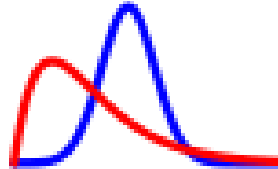
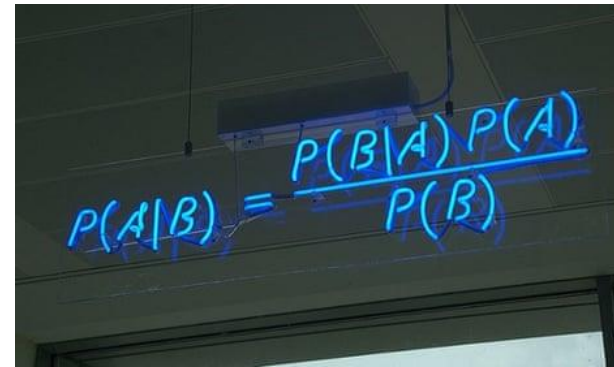


Gaussian Naive Bayes

The past is the only key to the future.

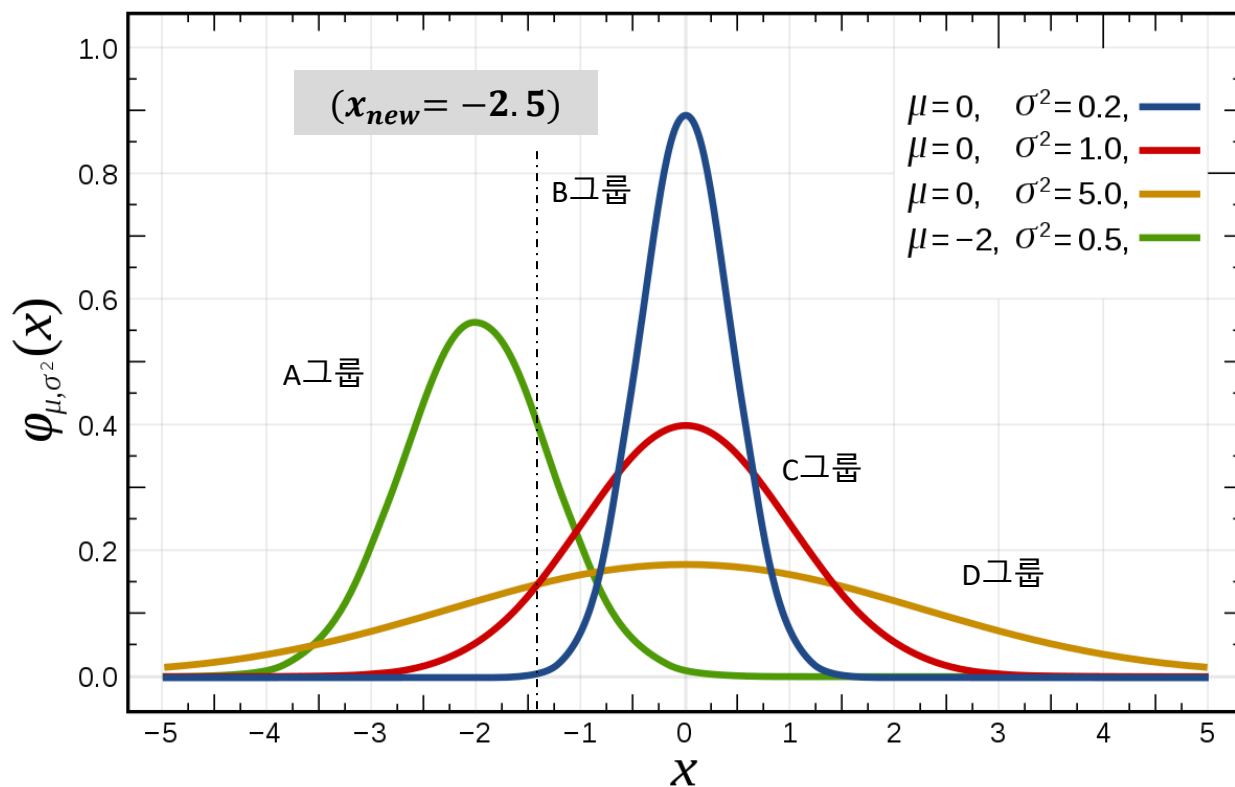


A photograph of a whiteboard with the formula for Bayes' theorem written in blue marker. The formula is
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The whiteboard shows the formula for conditional probability, Bayes' theorem, written in blue marker. The formula is
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

과거는 미래를 여는 열쇠다.

관측된 자료를 바탕으로 이산형 히스토그램을 통해 연속형 확률분포를 다음과 같이 그렸다. 만약 새로운 데이터 '-2.5'인 경우에는 어느 그룹(A,B,C, D)에 속할 확률이 가장 높은가?



생각할 점(1/2)

- 매일 아침 출근시간에 지하철을 타는 남자가 금융 및 보험업에 종사할 확률은?



산업별(1)	산업별(2)	2018
▲ ▼	▲ ▼ ▢	▲ ▼ ▢
합계	사업체수 (개)	14,648
	종사자수 (명)	66,759
정보통신업	사업체수 (개)	50
	종사자수 (명)	573
금융 및 보험업	사업체수 (개)	185
	종사자수 (명)	3,369
부동산업	사업체수 (개)	592

사전확률(Prior probability)
= 3369 / 66759 = 0.050 (5.0%)

https://kosis.kr/statHtml/statHtml.do?orgId=622&tblId=DT_62201_D000003

생각할 점(2/2)

- 금융보험업에 종사하는 사람의 넥타이 착용률이 90%이고 기타 다른 업종 종사자의 넥타이 착용률은 15%라는 새로운 사실을 알게 되었을 경우에는 달라지나?



산업별(1)	산업별(2)	2018
합계	사업체 수 (개)	14,648
		5,733
	종사자 수 (명)	66,759
정보통신업	사업체 수 (개)	50
	종사자 수 (명)	573
금융 및 보험업	사업체 수 (개)	185
	종사자 수 (명)	3,369
부동산업	사업체 수 (개)	592

■ 새로운 정보: 업종별 넥타이 착용률



- 넥타이 착용률
 - 금융/보험업: 90%
 - 기타업종 평균: 15%
- 예상 넥타이 착용자수
 - 금융/보험업: 3032 (=3369 x 90%)
 - 기타업종 평균: 10014 (=66759 x 15%)

사후확률 (Posterior probability)

$$= 3032 / (3032 + 10014) = 0.232 (23.2\%)$$

Frequentist vs. Bayesian



Are you Bayesian or Frequentist?

137K views • 1 year ago



Cassie Kozyrkov

What if I told **you** I can show **you** the difference bet
SUMMARY ...

CC

<https://www.youtube.com/watch?v=GEFxFVESQXc&t=60s>


Gaussian Naïve Bayes Sun Rising Problem

- 누군가 당신에게 ‘내일 해가 뜰 확률을 묻는다면?’

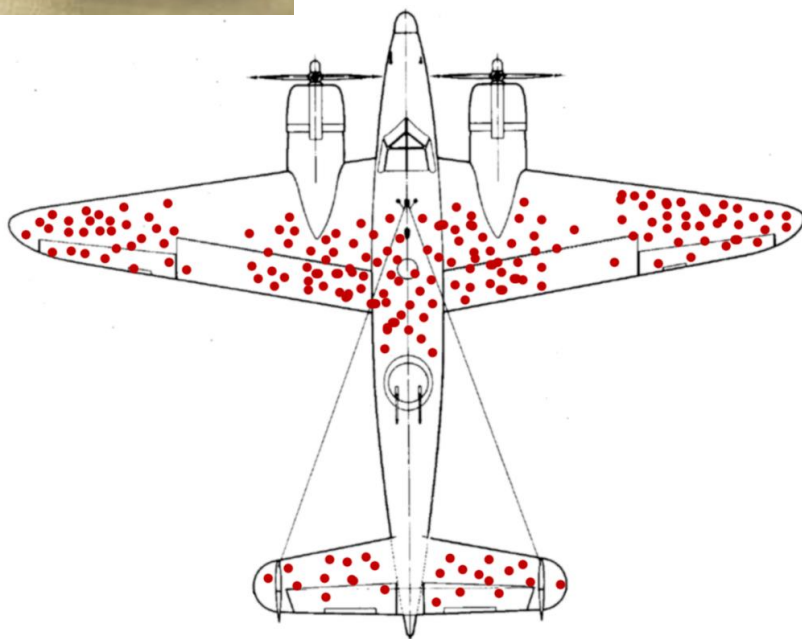
The **sunrise problem** can be expressed as follows: "What is the probability that the sun will rise tomorrow?" The sunrise problem illustrates the difficulty of using [probability theory](https://en.wikipedia.org/wiki/Probability_theory) when evaluating the plausibility of statements or beliefs.

https://en.wikipedia.org/wiki/Sunrise_problem



Usually inferred from repeated observations: *"The sun always rises in the east"*. 

Missing from data-survivorship bias



비행기	손상부위	결과
1) 헬캣아고네스	동체	귀환
2) 브롱크스파머	?	격추
3) 피스톨패킹파	엔진	귀환
.....	
375) 홈시크엔젤	?	격추
376) 컬래미티제인	없음	귀환



손상부위	귀환(총 316기)	격추 (총 60기)
엔진	29	?
조종석	36	?
동체	50	?
앞날개	55	?
없음	146	0

$$P(\text{동체손상/귀환}) = 50/316 = 15.8\%$$

$$P(\text{귀환/동체손상}) = 50/(50+?) = \text{??\%}$$

https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military

Missing from data-survivorship bias

- 원래 데이터를 가공, 조합, 정제 등의 처리 작업뿐만 아니라 존재하지 않는 자료를 만들 경우 예측 성능을 혁신적으로 높일 수 있음 (derivative features)

손상부위	귀환(총 316기)	격추 (총 60기)
엔진	29	31
조종석	36	21
동체	50	4
앞날개	55	4
없음	146	0

B-17이 적과 조우하는 전형적인 양상을 공군조종사와 엔지니어가 재현하여 가상의 데이터 생성


$$P(\text{귀환}/\text{엔진}) = 29/(29+31) = 48\%$$

$$P(\text{귀환}/\text{동체손상}) = 50/(50+4) = 93\%$$

$$P(\text{귀환}/\text{조종석}) = 36/(36+21) = 63\%$$

Gaussian Naïve Bayes

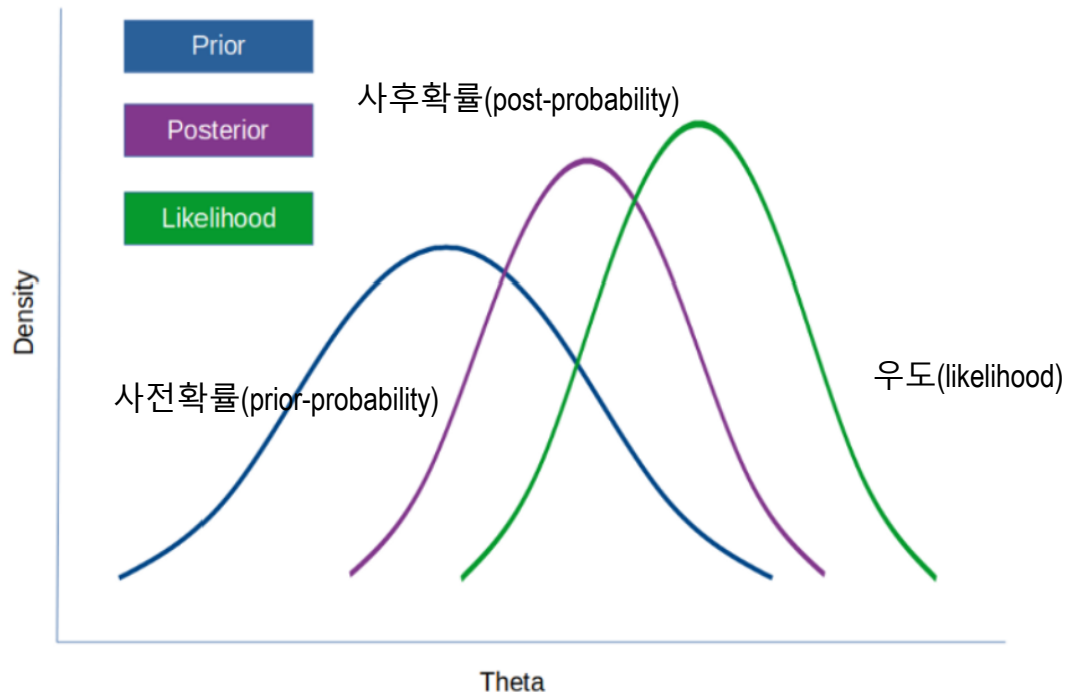
사후확률은 신념(Credibility)

- 미지의 세계에 대한 구체적인 사실 확인, 관측치 발견, 경험을 통해 나의 신념은 변한다.


Posterior Distribution (Credibility)

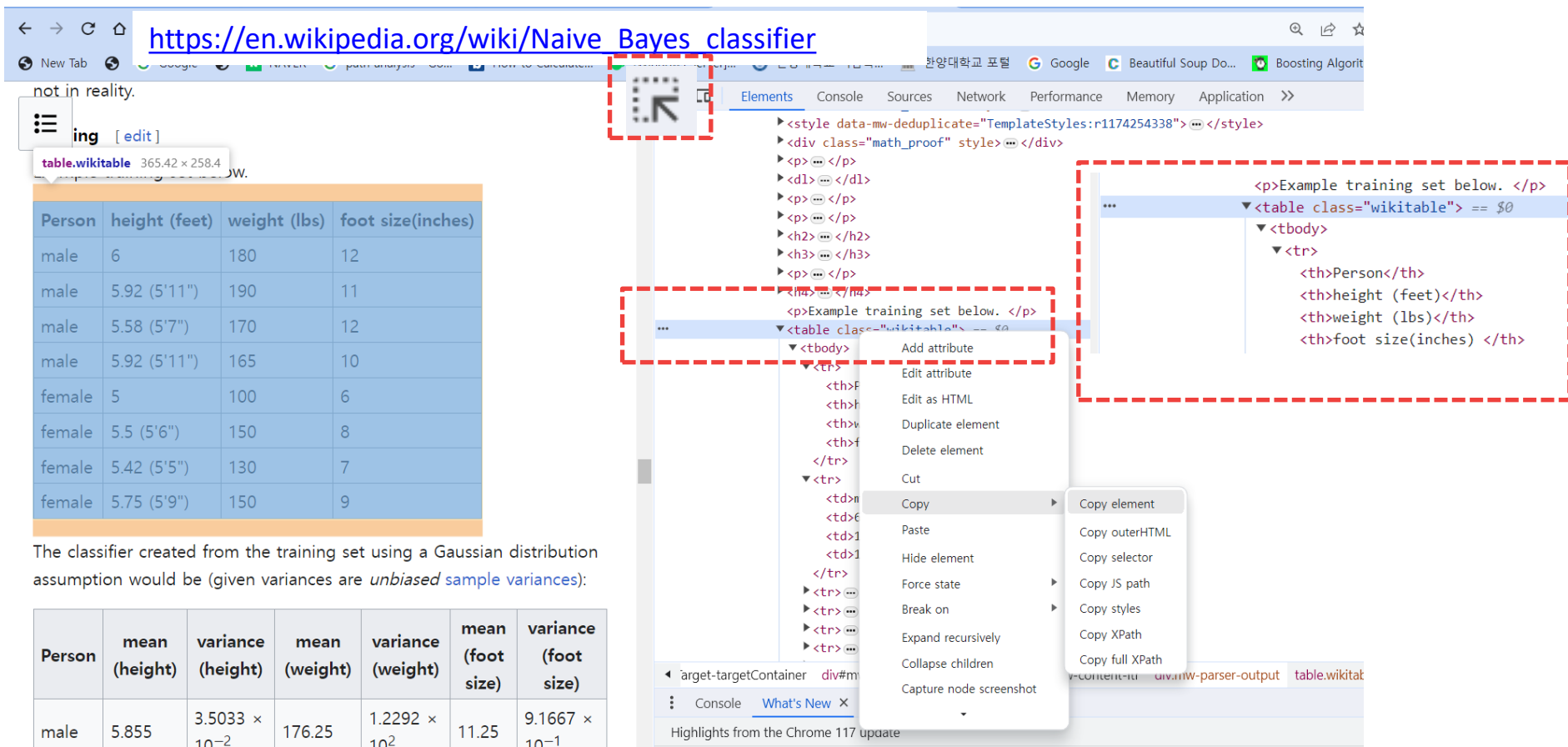
$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')}$$

- This distribution is comprised of the prior distribution (previous data) and likelihood function (probabilities inferred through Bayesian statistics).
- COVID-19 has demonstrated the need to account for **uncertainty** when making forecasts.



Gaussian Naïve Bayes 베이지안 예제

아래 사이트로 이동하여 'F12'키를 누르고  를 클릭한 후 왼쪽 사이트 화면의 테이블을 가리키면 '<table class="wikitable">'이 나타나는 것을 확인하고 'Copy element'를 하고



The classifier created from the training set using a Gaussian distribution assumption would be (given variances are *unbiased sample variances*):

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^2	11.25	9.1667×10^{-1}

Gaussian Naïve Bayes 베이지안 예제

chat GPT에게 해당되는 테이블을 웹스크롤링하여 dataframe으로 받는 코드를 요청하면



```
<table class="wikitable">
```

```
<tbody><tr>
```

```
<th>Person</th>
```

```
<th>height (feet)</th>
```

```
<th>weight (lbs)</th>
```

```
<th>foot size(inches)
```

```
</th></tr>
```

```
<tr>
```

```
<td>male</td>
```

```
<td>6</td>
```

```
<td>180</td>
```

```
<td>12
```

```
</td></tr>
```

```
<tr>
```

```
<td>male</td>
```

```
<td>5.92 (5'11")</td>
```

python

Copy code

```
import requests
from bs4 import BeautifulSoup
import pandas as pd

# 웹 페이지 URL
url = '여기에_웹_페이지_URL_입력'

# 웹 페이지 내용을 가져옵니다.
response = requests.get(url)
html = response.text

# BeautifulSoup을 사용하여 HTML 파싱
soup = BeautifulSoup(html, 'html.parser')

# 원하는 테이블을 선택합니다.
# 예를 들어, 첫 번째 테이블을 선택하려면 다음과 같이 합니다.
table = soup.find('table', class_='wikitable')

# 테이블을 데이터프레임으로 변환
df = pd.read_html(str(table))[0]

# 데이터프레임 출력
print(df)
```

Gaussian Naïve Bayes

베이지안 예제

chat GPT에게 해당되는 테이블을 웹스크롤링하여 dataframe으로 받는 코드를 요청한 후

```
import requests
from bs4 import BeautifulSoup
# 웹 페이지 URL
url = 'https://en.wikipedia.org/wiki/Naive_Bayes_classifier'
# 웹 페이지 내용을 가져옵니다.
response = requests.get(url)
html = response.text
# BeautifulSoup을 사용하여 HTML 파싱
soup = BeautifulSoup(html, 'html.parser')
# 원하는 테이블을 선택합니다.
# 예를 들어, 첫 번째 테이블을 선택
table = soup.find('table', class_='wikitable')

import pandas as pd
# 테이블을 데이터프레임으로 변환
df = pd.read_html(str(table))[0]
# 데이터프레임 출력
df
```

	Person	height (feet)	weight (lbs)	foot size(inches)
0	male	6	180	12
1	male	5.92 (5'11")	190	11
2	male	5.58 (5'7")	170	12
3	male	5.92 (5'11")	165	10
4	female	5	100	6
5	female	5.5 (5'6")	150	8
6	female	5.42 (5'5")	130	7
7	female	5.75 (5'9")	150	9

Gaussian Naïve Bayes 베이지안 예제

chat GPT에게 해당되는 테이블을 웹스크롤링하여 dataframe으로 받는 코드를 요청하면

```
# 데이터 전처리
```

```
df['height (feet)'] = df['height (feet)'].apply(lambda x : float(x.split('.')[0]))
```

```
# 시각화
```

```
import matplotlib.pyplot as plt
```

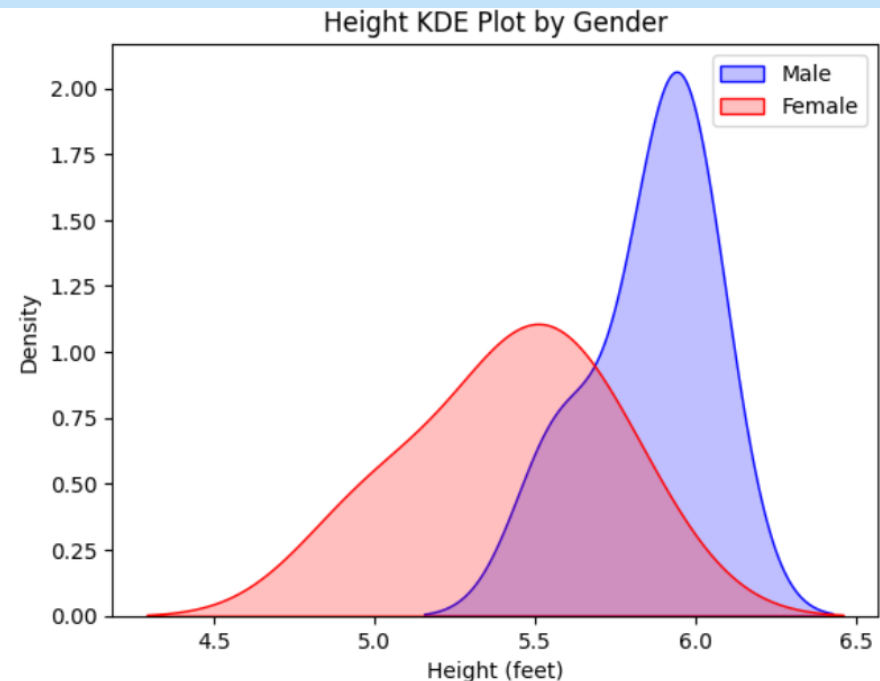
```
sns.kdeplot(data=df[df['Person'] == 'male']['height (feet)'], label='Male', color='blue', shade=True)
```

```
sns.kdeplot(data=df[df['Person'] == 'female']['height (feet)'], label='Female', color='red', shade=True)
```

```
plt.title('Height KDE Plot by Gender')
```


```
plt.xlabel('Height (feet)'); plt.ylabel('Density')
```

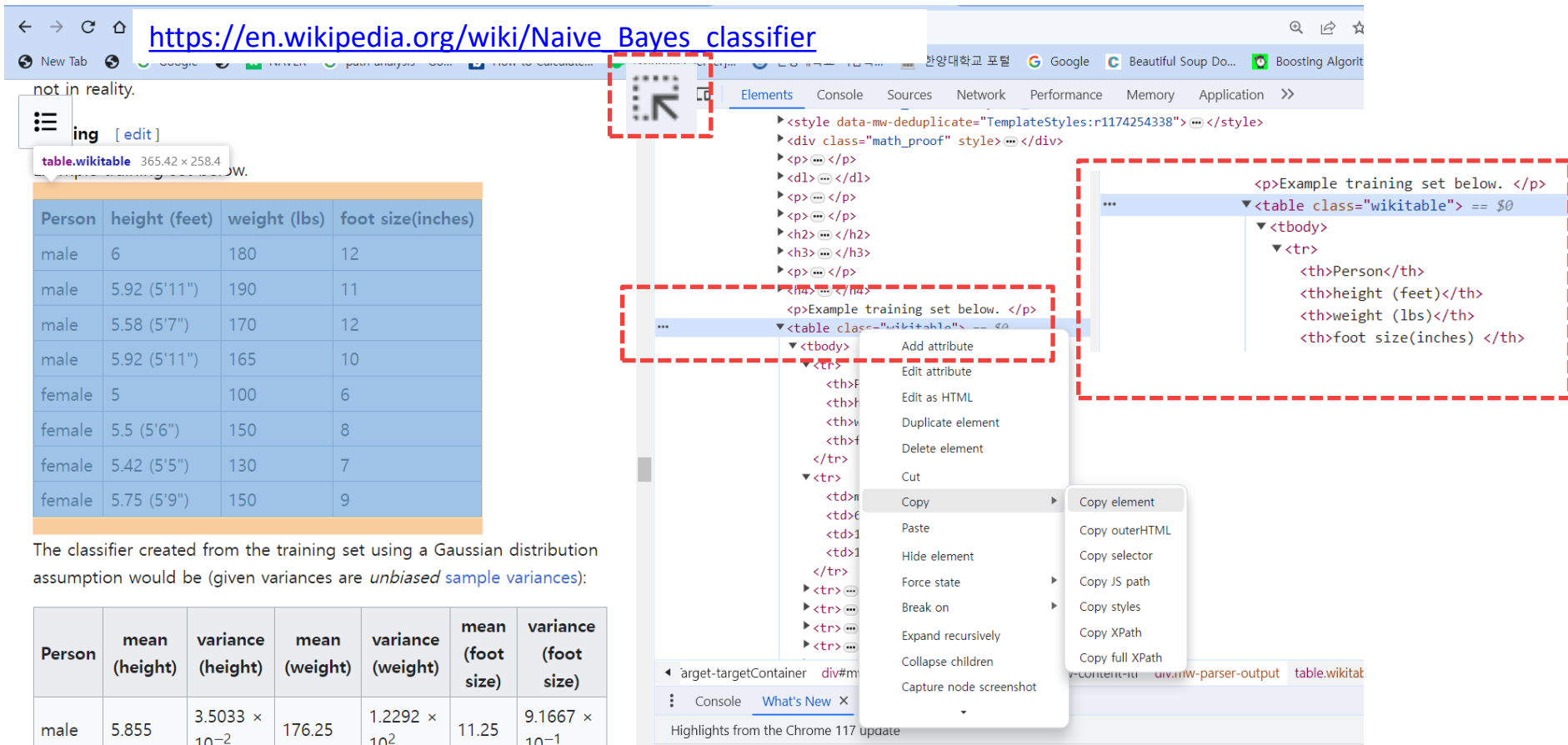
```
plt.legend(); plt.show()
```



Gaussian Naïve Bayes

베이지안 예제

아래 사이트로 이동하여 'F12'키를 누르고  를 클릭한 후 왼쪽 사이트 화면의 테이블을 가리키면 '<table class="wikitable">'이 나타나는 것을 확인한다.



not in reality.

ing [edit]

table.wikitable 365.42 x 258.4

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

The classifier created from the training set using a Gaussian distribution assumption would be (given variances are *unbiased sample variances*):

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^2	11.25	9.1667×10^{-1}

target-targetContainer div#m

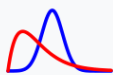
Console What's New

Highlights from the Chrome 117 update

Gaussian Naïve Bayes

구글 예제

Part of a series on
Bayesian statistics



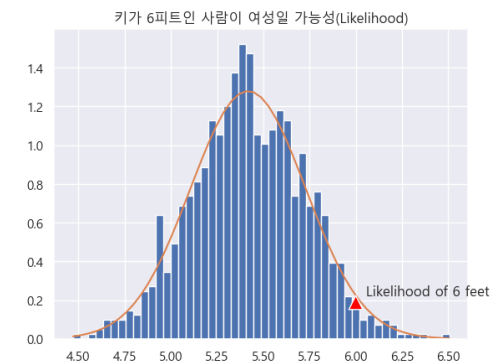
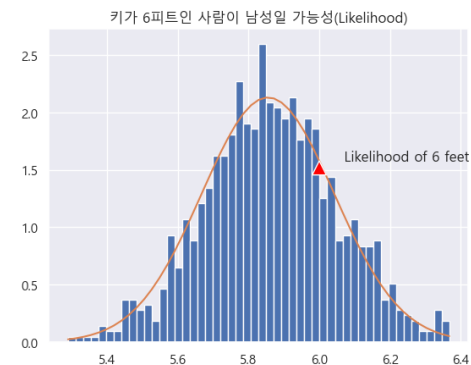
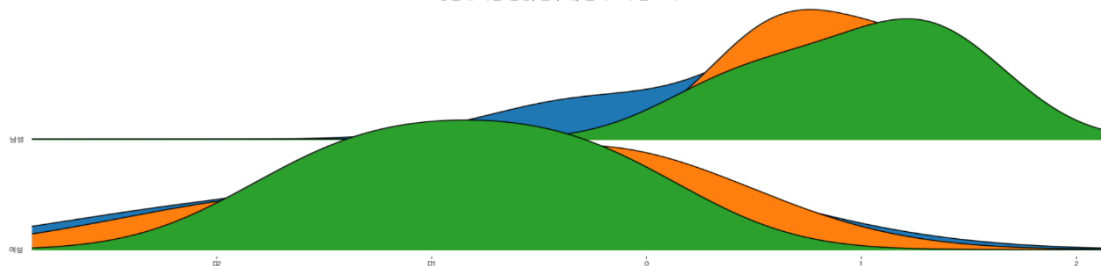
Theory

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

$$\begin{aligned}
 p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2, \dots, x_n, C_k) \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2 | x_3, \dots, x_n, C_k) p(x_3, \dots, x_n, C_k) \\
 &= \dots \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2 | x_3, \dots, x_n, C_k) \dots p(x_{n-1} | x_n, C_k) p(x_n | C_k) p(C_k)
 \end{aligned}$$

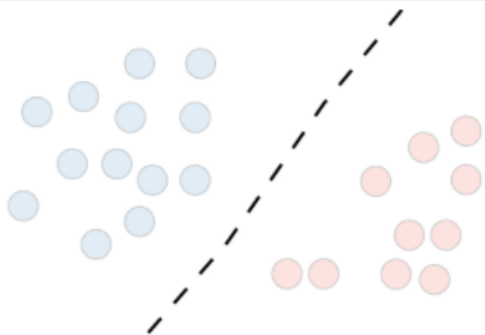
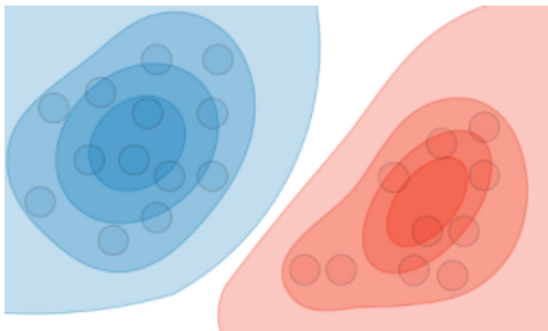
	성별	신장	무게	발의크기	(신장, mean)	(신장, var)	(무게, mean)	(무게, var)	(발의크기, mean)	(발의크기, var)
0	남성	6.00	180.0	12.0	5.8550	0.035033	176.25	122.916667	11.25	0.916667
1	남성	5.92	190.0	11.0	5.8550	0.035033	176.25	122.916667	11.25	0.916667
2	남성	5.58	170.0	12.0	5.8550	0.035033	176.25	122.916667	11.25	0.916667
3	남성	5.92	165.0	10.0	5.8550	0.035033	176.25	122.916667	11.25	0.916667
4	여성	5.00	100.0	6.0	5.4175	0.097225	132.50	558.333333	7.50	1.666667
5	여성	5.50	150.0	8.0	5.4175	0.097225	132.50	558.333333	7.50	1.666667
6	여성	5.42	130.0	7.0	5.4175	0.097225	132.50	558.333333	7.50	1.666667
7	여성	5.75	150.0	9.0	5.4175	0.097225	132.50	558.333333	7.50	1.666667
8	NaN	6.00	130.0	8.0	NaN	NaN	NaN	NaN	NaN	NaN

성별에 따른 신장, 몸무게, 발의크기 분포 비교도



Gaussian Naïve Bayes

Discriminant or Generative ?

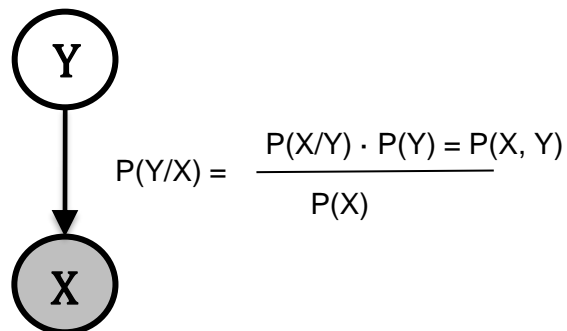
	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

- 데이터로부터 직접 조건부 확률을 계산
- 확률모형에는 관심이 없고 x 와 y 의 패턴을 파악하여 직접 분류를 하기에 y 가 반드시 필요
- 선형회귀분석, SVM, 의사결정나무와 같이 확률적 모델을 가정하지 않고 간단하게 직선, 커브 등으로 사후확률을 직접 예측
- 두 개의 확률 모형 사전 확률과 우도를 정의하여 조건부확률인 사후 확률 생성
- 가우시안 믹스처 모델, 토픽 모델과 같은 비지도학습에도 적용 가능
- 특성 변수간 독립이라는 확률적 모형을 가정하기 때문에 예측 성능이 차별모형보다 낮지만, 데이터의 크기가 충분히 크면 성능은 비슷
- 가우시안 믹스처, 나이브 베이지안, GAN, 딥러닝

Gaussian Naïve Bayes Discriminant or Generative ?

- 단어 시퀀스에 조건부 확률을 할당하여 가장 자연스러운 단어 시퀀스를 찾는 RNN, CBOW
- 기계번역, 오타교정, 음성인식, 셰익스피어 문체 글쓰기, 바하 스타일의 작곡

Generative Model



기계번역 :

$P(\text{탔다/버스를}) > P(\text{태웠다/버스를})$ 이 되도록 조건부 확률이 할당되어 학습하면, 'I took a bus' 는 '나는 버스를 태웠다'가 아니라 나는 버스를 탔다'로 번역된다.

Discriminative Model

