보가 빠져있다.
- 표본번호 3과 339, 그리고 sex 변수를 삭제하고 분석한다.

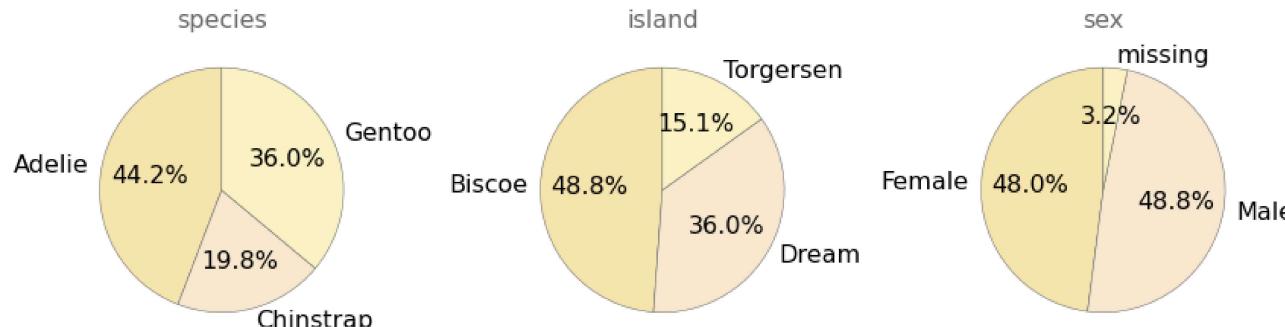
Gaussian Naïve Bayes 모형이므로 Categorical feature를 포함시킬 수 없다.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   species     344 non-null    object 
 1   island      344 non-null    object 
 2   bill_length  342 non-null    float64
 3   bill_depth   342 non-null    float64
 4   flipper_length  342 non-null    float64
 5   body_mass    342 non-null    float64
 6   sex          333 non-null    object 
dtypes: float64(4), object(3)
memory usage: 18.9± KB
```

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 물무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

Figure 10. Categorical Variables



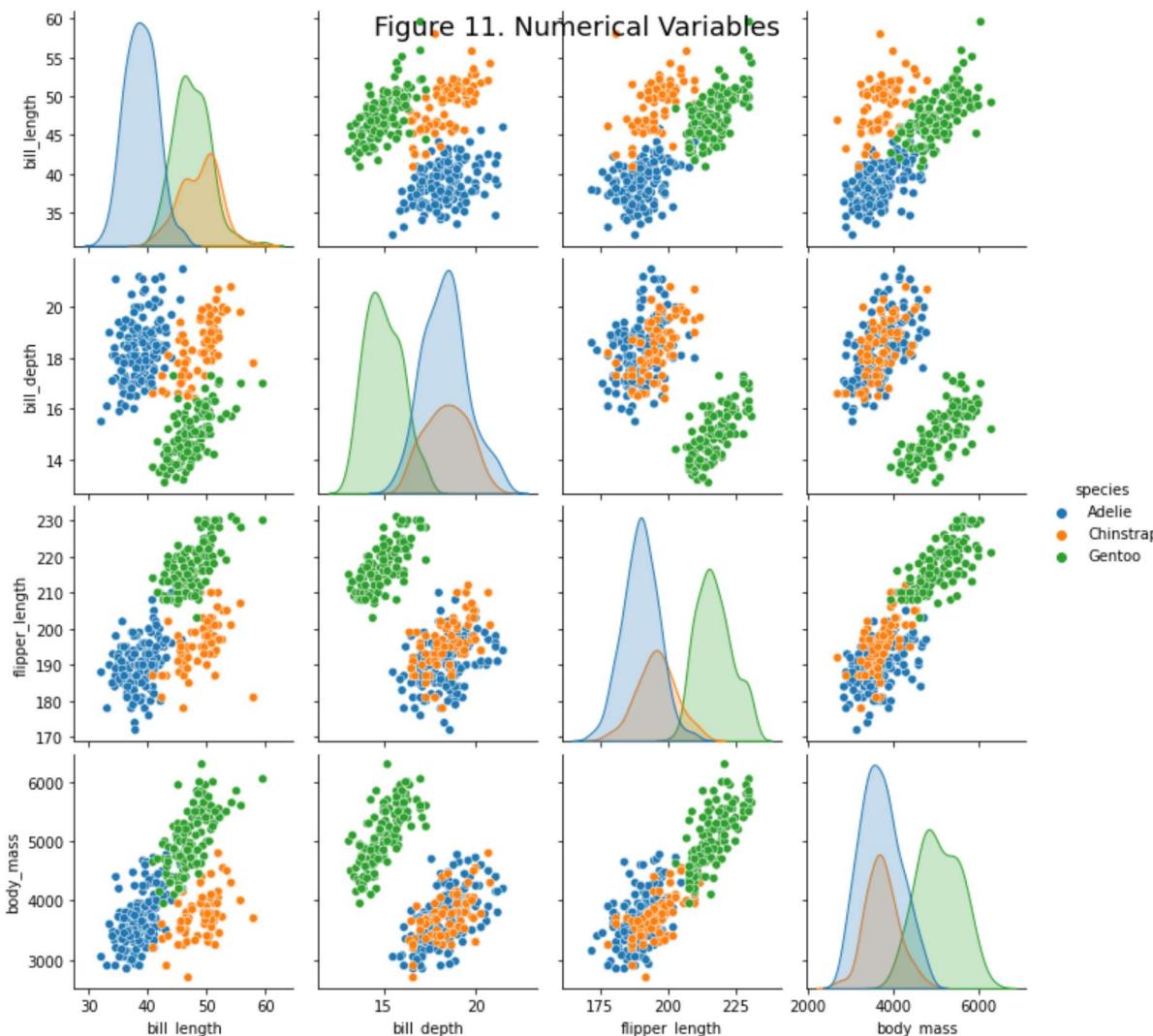
### 7.3 Data Cleaning

- Target인 'species'를 제외하고 categorical variable은 사용할 수 없으므로 분석에서 제외한다.
- 종과 서식지를 제외한 모든 정보들이 missing인 표본이 2개, sex 만 확인할 수 없는 표본이 9개 있다.
- sex 는 logistic regression으로 imputation을 하여 채워넣고, 이후 어떤 역할을 하는지 살펴보자.
  - Gaussian Naïve Bayes에서는 확인할 수 없지만 성별에 따라 칫수에 체계적인 차이가 있을 수 있다.

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve
  - ▼ 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

Figure 11. Numerical Variables



- 종과 서식지를 제외한 모든 정보들이 missing인 2개의 표본은 삭제한다.
- Logistic regression으로 성별을 예측하고 예측값으로 성별을 대체한다.

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature가 하나이면서 조건부분산
  - 4.2 Feature가 하나이면서 조건부분산
  - 4.3 Feature가 여러 개이면서 조건부분산
  - 4.4 Feature가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

|     | species | island    | bill_length | bill_depth | flipper_length | body_mass | sex |
|-----|---------|-----------|-------------|------------|----------------|-----------|-----|
| 3   | Adelie  | Torgersen | NaN         | NaN        | NaN            | NaN       | NaN |
| 8   | Adelie  | Torgersen | 34.1        | 18.1       | 193.0          | 3475.0    | NaN |
| 9   | Adelie  | Torgersen | 42.0        | 20.2       | 190.0          | 4250.0    | NaN |
| 10  | Adelie  | Torgersen | 37.8        | 17.1       | 186.0          | 3300.0    | NaN |
| 11  | Adelie  | Torgersen | 37.8        | 17.3       | 180.0          | 3700.0    | NaN |
| 47  | Adelie  | Dream     | 37.5        | 18.9       | 179.0          | 2975.0    | NaN |
| 246 | Gentoo  | Biscoe    | 44.5        | 14.3       | 216.0          | 4100.0    | NaN |
| 286 | Gentoo  | Biscoe    | 46.2        | 14.4       | 214.0          | 4650.0    | NaN |
| 324 | Gentoo  | Biscoe    | 47.3        | 13.8       | 216.0          | 4725.0    | NaN |
| 336 | Gentoo  | Biscoe    | 44.5        | 15.7       | 217.0          | 4875.0    | NaN |
| 339 | Gentoo  | Biscoe    | NaN         | NaN        | NaN            | NaN       | NaN |

|     | species | island    | bill_length | bill_depth | flipper_length | body_mass | sex |
|-----|---------|-----------|-------------|------------|----------------|-----------|-----|
| 8   | Adelie  | Torgersen | 34.1        | 18.1       | 193.0          | 3475.0    | NaN |
| 9   | Adelie  | Torgersen | 42.0        | 20.2       | 190.0          | 4250.0    | NaN |
| 10  | Adelie  | Torgersen | 37.8        | 17.1       | 186.0          | 3300.0    | NaN |
| 11  | Adelie  | Torgersen | 37.8        | 17.3       | 180.0          | 3700.0    | NaN |
| 47  | Adelie  | Dream     | 37.5        | 18.9       | 179.0          | 2975.0    | NaN |
| 246 | Gentoo  | Biscoe    | 44.5        | 14.3       | 216.0          | 4100.0    | NaN |
| 286 | Gentoo  | Biscoe    | 46.2        | 14.4       | 214.0          | 4650.0    | NaN |
| 324 | Gentoo  | Biscoe    | 47.3        | 13.8       | 216.0          | 4725.0    | NaN |
| 336 | Gentoo  | Biscoe    | 44.5        | 15.7       | 217.0          | 4875.0    | NaN |

```
array([[149, 16],  
       [14, 154]], dtype=int64)
```

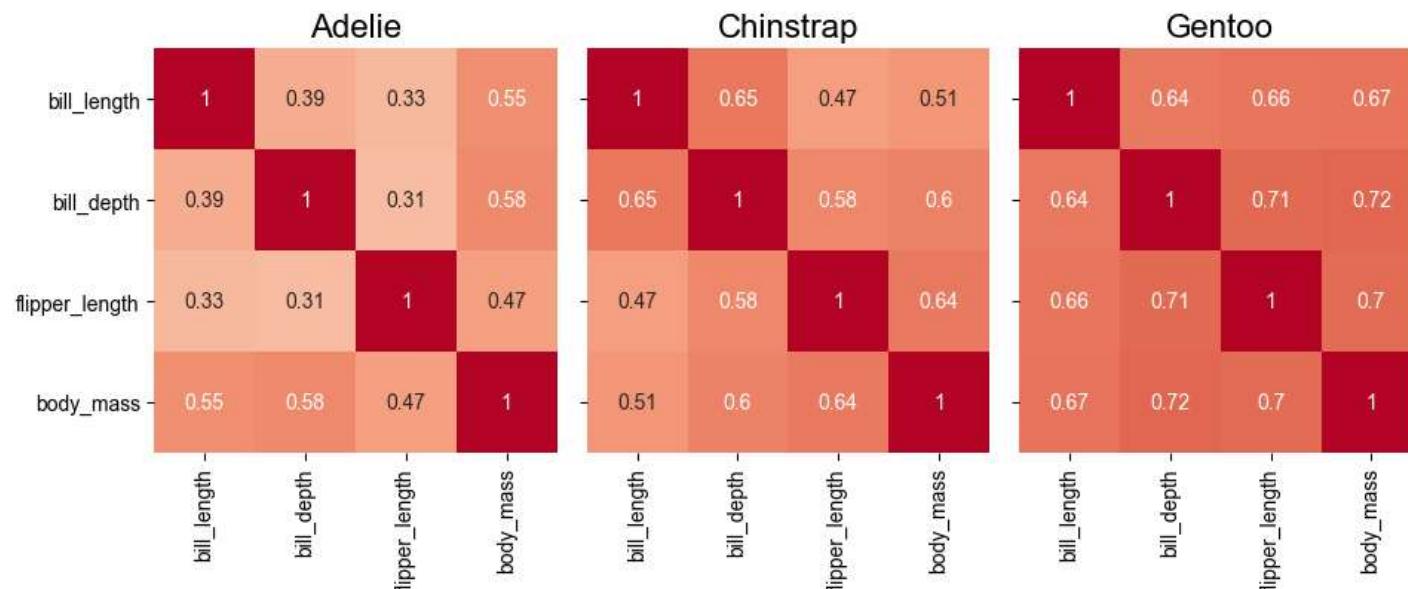
## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 물무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하여 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 7.4 조건부 분포와 조건부 독립성

- 조건부 독립성은 확인하기 어렵다.
  - 앞의 scatter plot을 보면 각 변수 사이에 어느 정도의 상관관계가 있다.
  - 특히 Gentoo의 경우에는 변수들 사이에 상당한 수준의 상관관계가 있지만 다른 종들과 상당히 분리가 되어있어 다중공선성 문제는 크지 않을 것으로 보인다.
- 조건부분포에서 정규분포와 너무 다르면 GaussianNB의 추정결과를 신뢰하기 어렵다.
- 다음의 kernel density를 보면 정규분포라고 보긴 어렵지만 모두 어느 정도의 대칭적이므로 GaussianNB를 적용해 보자.

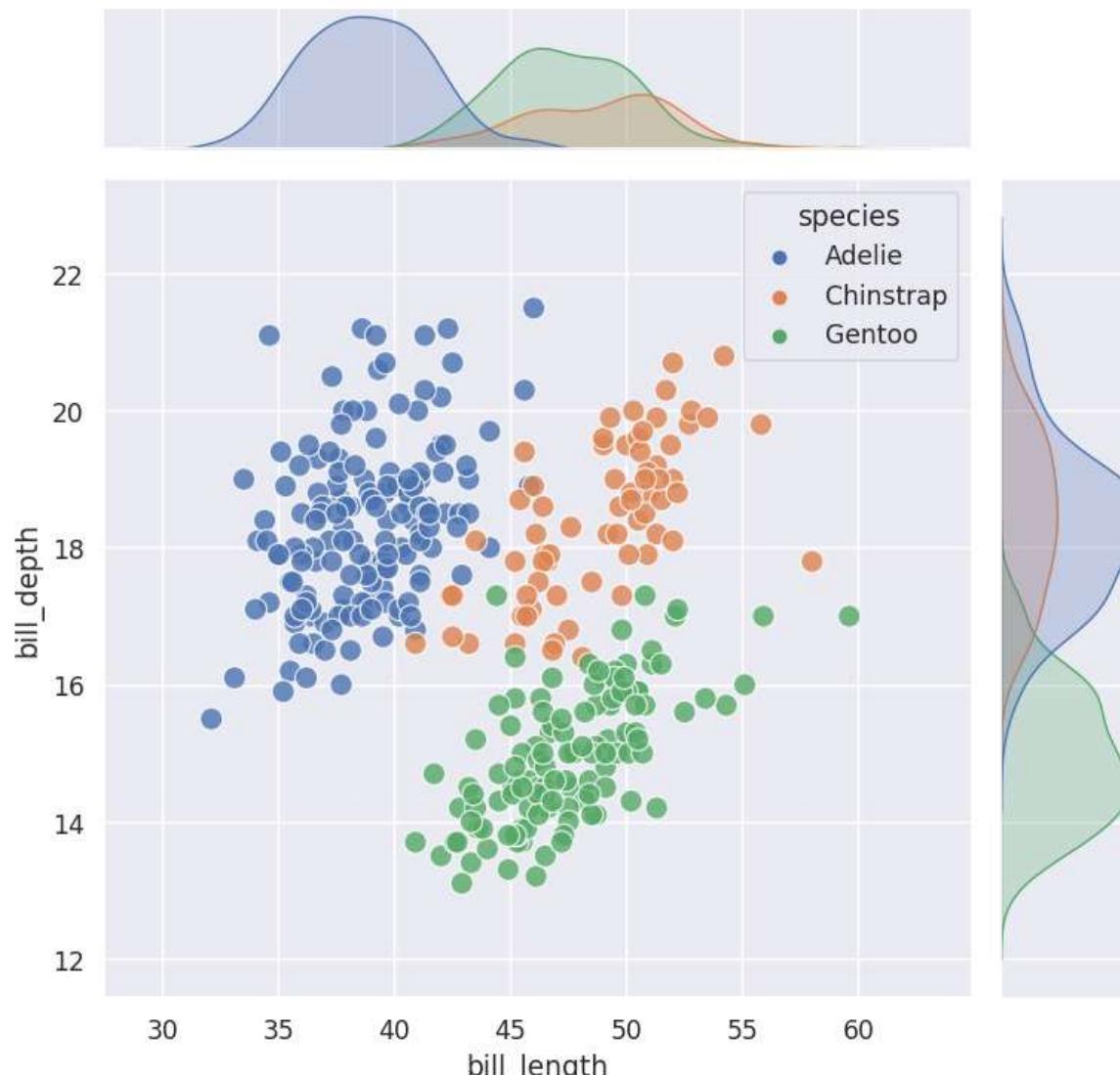
Figure 12. Conditional Independence



## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naï
  - ▼ 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bay
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning cur
  - 9.3 Gaussian Naïve Bayes vs. Logistic
  - 9.4 update and maintenance

Figure 13. Joint density of bill length and bill depth



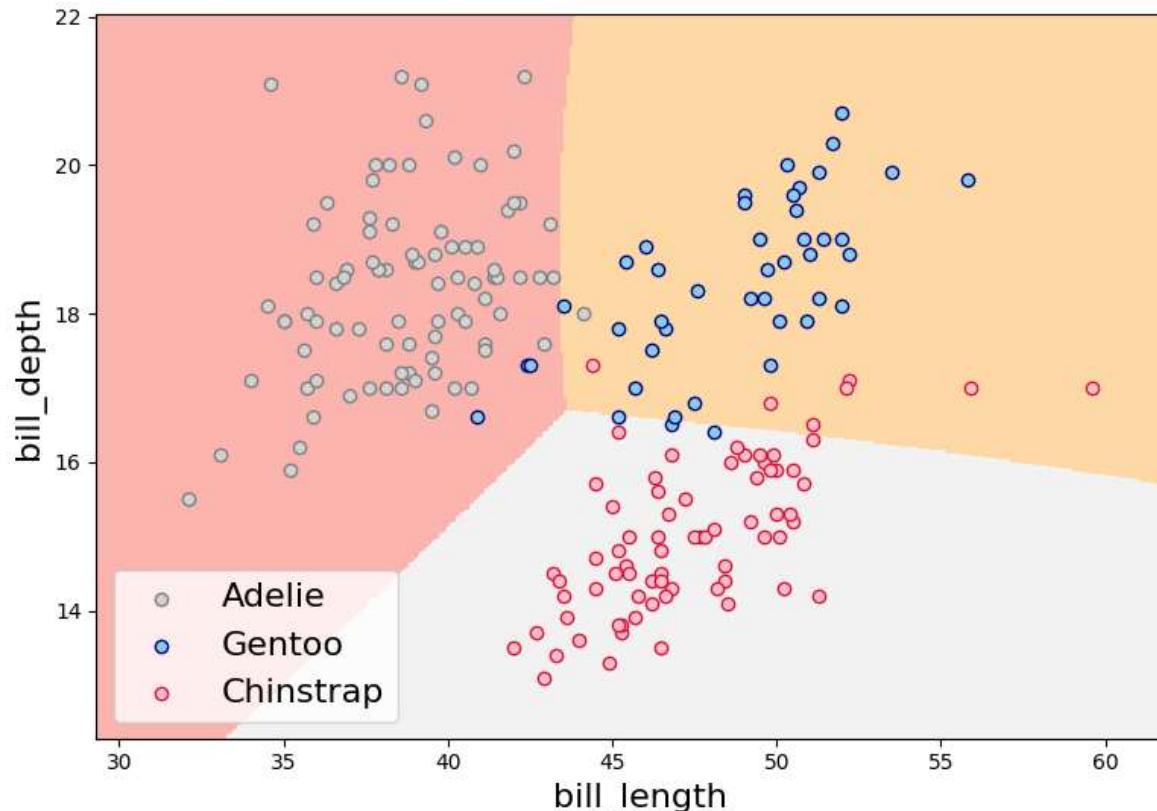
```
{'bill_length': 8.484394986868142e-09,  
 'bill_depth': 3.124442564963218e-10,  
 'flipper_length': 2.0885338150804494e-15,  
 'body_mass': 3.0689875436374184e-07}
```

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시작화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

C:\Users\K5\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but GaussianNB was fitted with feature names  
warnings.warn(

Figure 14. Decision Region of GaussianNB for Penguin dataset



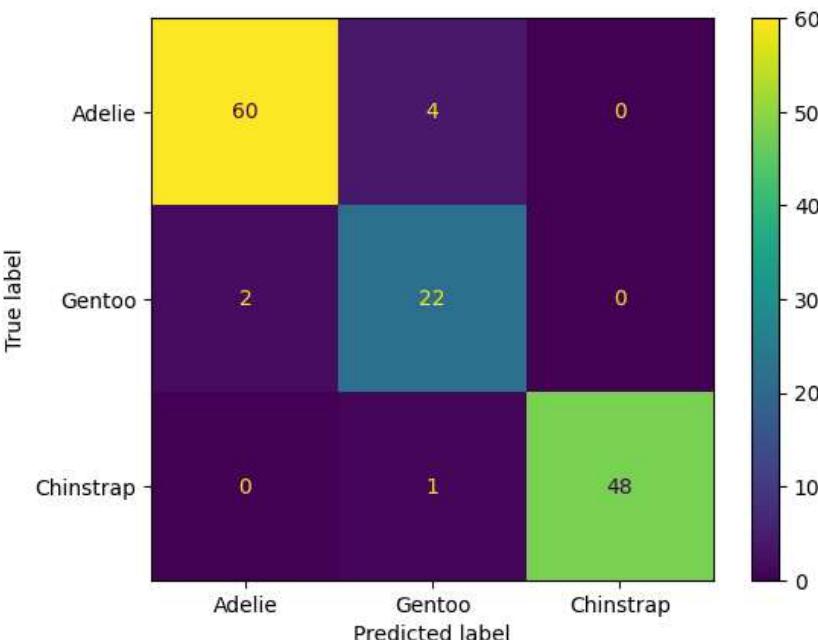
## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

precision recall f1-score support

|              |      |      |      |     |
|--------------|------|------|------|-----|
| Adelie       | 0.97 | 0.94 | 0.95 | 64  |
| Chinstrap    | 0.81 | 0.92 | 0.86 | 24  |
| Gentoo       | 1.00 | 0.98 | 0.99 | 49  |
| accuracy     |      |      | 0.95 | 137 |
| macro avg    | 0.93 | 0.94 | 0.93 | 137 |
| weighted avg | 0.95 | 0.95 | 0.95 | 137 |

C:\Users\K5\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function `plot\_confusion\_matrix` is deprecated; Function `plot\_confusion\_matrix` is deprecated in 1.0 and will be removed in 1.2. Use one of the class methods: ConfusionMatrixDisplay.from\_predictions or ConfusionMatrixDisplay.from\_estimator.  
warnings.warn(msg, category=FutureWarning)



- 모든 연속변수들을 고려해도 예측력엔 별다른 차이가 없다.
- bill\_length 와 bill\_depth 가 대부분의 변화를 설명하고 있다.

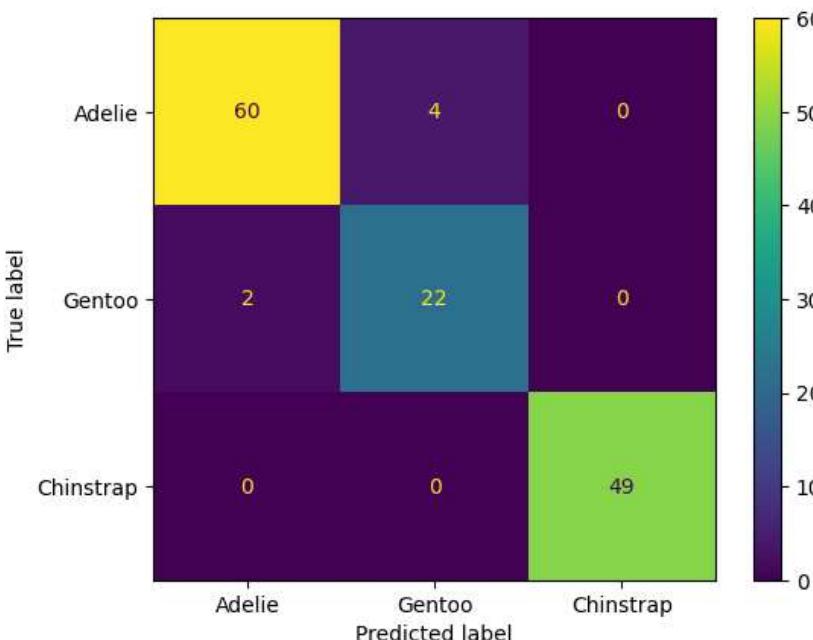
## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

precision recall f1-score support

|              |      |      |      |     |
|--------------|------|------|------|-----|
| Adelie       | 0.97 | 0.94 | 0.95 | 64  |
| Chinstrap    | 0.85 | 0.92 | 0.88 | 24  |
| Gentoo       | 1.00 | 1.00 | 1.00 | 49  |
| accuracy     |      |      | 0.96 | 137 |
| macro avg    | 0.94 | 0.95 | 0.94 | 137 |
| weighted avg | 0.96 | 0.96 | 0.96 | 137 |

C:\Users\K5\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function `plot\_confusion\_matrix` is deprecated; Function `plot\_confusion\_matrix` is deprecated in 1.0 and will be removed in 1.2. Use one of the class methods: ConfusionMatrixDisplay.from\_predictions or ConfusionMatrixDisplay.from\_estimator.  
warnings.warn(msg, category=FutureWarning)



- 동일한 feature set으로도 Logistic regression의 예측은 현저히 좋다.
- 표본의 balance가 어느정도 기울어져 있고 정규분포 가정을 만족하지 못한 결과로 보인다.

## Contents

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

precision recall f1-score support

|           |      |      |      |    |
|-----------|------|------|------|----|
| Adelie    | 1.00 | 1.00 | 1.00 | 87 |
| Chinstrap | 1.00 | 1.00 | 1.00 | 44 |
| Gentoo    | 1.00 | 1.00 | 1.00 | 74 |

accuracy

macro avg

weighted avg

Logistic Regression, test dataset

precision recall f1-score support

|           |      |      |      |    |
|-----------|------|------|------|----|
| Adelie    | 1.00 | 0.98 | 0.99 | 64 |
| Chinstrap | 0.92 | 1.00 | 0.96 | 24 |
| Gentoo    | 1.00 | 0.98 | 0.99 | 49 |

accuracy

macro avg

weighted avg

## Confusion Matrices for Logistic Regression

Training Dataset

| True label |    | Predicted label |        |           |
|------------|----|-----------------|--------|-----------|
|            |    | Adelie          | Gentoo | Chinstrap |
| Adelie     | 87 | 0               | 0      |           |
| Gentoo     | 0  | 44              | 0      |           |
| Chinstrap  | 0  | 0               | 74     |           |

Test Dataset

| True label |    | Predicted label |        |           |
|------------|----|-----------------|--------|-----------|
|            |    | Adelie          | Gentoo | Chinstrap |
| Adelie     | 63 | 1               | 0      |           |
| Gentoo     | 0  | 24              | 0      |           |
| Chinstrap  | 0  | 1               | 48     |           |

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
    - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## Correlation Coefficients of Features

|                  | bill_length | bill_depth | flipper_length | body_mass | island_Dream | island_Torgersen | sex_Male |
|------------------|-------------|------------|----------------|-----------|--------------|------------------|----------|
| bill_length      | 1           | -0.24      | 0.66           | 0.6       | 0.034        | -0.38            | 0.35     |
| bill_depth       | -0.24       | 1          | -0.58          | -0.47     | 0.46         | 0.27             | 0.37     |
| flipper_length   | 0.66        | -0.58      | 1              | 0.87      | -0.42        | -0.29            | 0.25     |
| body_mass        | 0.6         | -0.47      | 0.87           | 1         | -0.46        | -0.26            | 0.43     |
| island_Dream     | 0.034       | 0.46       | -0.42          | -0.46     | 1            | -0.320           | 0.044    |
| island_Torgersen | -0.38       | 0.27       | -0.29          | -0.26     | -0.32        | 1                | -0.022   |
| sex_Male         | 0.35        | 0.37       | 0.25           | 0.430     | 0.044        | 0.022            | 1        |

## Contents

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gausian Bayes, Gaussian Non-Naï
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분
  - 4.4 Feature 가 여러 개이면서 조건부분
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bay
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning cur
  - 9.3 Gaussian Naïve Bayes vs. Logistic
  - 9.4 update and maintenance

```
const          1205.104132
bill_length    2.348039
bill_depth     3.367422
flipper_length 6.006130
body_mass      6.155328
island_Dream   2.458932
island_Torgersen 1.730616
sex_Male       2.586694
Name: VIF, dtype: float64

drop body_mass
const          1165.826304
bill_length    2.333336
bill_depth     3.273408
flipper_length 3.919854
island_Dream   2.235179
island_Torgersen 1.679133
sex_Male       2.003864
Name: VIF, dtype: float64

drop body_mass and flipper_length
const          339.188170
bill_length    1.571883
bill_depth     2.528312
island_Dream   1.962146
island_Torgersen 1.653592
sex_Male       1.733400
Name: VIF, dtype: float64
```

## 8 Mixed Naïve Bayes

### 8.1 상이한 유형의 features

- Naïve Bayes 모형은 feature의 분포에 대한 가정을 기반으로 관련 모수를 추정
- Multinomial, Categorical, Gaussian 분포 등 feature의 특성에 따라 적합한 분포를 선택
- 조건부독립성으로 인해 Naïve Bayesian의 추정결과는 쉽게 결합이 가능하다.
- 여러 유형의 feature를 같이 사용할 때는 다음과 같은 방법으로 추정할 수 있다.
  1. 4장의 Titanic 자료에 적용한 방법으로 연속변수를 범주형으로 변환하여 Categorical Naïve Bayes로 추정
  2. 여러 모형으로 추정량을 구한 후 sklearn.ensemble 방법 적용
    - Stacking 같이 각각의 모형을 별도로 추정하여 얻은 예측확률( predict\_proba )을 feature로 사용
  3. BernoulliNB, MultinomialNB/ComplementNB, CategoricalNB, Gaussian NB의 결과를 조건부독립성 가정 하에 결합

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 8.2 독립성과 추정결과의 결합

$$\Pr(\text{survive}|\text{sex}, \text{age}) = \frac{\Pr(\text{sex}, \text{age}|\text{survive}) \times \Pr(\text{survive})}{\Pr(\text{sex}, \text{age})}$$
$$\propto \underbrace{\Pr(\text{sex}|\text{survive}) \times \Pr(\text{survive})}_{\text{CategoricalNB}} \times \underbrace{\Pr(\text{age}|\text{survive}) \times \Pr(\text{survive})}_{\text{GaussianNB}} \times \frac{1}{\Pr(\text{survive})}$$

- 이 추정방법은 ensemble 모형이 아니고 조건부독립성을 이용해 Naïve Bayes 추정결과를 "결합"한 것이다.
- Categorical variables과 continuous variable 인 경우
  - [mixed-naïve Bayes module \(<https://github.com/remykarem/mixed-naive-bayes>\)](https://github.com/remykarem/mixed-naive-bayes)는 sklearn에 의존하지 않는다.
  - [Titanic 예제 \(<https://www.kaggle.com/guruprasad91/titanic-naive-bayes>\)](https://www.kaggle.com/guruprasad91/titanic-naive-bayes)

## 8.3 결합확률분포를 이용한 방법

- 세 가지 유형의 feature set을  $X_1 = (x_{11}, x_{12}, \dots)$ ,  $X_2 = (x_{21}, x_{22}, \dots)$ ,  $X_3 = (x_{31}, x_{32}, \dots)$ 라고 하자.

$$\ln \Pr(Y|X_1, X_2, X_3) = \ln \Pr(X_1, X_2, X_3|Y) + \ln \Pr(Y) - \ln \Pr(X_1, X_2, X_3)$$

- 각 feature는 조건부독립이므로

$$\begin{aligned}\ln \Pr(Y|X_1, X_2, X_3) &= \ln \Pr(X_1|Y) + \ln \Pr(X_2|Y) + \ln \Pr(X_3|Y) + \ln \Pr(Y) - \ln \Pr(X_1, X_2, X_3) \\ &= \ln \Pr(X_1, Y) + \ln \Pr(X_2, Y) + \ln \Pr(X_3, Y) - 2 \times \ln \Pr(Y) - \ln \Pr(X_1, X_2, X_3)\end{aligned}$$

$$\ln \Pr(Y|X_1, X_2, X_3) \propto \ln \Pr(X_1, Y) + \ln \Pr(X_2, Y) + \ln \Pr(X_3, Y) - 2 \times \ln \Pr(Y)$$

- $\Pr(X, Y)$ 는 `clf._joint_log_likelihood()`,  $\ln \Pr(Y)$ 는 CategoricalNB 나 MultinomialNB에서 `clf.class_log_prior_`로 구한다.

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
    - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

## 8.4 Implementation of mixed naïve Bayes

### 1. 모형을 추정

```
gnb = GaussianNB()  
gnb.fit(x_con, y)  
  
cnb = CategoricalNB()  
cnb.fit(x_cat, y)
```

1. 추정한 로그결합확률을 더하고 log prior를 빼준다. (세 개의 모형을 결합하면 두 번 빼야한다.)

```
jlls = []  
jlls.append(gnb._joint_log_likelihood(x_num)) # shape = (n, C)  
jlls.append(cnb._joint_log_likelihood(x_cat.astype(int)))  
  
jlls = np.r_[0, 3, jlls] # shape = (# models, n, C)  
jll_sum = np.sum(jlls, axis=0) - cnb.class_log_prior_
```

- 여기서 `_joint_log_likelihood()` 는  $\ln \Pr(c) + \ln \Pr(x|c)$  이므로 prior를 한번 빼야한다.
- `cnb.class_log_prior_` 은 prior의 log 값 (GaussianNB에는 없는 attribute이다.)
- $\ln \Pr(Y|x_{\text{num}}, x_{\text{cat}}) \propto \ln \Pr(x_{\text{num}}, Y) + \ln \Pr(x_{\text{cat}}, Y) - \ln \Pr(Y)$ 의 우변에 해당

1. 확률을 정규화한다.

```
predict_log_prob = jll_sum - np.logaddexp(jll_sum[:,0], jll_sum[:,1])
```

[428 46]  
[132 177]]

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.76      | 0.90   | 0.83     | 474     |
| 1            | 0.79      | 0.57   | 0.67     | 309     |
| accuracy     |           |        | 0.77     | 783     |
| macro avg    | 0.78      | 0.74   | 0.75     | 783     |
| weighted avg | 0.78      | 0.77   | 0.76     | 783     |

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic Regression
  - 9.4 update and maintenance

[[428 46]

[132 177]]

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.76      | 0.90   | 0.83     | 474     |
| 1            | 0.79      | 0.57   | 0.67     | 309     |
| accuracy     |           |        | 0.77     | 783     |
| macro avg    | 0.78      | 0.74   | 0.75     | 783     |
| weighted avg | 0.78      | 0.77   | 0.76     | 783     |

## 9 Generative vs Discriminant classifiers

### 9.1 추정방법

- Generative classifiers
  - $\Pr(X|Y)$ 의 분포, 즉 함수형태를 가정한다.
  - 자료에서  $\Pr(X|Y)$ 과  $\Pr(Y)$ 를 계산한 후 Bayes 규칙을 이용하여  $\Pr(Y|X)$ 를 계산
  - 소표본에선 prior의 선택이 예측결과에 영향을 미칠 수 있다.
  - Overfitting을 줄이기 위한 regularization과 관련이 있다.
- Discriminant Classifiers
  - $\Pr(Y|X)$ 의 분포를 가정한다.
  - 자료에서  $\Pr(Y|X)$ 를 직접 계산한다.

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curve
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 9.2 Rate of convergence: Learning curve

- Discriminant 모형과 Generative 모형의 asymptotic properties
- $\mathcal{O}(\cdot)$ 은 learning curve 혹은 추정오차의 변화율로 해석할 수 있다.

Proposition 1.

$$\varepsilon_{\text{Dis},\infty} \leq \varepsilon_{\text{Gen},\infty} \quad \text{if not correctly specified}$$

Proposition 2.  $\varepsilon_{\text{Dis},n} \leq \varepsilon_{\text{Gen},n}$

- $\mathcal{O}(\varepsilon_{\text{Dis},n}) = \mathcal{O}(\sqrt{\log n})$
  - $\mathcal{O}(\varepsilon_{\text{Gen},n}) = \mathcal{O}(\sqrt{n})$
- < br > Theorem4.*
- Feature의 수가 증가할 때 표본의 수가 증가하면서 generative 모형의 오차가 더 빠르게 asymptotic error로 수렴한다.
    - 오차가 감소하는 것이 아니라 asymptotic error로 수렴하는 속도이므로 정확도를 의미하지 않는다.
    - 단순한 가정으로 인해 일반적으로 편차는 더 큰 것으로 알려져 있지만 마찬가지 이유로 variance가 상대적으로 작다.
  - Ng, A.Y. and Jordan, M.I. (2002). (<https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf>) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, *Neural Information Processing Systems*, 14, 841.

## Contents ⚙

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
  - 2 예제: 봄무개와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시각화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 9.3 Gaussian Naïve Bayes vs. Logistic Regression

- 가정 1 (GNB1):  $X_k | Y$  conditionally independent
- 가정 2 (GNB2): GNB1 + ( $\Sigma_c = \Sigma$ )
- 비교를 위한 가정
  - 무한히 많은 학습 표본
  - Logistic regression의 선형 decision boundary
- 1과 2의 가정 성립 (benchmark)
  - Logistic regression ~ GNB1 ~ GNB2
- 1과 2 모두 성립하지 않음 ( $X$ 에 대한 가정)
  - Logistic regression  $\succcurlyeq$  GNB2, GNB1  $\succcurlyeq$  GNB2
- 1만 성립 (비선형 decision boundary)
  - GNB1  $\succcurlyeq$  Logistic regression  $\succcurlyeq$  GNB2
- Gaussian Bayes가 data generating process에 대해 가정하는 것은 정규분포 뿐이다.
- 이에 반해 logistic regression은 log-odd가 선형이라는 제약을 사용한다.
- 완화될 수 있는 제약이지만 대안의 선택이 쉽지 않다. DNN은 여러 개의 logistic regression을 병렬로 동시에 사용

## Contents ↗

- ▼ 1 Gaussian Bayes
  - 1.1 Gaussian Naïve Bayes
  - 1.2 연속확률변수의 Bayes' theorem
- 2 예제: 몸무게와 키를 이용한 성별예측
- ▼ 3 Gaussian Bayes classifier
  - 3.1 Gaussian Bayes, Gaussian Non-Naïve Bayes
  - 3.2 Gaussian Naïve Bayes
    - 3.2.1 Posterior 계산에 필요한 정보
    - 3.2.2 Class마다 공분산이 다를 경우
    - 3.2.3 모든 class의 공분산이 같을 경우
  - 3.3 Gaussian Naïve Bayes의 추정
  - 3.4 Gaussian Naïve Bayes Classifier
- ▼ 4 Binary classifier와 decision boundary
  - 4.1 Feature 가 하나이면서 조건부분산
  - 4.2 Feature 가 하나이면서 조건부분산
  - 4.3 Feature 가 여러 개이면서 조건부분산
  - 4.4 Feature 가 여러 개이면서 조건부분산
  - 4.5 Logistic Regression과의 관계
- ▼ 5 Gaussian Bayes - 시작화
  - 5.1 조건부공분산이 동일하며 조건부독립
  - 5.2 조건부공분산이 다르며 조건부독립
  - 5.3 Multiclass Gaussian Naïve Bayes
- ▼ 6 연속확률변수의 독립성 검증
  - 6.1 Chi-Square test 복습
  - 6.2 Chi-square test를 연속확률변수에 적용
  - 6.3 mutual information
- ▼ 7 Palmer Penguins
  - 7.1 Palmer Penguins dataset
  - 7.2 Sample distribution
  - 7.3 Data Cleaning
  - 7.4 조건부 분포와 조건부 독립성
- ▼ 8 Mixed Naïve Bayes
  - 8.1 상이한 유형의 features
  - 8.2 독립성과 추정결과의 결합
  - 8.3 결합확률분포를 이용한 방법
  - 8.4 Implementation of mixed naïve Bayes
- ▼ 9 Generative vs Discriminant classifiers
  - 9.1 추정방법
  - 9.2 Rate of convergence: Learning curves
  - 9.3 Gaussian Naïve Bayes vs. Logistic regression
  - 9.4 update and maintenance

## 9.4 update and maintenance

- Logistic regression
  - 표본의 크기와 feature의 수가 많아지면 logistic regression의 추정은 어려워진다.
  - 새로운 표본이 추가될 때마다 전체 표본을 이용하여 추정해야 한다.
  - 과거의 추정결과는 새로운 추정에 별다른 도움이 되지 않는다. 다음과 같이 초기값으로 사용하는 정도이다.

```
clf = LogisticRegression(warm_start=True)
clf.fit(X,y)
clf.fit(X_new, y_new)
```

- Naïve Bayes
  - 자료의 수나 feature의 수가 많아 기존 memory 용량으로 추정이 어려울 때 표본을 여러 개의 batch로 나누어 추정이 가능하다.
  - 동일한 방법으로 새로운 자료가 추가될 때마다 간단하게 기존 estimator를 update할 수 있다.

```
from joblib import dump, load

clf = load('filename.joblib')
clf.partial_fit(X, y, classes=None, sample_weight=None)
dump(clf, 'filename.joblib', compress=3)
```

- 크지 않은 model이라면 pickle 을 사용할 수도 있다.

```
import pickle

clf = pickle.loads(model)
clf.partial_fit(X, y)
model = pickle.dumps(clf)
```

- joblib이나 pickle 은 모든 python object를 저장할 수 있다.