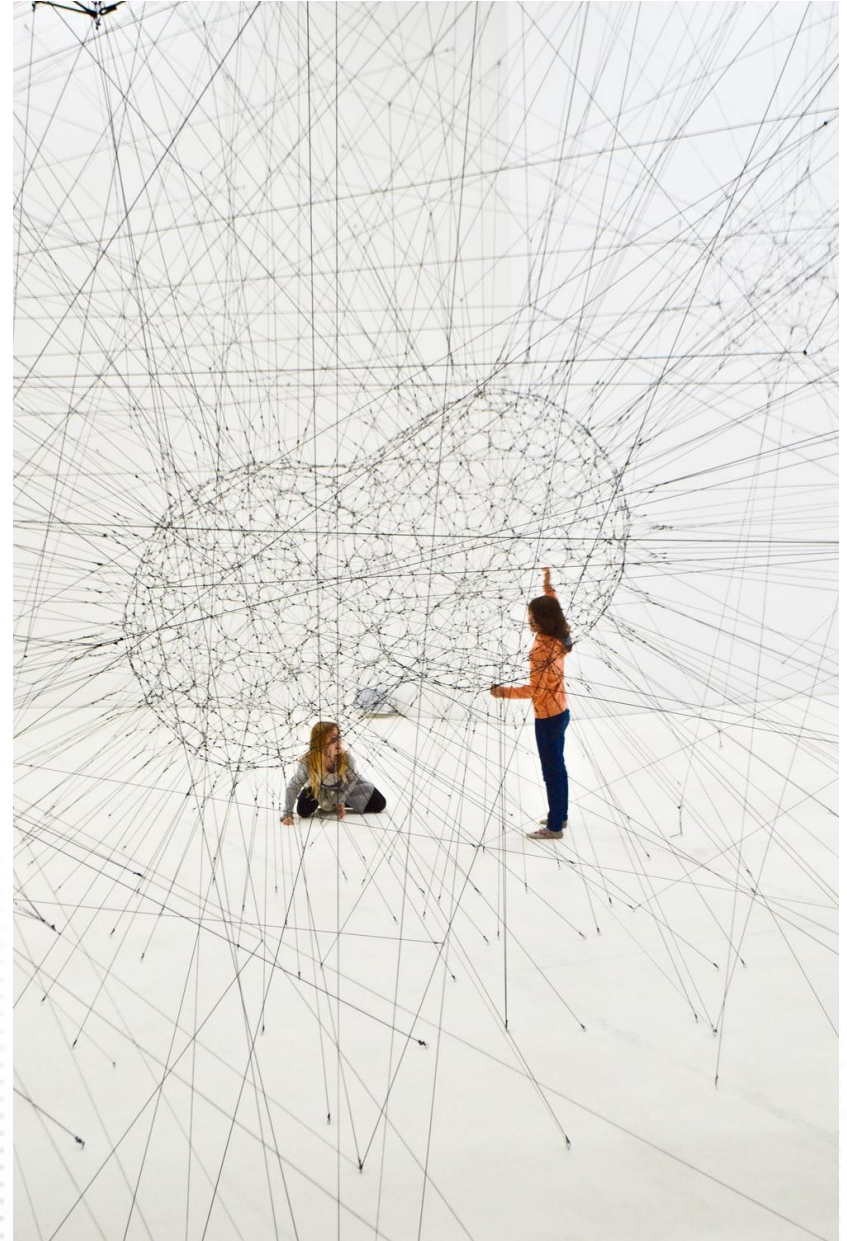


# 머신러닝 배우기

2025.05.15

조 상 구

ancestor9@kbu.ac.kr



# 3. Supervised Learning

- Bayesian algorithm
- Support Vector Machine

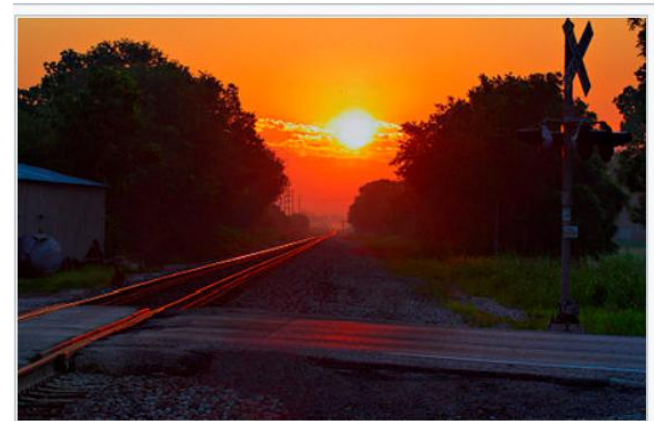



## Gaussian Naïve Bayes Sun Rising Problem

- 누군가 당신에게 ‘내일 해가 뜰 확률’을 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](https://en.wikipedia.org/wiki/Sunrise_problem)



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

## 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>

## Frequentist vs. Bayesian

- 만약 어떤 사람이 5회 연속으로 카드의 색깔을 맞추었다면 그 사람이 초능력자라고 얼마나 믿는가?  
3명의 친구 (Stub, Freq, Bays) 들의 관점을 비교



### Mr. Stub

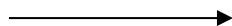
Prior  
Knowledge  
is everything



$$\left\{ \begin{array}{l} P(\theta = \text{super man}) = 0.001 \\ P(\theta = \text{Not superman}) = 0.999 \end{array} \right.$$

### Mr. Freq

The truth is  
somewhere  
in between



$$\left\{ \begin{array}{l} P(5 \text{ rights} / \theta = \text{super man}) = 1 \\ P(5 \text{ rights} / \theta = \text{Not superman}) = \left(\frac{1}{2}\right)^5 = 0.03125 \end{array} \right.$$

- 초능력자라면 무조건 맞추었으니 100%
- 초능력자가 아닌 일반 사람이라면 3.1%의 확률로 5회 연속으로 맞춘 것임
  - ▶ 초능력자가 아닌지는 확률로 나타내지 못함(Maximum likelihood Estimation은 제공하지 않음)

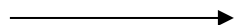


## Frequentist vs. Bayesian

- Stub는 너무 고집스럽고 Freq는 data를 너무 신봉하는 경향이 있지만 Bays는

### Mr. Bays

Observed  
data is  
everything



$$\left\{ \begin{aligned} P(\theta/5 \text{ rights}) &= \frac{P(5 \text{ rights}/\theta) * P(\theta)}{P(5 \text{ rights}/\theta) * P(\theta) + P(5 \text{ rights}/\sim\theta) * P(\sim\theta)} \\ &= \frac{1 * 0.001}{1 * 0.001 + 0.999 * (\frac{1}{2})^5} = 0.031038 = \mathbf{3.1\%} \\ P(\sim\theta/5 \text{ rights}) &= 1 - P(\theta/5 \text{ rights}) = \mathbf{96.9\%} \end{aligned} \right.$$

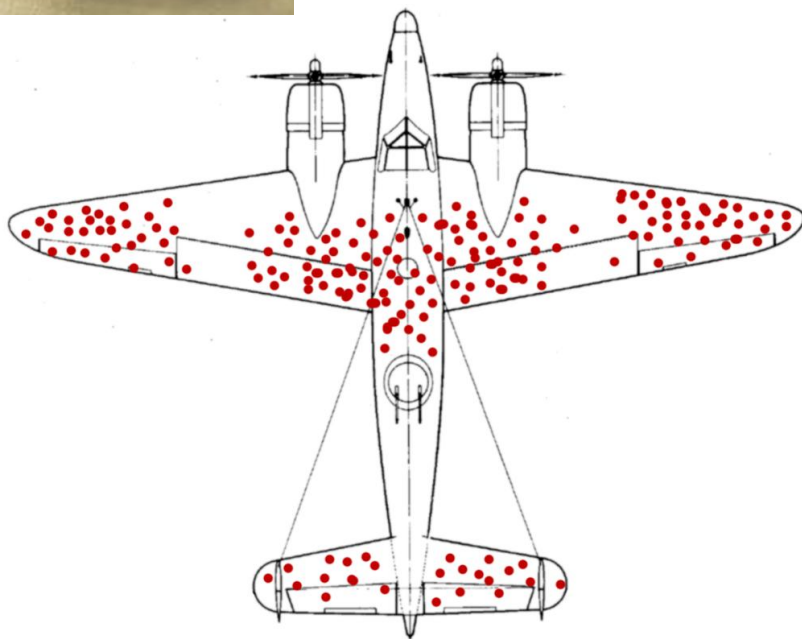
$$p(\theta | \text{data}) = \frac{p(\text{data} | \theta) \cdot p(\theta)}{p(\text{data})}$$

Posterior      Likelihood      Prior  
Normalization

$$p(\text{data}) = \sum_{\theta} p(\text{data} | \theta) \cdot p(\theta)$$

- 5회 연속으로 맞춘 걸 보니 초능력자일 가능성(믿음)이 96.9%, 일반인일 가능성(믿음)은 3.1%
- Bays의 관점은 Freq의 likelihood와 Stub의 믿음(Prior)을 곱한 값으로 확률이라가보다는 Credibility

# 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](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 가상  
의 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 사례

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

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



산업별(1)	산업별(2)	2018
합계	사업체 수 (개)	14,648
	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: 넥타이 착용률



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



[https://kosis.kr/statHtml/statHtml.do?orgId=622&tblId=DT\\_62201\\_D000003](https://kosis.kr/statHtml/statHtml.do?orgId=622&tblId=DT_62201_D000003)

# 사후확률 (Posterior probability)  
=  $3032 / (3032 + 10014) = 0.232$   
(23.2%)

## Gaussian Naïve Bayes

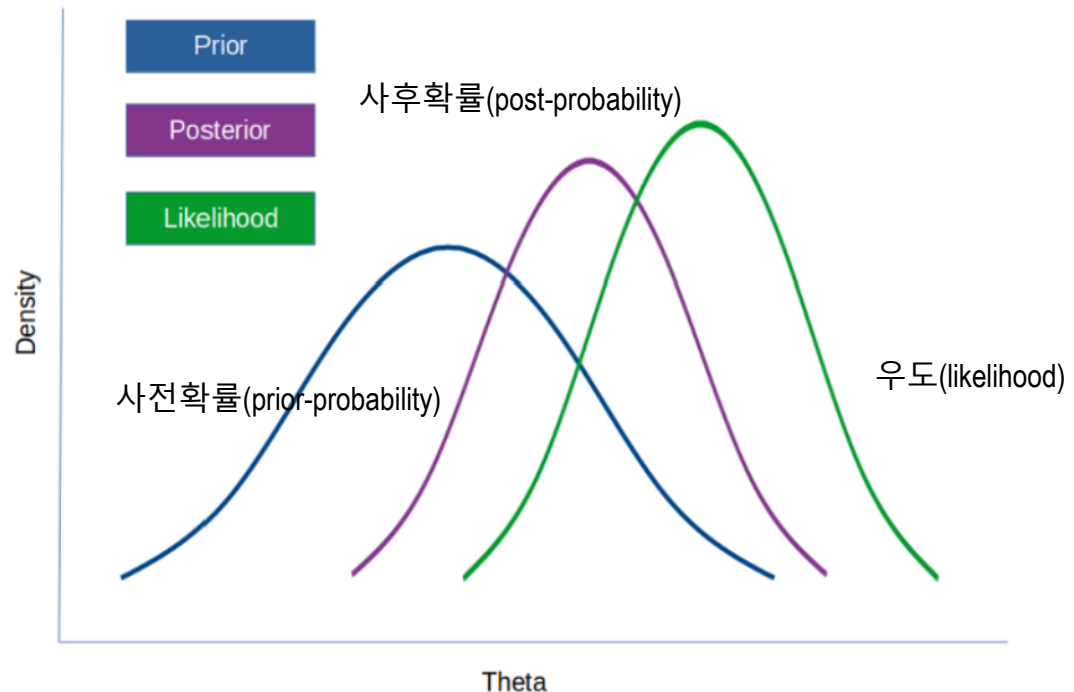
### 사후확률은 신념(Credibility)

- 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

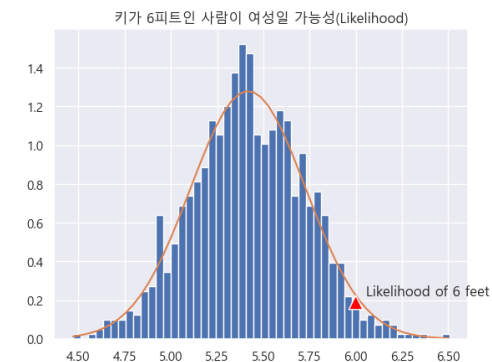
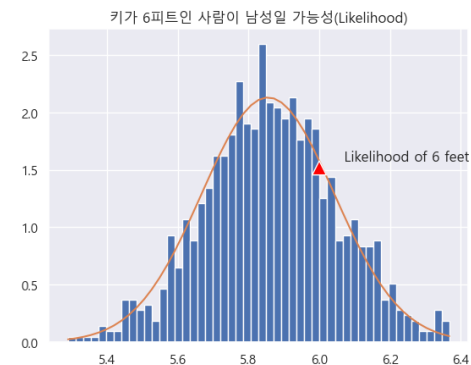
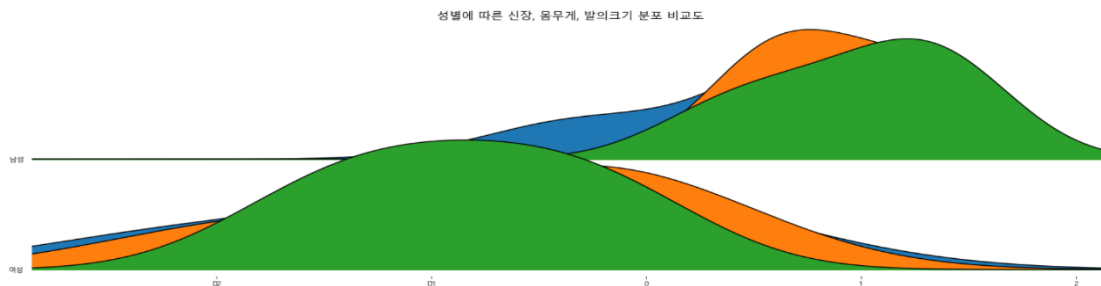
## 구글 예제



[https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)

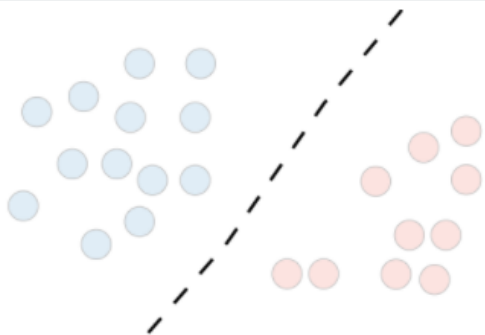
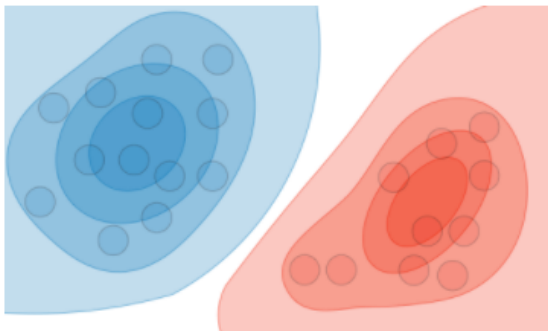
$$\begin{aligned}
 p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2, \dots, x_n, C_k) \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2 | x_3, \dots, x_n, C_k) p(x_3, \dots, x_n, C_k) \\
 &= \dots \\
 &= p(x_1 | x_2, \dots, x_n, C_k) p(x_2 | x_3, \dots, x_n, C_k)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

	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



# Gaussian Naïve Bayes

## Discriminant or Generative ?

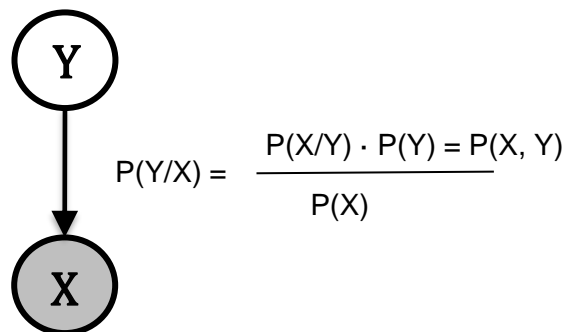
	Discriminative model	Generative model
<b>Goal</b>	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
<b>What's learned</b>	Decision boundary	Probability distributions of the data
<b>Illustration</b>		
<b>Examples</b>	Regressions, SVMs	GDA, Naive Bayes

- 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

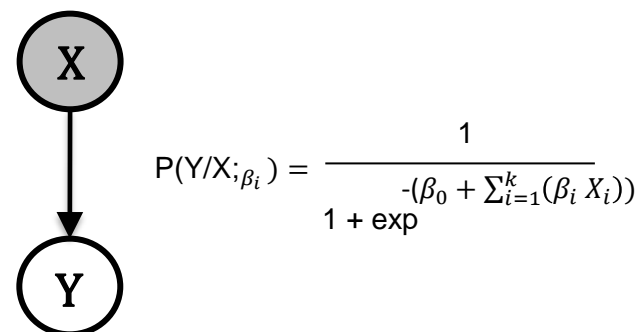
## Gaussian Naïve Bayes Discriminant or Generative ?

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

**Generative Model**



**Discriminative 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.

