

17기 정규세션

ToBig's 16기 김송민

Naïve Bayes Classifier

* Naïve : 순진하다 Bayes : 베이즈정리

Contents

Unit 01 | 확률 기초 (Probability Overview)

Unit 02 | 베이즈 정리

Unit 03 | Naïve Bayes Classification

목표

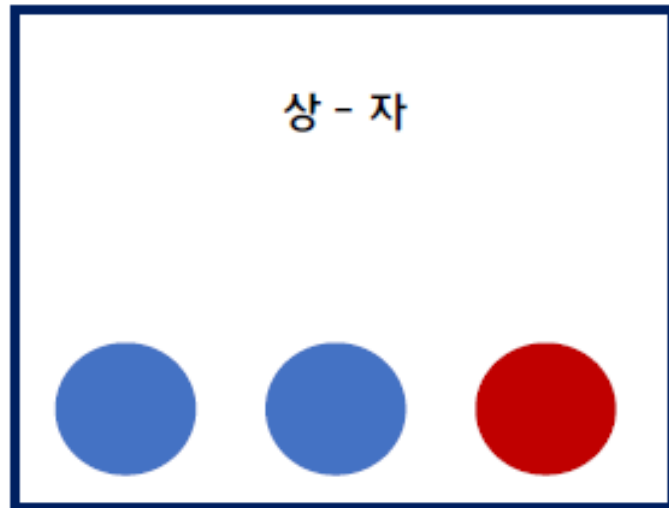
[목표]

- 기본적인 확률 공식들을 바탕으로 베이즈 정리를 이해한다.
- 앞서 이해한 베이즈 정리를 바탕으로 나이브 베이즈에 대해서 이해한다.

Unit 01 | 확률 기초(Probability Overview)

1-1) 확률이란 (Probability)

- 특정한 사건이 일어날 가능성을 나타낸 것



파란 공을 뽑을 확률 : $2/3$
빨간 공을 뽑을 확률 : $1/3$

$$1/3 + 2/3 = 1$$

Unit 01 | 확률 기초(Probability Overview)

1-2) 조건부 확률 (conditional probability)

- 어떤 사건이 일어난 조건 하에서, 다른 사건이 일어날 확률

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

사건 A가 일어났을 때, 사건 B가 일어날 확률

Ex. 2주차 세션에 출석했을 때 (A) 3주차 세션도 출석할 확률(B)
= $P(B|A)$
= 2주차, 3주차 모두 출석할 확률 (A&B) / 2주차 세션에 출석할 확률 (A)

- 곱셈 공식 교집합을 조건부 확률로 나타내기

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

Unit 01 | 확률 기초(Probability Overview)

1-3) 독립과 조건부 독립 (Independent & conditional independent)

- 독립 : 한 사건이 일어날 확률이 다른 사건이 일어날 확률에 영향을 미치지 않는 상태

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

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

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

- 조건부 독립 : 한 사건이 일어났다는 가정하에서, 서로 다른 두 사건은 독립인 상황

$$P(A, B|C) = P(A|C)P(B|C)$$

C 사건이 일어났을 때,
사건 A가 일어날 확률은 사건 B가 일어날 확률에 영향을 주지 않는다.

Ex. 1주차 과제를 모두 제출했을 때(C), 2주차 출석 확률(A)은 3주차 출석 확률(B)에 영향을 주지 않는다.

1주차 과제를 모두 제출했을 때(C), 2주차를 출석하면 3주차는 결석하고 놀러 나가게 되는 경향이 있다면? (A->B) 조건부 독립 X

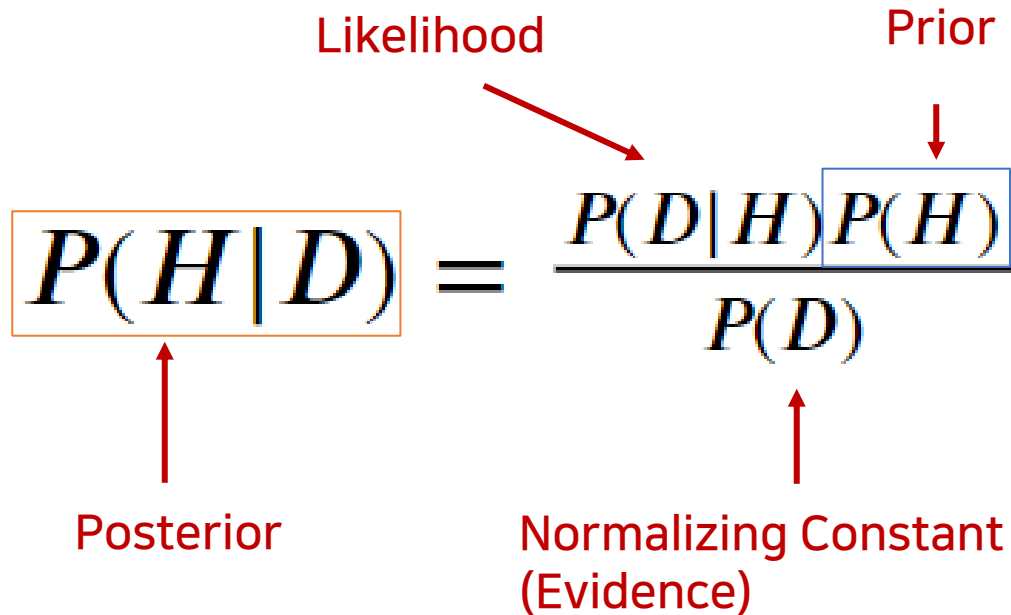
Unit 02 | 베이즈 정리 (Bayes' Rule)

2-1) 베이즈 정리

두 확률 변수의 **사전 확률(prior)**과 **사후 확률(posterior)** 사이의 관계를 나타내는 정리
사전 확률(prior)로부터 **사후 확률(posterior)**을 구하고자 한다

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-1) 베이즈 정리



The diagram shows the Bayes' Rule formula with labels and arrows indicating the components:

$$\boxed{P(H|D)} = \frac{P(D|H) \boxed{P(H)}}{P(D)}$$

- Posterior**: Points to $P(H|D)$ (enclosed in an orange box).
- Likelihood**: Points to $P(D|H)$ (enclosed in a blue box).
- Prior**: Points to $P(H)$ (enclosed in a blue box).
- Normalizing Constant (Evidence)**: Points to $P(D)$ (enclosed in a blue box).

H - 알고 싶은 정보 ex. 나가서 운동하기 적합한 날씨인가
D - 관찰하여 알고 있는 정보 ex. 기상정보

- Prior
 - 사전 확률
 - 과거의 경험을 토대로 나름대로 추정한 파라미터(H)의 확률
- Posterior
 - 사후 확률
 - 관측결과 사건 D가 일어난 조건 하의 파라미터(H)의 확률
- Likelihood
 - 사전 확률의 '과거 경험을 잘 설명하는 정도'
 - 모델 파라미터(H)를 바탕으로 하는 관측결과 사건(D)의 확률
- Normalizing Constant (Evidence)
 - 사건 D의 발생 가능성
 - 모델 파라미터(H)와 무관한 관측결과(D) 자체의 확률
 - 우리가 관심있는 H와 무관한 값. 보통 상수로 생각하고 무시하고 계산

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-1) 베이즈 정리

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Diagram labels for the equation:

- 2) Likelihood** (blue arrow pointing to $P(D|H)$)
- 1) Prior** (blue arrow pointing to $P(H)$)
- Posterior** (red arrow pointing to $P(H|D)$)
- 3) Normalizing Constant (Evidence)** (blue arrow pointing to $P(D)$)

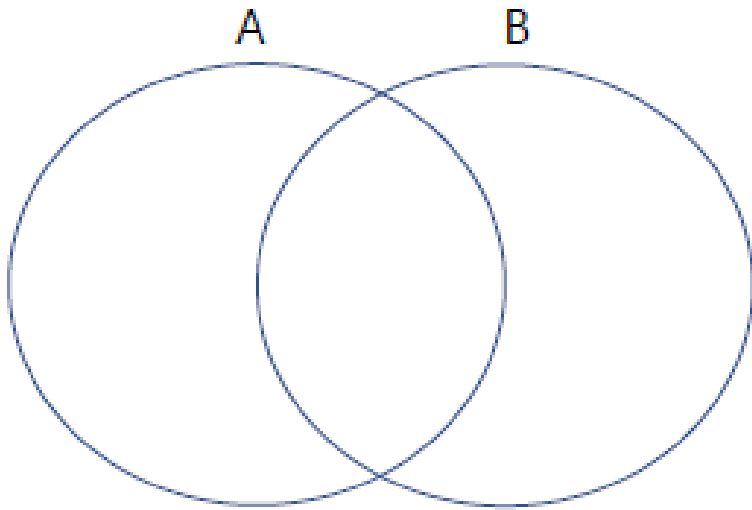
Posterior

- 직접적(데이터, 실험 등)으로 관찰 및 표현이 불가능하지만, 계산을 통해 '사후적으로' 알 수 있는 확률
- 사전에 예상한 가설 H 에 대한 확률(사전확률)에
 - 관찰된 데이터 D로 계산한 가능성 정도(likelihood)를 곱하고
 - 모델 파라미터와 무관한, 관측결과가 발생할 확률 (Normalizing Constant)로 나눴서 구한 값

H - 알고 싶은 정보 ex. 나가서 운동하기 적합한 날씨인가
D - 관찰하여 알고 있는 정보 ex. 기상정보

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-2) 베이즈 정리 증명



$$\begin{aligned}P(A|B) &= P(A \cap B) / P(B) \\P(A \cap B) &= P(A|B) * P(B) \\P(B \cap A) &= P(B|A) * P(A) \\P(B \cap A) &= P(A \cap B) \\P(A \cap B) &= P(B|A) * P(A) = P(A|B) * P(B) \\P(B|A) * P(A) &= P(A|B) * P(B)\end{aligned}$$



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

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-3) 연습 날씨(Sky) 정보로 운동하기 적합한지 여부 (Enjoy Point) 알아보기

Sky (X)	Enjoy Point (Y)
Sunny	Yes
Sunny	Yes
Rainy	No
Sunny	No
Rainy	Yes

$$P(Y = yes | X = sunny) = \frac{P(X = sunny | Y = yes) \times P(Y = yes)}{P(X = sunny)}$$

$$P(Y = no | X = sunny) = \frac{P(X = sunny | Y = no) \times P(Y = no)}{P(X = sunny)}$$

by 베이즈 정리 $P(B|A) = \frac{P(A|B) P(B)}{P(A)}$

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-3) 연습

$$P(Y = \text{yes} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{yes}) \times P(Y = \text{yes})}{P(X = \text{sunny})}$$

$$P(Y = \text{no} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{no}) \times P(Y = \text{no})}{P(X = \text{sunny})}$$

Sky (X)	Enjoy Point (Y)
Sunny	Yes
Sunny	Yes
Rainy	No
Sunny	No
Rainy	Yes

$$P(Y = \text{yes}) = 3/5$$

$$P(X = \text{sunny} | Y = \text{yes}) = 2/3$$

$$P(Y = \text{no}) = 2/5$$

$$P(X = \text{sunny} | Y = \text{no}) = 1/2$$

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-3) 연습

$$P(Y = \text{yes} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{yes}) \times P(Y = \text{yes})}{P(X = \text{sunny})}$$

$$P(Y = \text{no} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{no}) \times P(Y = \text{no})}{P(X = \text{sunny})}$$

Sky (X)	Enjoy Point (Y)
Sunny	Yes
Sunny	Yes
Rainy	No
Sunny	No
Rainy	Yes

$$P(Y = \text{yes} | X = \text{sunny}) = \frac{2/3 \times 3/5}{P(X = \text{sunny})}$$

$$P(Y = \text{no} | X = \text{sunny}) = \frac{1/2 \times 2/5}{P(X = \text{sunny})}$$

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-3) 연습

$$P(Y = yes | X = sunny) = \frac{P(X = sunny | Y = yes) \times P(Y = yes)}{P(X = sunny)}$$

$$P(Y = no | X = sunny) = \frac{P(X = sunny | Y = no) \times P(Y = no)}{P(X = sunny)}$$

Sky (X)	Enjoy Point (Y)
Sunny	Yes
Sunny	Yes
Rainy	No
Sunny	No
Rainy	Yes

$$P(Y = yes | X = sunny) = \frac{2/3 \times 3/5}{P(X = sunny)}$$

$$P(Y = no | X = sunny) = \frac{1/2 \times 2/5}{P(X = sunny)}$$

$$P(Y = yes | X = sunny) + P(Y = no | X = sunny) = 1$$

굳이 노란 박스를 구할 필요가 없다~!

Unit 02 | 베이즈 정리 (Bayes' Rule)

2-3) 연습

$$P(Y = \text{yes} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{yes}) \times P(Y = \text{yes})}{P(X = \text{sunny})}$$

$$P(Y = \text{no} | X = \text{sunny}) = \frac{P(X = \text{sunny} | Y = \text{no}) \times P(Y = \text{no})}{P(X = \text{sunny})}$$

Sky (X)	Enjoy Point (Y)
Sunny	Yes
Sunny	Yes
Rainy	No
Sunny	No
Rainy	Yes

$$P(Y = \text{yes} | X = \text{sunny}) = \frac{2/3 \times 3/5}{P(X = \text{sunny})}$$

$$P(Y = \text{no} | X = \text{sunny}) = \frac{1/2 \times 2/5}{P(X = \text{sunny})}$$

$$P(Y = \text{yes} | X = \text{sunny}) = 2/3$$

$$P(Y = \text{no} | X = \text{sunny}) = 1/3$$

Unit 03 | Naive Bayes Classification

3-1) 계산의 한계

$$f^*(x) = \operatorname{argmax}_{Y=y} P(X = x|Y = y)P(Y = y)$$

2. 예시 – 분류문제

d=관측치 개수 (관측된 날의 수)
K=class 개수 (Yes, no)

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$P(X = x|Y = y)$$

$$= P(x_1 = \text{sunny}, x_2 = \text{warm}, x_3 = \text{normal}, x_4 = \text{strong}, x_5 = \text{warm}, x_6 = \text{same} | y = \text{Yes})$$

$$P(Y=y) = (y=\text{Yes})$$

$$P(X = x|Y = y) = \text{for all } x, y \rightarrow (2^d - 1)k$$

$$P(Y=y) \text{ for all } y \rightarrow k-1$$

Unit 03 | Naive Bayes Classification

3-1) 계산의 한계

문제점: 계산량이 많아짐

$$P(X = x|Y = y) = \text{for all } x, y \rightarrow (2^d - 1)k$$

변수가 늘어날 수록 기하급수적으로 연산량이 증가함

어떻게 얻은 변수들인데..... d의 개수를 줄이는 건 피하고 싶어

->>해결책: 조건부 독립을 가정!!!!

Unit 03 | Naive Bayes Classification

3-2) Naive Bayes Classification

- 가정 : 종속변수(Y)가 주어졌을 때, **입력 변수들이 모두 독립이다!!!!**(조건부 독립 가정)
- **결과가 주어졌을 때**, 예측 변수 벡터의 정확한 **조건부 확률은 각 조건부 확률의 곱으로** 충분히 잘 추정 할 수 있다는 단순한 가정을 기초로 한다
-> 데이터셋을 순진하게 믿는다! -> Naïve Bayes!!

$$f^*(x) = \operatorname{argmax}_{Y=y} P(X = x|Y = y)P(Y = y)$$

$$\approx \operatorname{argmax}_{Y=y} P(Y = y) \prod_{1 \leq i \leq d} P(X = x_i|Y = y)$$

Unit 03 | Naive Bayes Classification

3-2) Naive Bayes Classification

$$\text{조건부 독립} \Rightarrow P(A \cap B | C) = P(A | C)P(B | C)$$

$$P(X = \mathbf{x} | Y = y)$$

$$= P(x_1 = \text{sunny}, x_2 = \text{warm}, x_3 = \text{normal}, x_4 = \text{strong}, x_5 = \text{warm}, x_6 = \text{same} | y = \text{Yes})$$

$$P(X=\mathbf{x}|Y=y)$$

$$= P(x_1=\text{sunny}|Y=\text{yes}) P(x_2=\text{warm}|Y=\text{yes}) P(x_3=\text{normal}|Y=\text{yes}) P(x_4=\text{strong}|Y=\text{yes}) P(x_5=\text{warm}|Y=\text{yes}) P(x_6=\text{same}|Y=\text{yes})$$

$$f^*(x) = \operatorname{argmax}_{Y=y} P(X = \mathbf{x} | Y = y) P(Y = y)$$

$$\approx \operatorname{argmax}_{Y=y} P(Y = y) \prod_{1 \leq i \leq d} P(X = x_i | Y = y)$$

Unit 03 | Naive Bayes Classification

3-2) Naive Bayes Classification

1. 알아야할 파라미터의 수가 대폭 줄어들게 된다.

$$P(X = x|Y = y) \Rightarrow dk$$

2. Feature들의 곱으로 바뀌면서 계산이 수월해진다.

Unit 03 | Naive Bayes Classification

3-3) Naive Bayes Classification 연습

Weather (날씨)	Temp (온도)	Play (경기 여부)
sunny	hot	No
sunny	hot	No
overcast	hot	Yes
rainy	mild	Yes
rainy	cool	Yes
rainy	cool	No
overcast	cool	Yes
sunny	mild	No
sunny	cool	Yes
rainy	mild	Yes
sunny	mild	Yes
overcast	mild	Yes
overcast	hot	Yes
rainy	mild	No

야외 스포츠 경기가 열리기 좋은 날씨인가? 아닌가?
 날씨가 overcast, 기온이 mild일 때 경기가 열릴 확률은?

1. 사전 확률

$$P(\text{Yes}) = 9 / 14 = 0.64$$

2. 사후 확률

$$P(\text{Overcast}|\text{Yes}) = 4 / 9 = 0.44$$

$$P(\text{Mild}|\text{Yes}) = 4 / 9 = 0.44$$

3. 베이즈 공식에 대입

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

$$\begin{aligned}
 &P(\text{Play}=\text{Yes} | \text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild}) = \\
 &P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild} | \text{Play}=\text{Yes}) P(\text{Play}=\text{Yes}) \\
 &/ P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild}) \\
 &= 0.1936 * 0.64 / 0.1224 = 1
 \end{aligned}$$

$$\begin{aligned}
 &P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild} | \text{Play}=\text{Yes}) = \\
 &P(\text{Overcast}|\text{Yes}) P(\text{Mild}|\text{Yes}) = 0.44 * 0.44 = 0.1936
 \end{aligned}$$

$$\begin{aligned}
 &P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild}) = \\
 &P(\text{Weather}=\text{Overcast}) P(\text{Temp}=\text{Mild}) = (4 / 14) * (6 / 14) \\
 &= 0.1224
 \end{aligned}$$

Unit 03 | Naive Bayes Classification

3-3) Naive Bayes Classification 연습

Weather (날씨)	Temp (온도)	Play (경기 여부)
sunny	hot	No
sunny	hot	No
overcast	hot	Yes
rainy	mild	Yes
rainy	cool	Yes
rainy	cool	No
overcast	cool	Yes
sunny	mild	No
sunny	cool	Yes
rainy	mild	Yes
sunny	mild	Yes
overcast	mild	Yes
overcast	hot	Yes
rainy	mild	No

야외 스포츠 경기가 열리기 좋은 날씨인가? 아닌가?
 날씨가 overcast, 기온이 mild일 때 경기가 없을 확률은?

1. 사전 확률

$$P(\text{No}) = 5 / 14 = 0.36$$

2. 사후 확률

$$P(\text{Overcast}|\text{No}) = 0 / 5 = 0$$

$$P(\text{Mild}|\text{No}) = 2 / 5 = 0.4$$

3. 베이즈 공식에 대입

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

$$P(\text{Play}=\text{No} \mid \text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild}) = \frac{P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild} \mid \text{Play}=\text{No}) P(\text{Play}=\text{No})}{P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild})}$$

$$= 0 * 0.36 / 0.1224 = 0$$

Likelihood가 0이 되면서
 결과값이 0이 되는 경우 발생!

$$\frac{P(\text{Weather}=\text{Overcast}, \text{Temp}=\text{Mild} \mid \text{Play}=\text{No})}{P(\text{Overcast}|\text{No}) P(\text{Mild}|\text{No})} = \frac{0}{0 * 0.4} = 0$$

Unit 03 | Naive Bayes Classification

3-4) 라플라스 스무딩 (Laplace Smoothing)

likelihood가 0이 되는 것을 방지하도록 최소한의 확률을 정해주자!!

$$P_{LAP} = \frac{c(x) + 1}{\sum_x [c(x)+1]}$$

실제보다 한 번씩 더 관찰되었다고 가정하기

$$P(x|c) = \frac{\text{count}(x, c) + 1}{\sum_{x \in v} \text{count}(x, c) + v}$$

분자에 1 더하고
분모에 입력변수들의 개수 v 더하는 꼴

Unit 03 | Naive Bayes Classification

3-4) 라플라스 스무딩 (Laplace Smoothing)

Ex. “단어가 스포츠를 나타내는 단어인가, 아닌가?”에 대한 NB 케이스
14개 단어로 이루어진 데이터

Word	P(word Sports)	P(word Not Sports)
a	$\frac{2 + 1}{11 + 14}$	$\frac{1 + 1}{9 + 14}$
very	$\frac{1 + 1}{11 + 14}$	$\frac{0 + 1}{9 + 14}$
close	$\frac{0 + 1}{11 + 14}$	$\frac{1 + 1}{9 + 14}$
game	$\frac{2 + 1}{11 + 14}$	$\frac{0 + 1}{9 + 14}$

정리

나이브 베이즈

장점

- 입력 공간의 차원이 높을 때 유리
- 텍스트에서 강점
- 가우시안 나이브베이즈를 활용하면 input이 연속형일때도 사용가능

단점

- 희귀한 확률이 나왔을 때 (라플라스 스무딩)
- 조건부 독립이라는 가정 자체가 비현실적

과제 및 데이터 설명

과제

week3_NaiveBayes_assignment.ipynb를 완성해주세요!!

Reference

참고자료

- 투빅스 12기 김태한님 강의자료
- 투빅스 13기 김미성님 강의자료
- 투빅스 15기 김동현님 강의자료
- <https://ratsgo.github.io/machine%20learning/2017/05/18/naive/>
- <https://datascienceschool.net/view-notebook/c19b48e3c7b048668f2bb0a113bd25f7/>
- <https://medium.com/@LSchultebraucks/gaussian-naive-bayes-19156306079b>
- https://www.edwith.org/machinelearning1_17/joinLectures/9738
- <https://www.youtube.com/watch?v=h09SVW6nnhM>
- <https://ratsgo.github.io/statistics/2017/09/23/MLE>
- https://yngie-c.github.io/machine%20learning/2020/04/08/naive_bayes/
- <https://bkshin.tistory.com/entry/dd?category=1042793>
- <https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-1%EB%82%98%EC%9D%B4%EB%B8%8C-%EB%B2%A0%EC%9D%B4%EC%A6%88-%EB%B6%84%EB%A5%98-Naive-Bayes-Classification>
- <https://m.blog.naver.com/61stu01/221277477927>