나이브 베이즈 알고리즘

베이즈 통계학의 창시자인

토머스 베이즈는

특별히 대학이나 연구기관에서 근무한적은 없었지만 뛰어난 수학자 였습니다

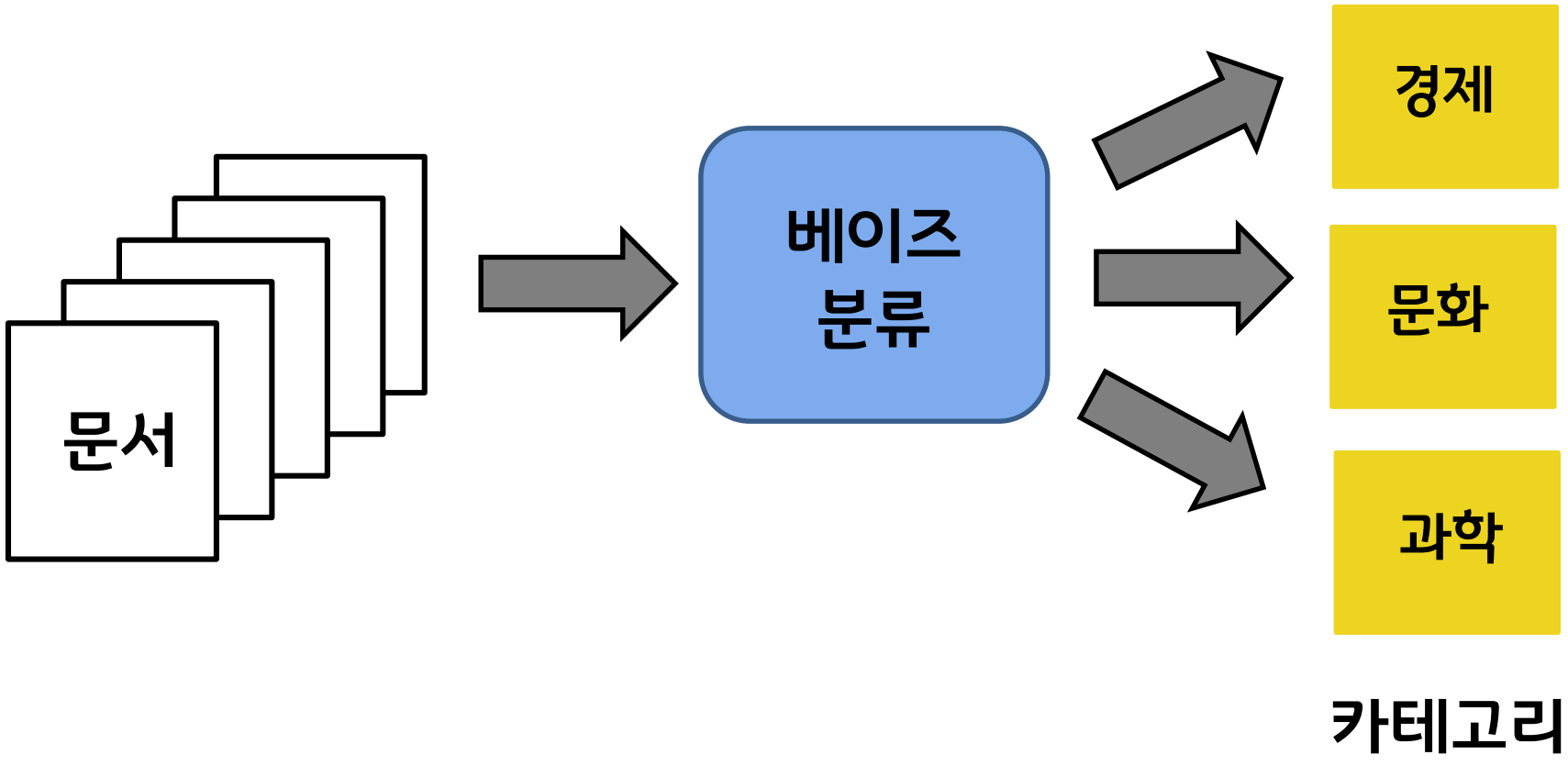
그가 남긴 가장 유명한 논문인
"확률론의 한 문제에 대한 에세이"는
그의 사후에 저명한 프랑스의 수학자인

라플라스에 의해 정리되고

베이즈 정리라는 이름이 붙여졌습니다

이 베이즈 이론은 21세기에 들어와
경제학, 정보과학, 심리학등
폭넓은 분야에서 급속하게 사용되기
시작했고 확률론, 통계론, 정보론에서
빼놓을 수 없는 입지를 굳혔습니다

베이즈 분류란 베이즈 이론을 이용해서 주어진 대상을 원하는 카테고리로 분류하는 방법을 말합니다



대표적으로 스팸메일을
분류할때 베이즈 분류를 사용하는데

이메일의 단어를 살펴보고

나이브 베이즈 필터는 대상이 되는 문서나 메일 속의 단어는 독립이라고 가정한다

문서

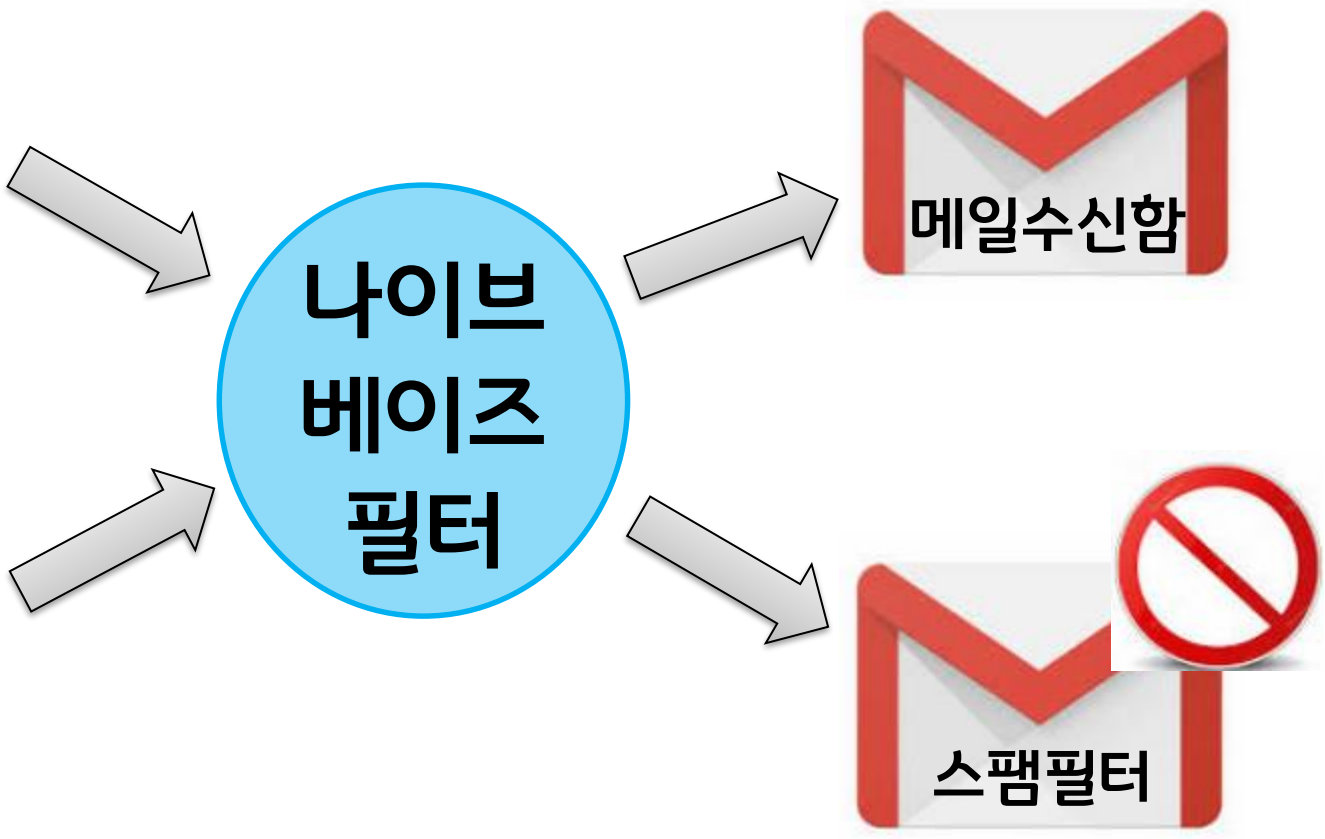
스팸성 단어들이 더 나왔을때 스팸일 확률이 높아짐으로 스팸으로 분류합니다

스팸메일

싸게 비아그라를 구입하기
정말 좋은 기회

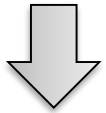
보통메일

비아그라 3정을 처방합니다

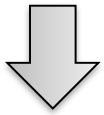


나이브 베이즈 분류를 이해하려면
아래의 내용을 순서데로 이해하면 됩니다

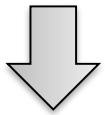
1. 확률의 시행과 사건



2. 결합 확률과 조건부 확률

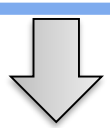


3. 독립사건과 종속사건

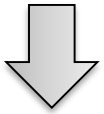


4. 베이즈 정리

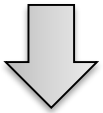
1. 확률의 시행과 사건



2. 결합 확률과 조건부 확률

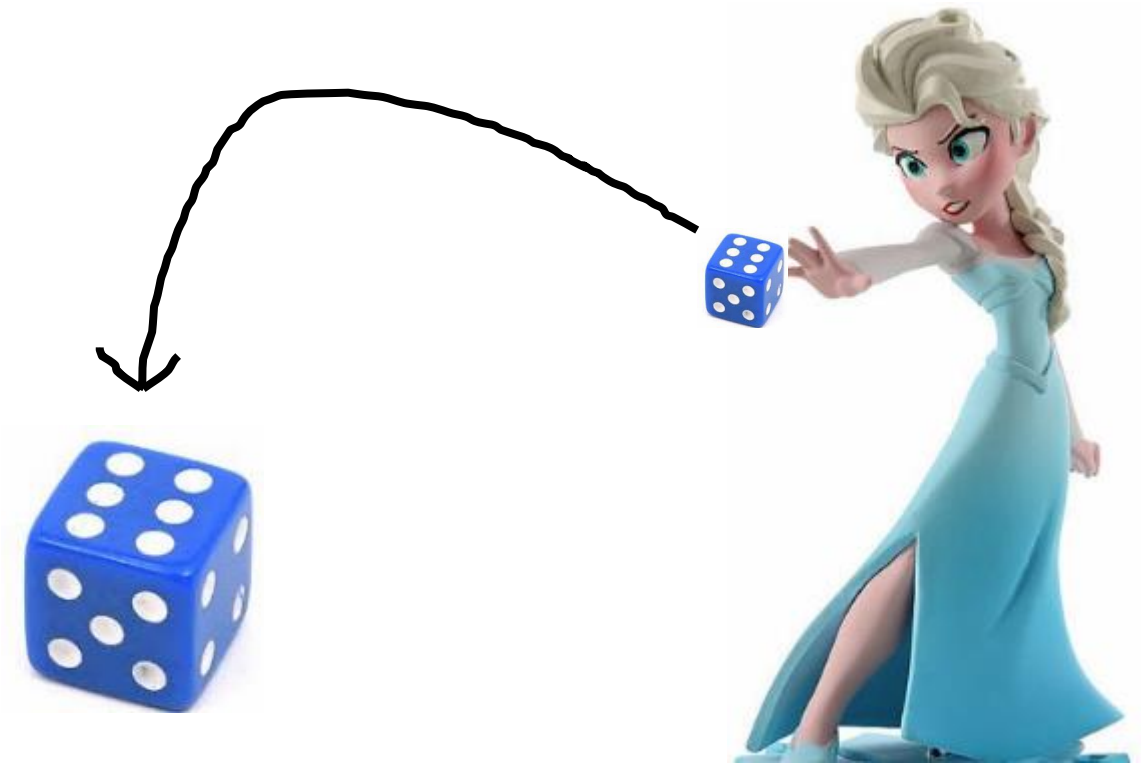


3. 독립사건과 종속사건

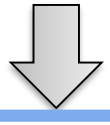


4. 베이즈 정리

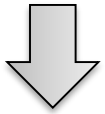
주사위를 던지는것을 시행이라고 하고
주사위를 던져서 6이 나온걸
사건이라고 합니다



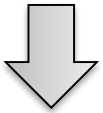
1. 확률의 시행과 사건



2. 결합 확률과 조건부 확률



3. 독립사건과 종속사건



4. 베이즈 정리

확률은
결합 확률과 조건부 확률이
있습니다

결합 확률은
두 사상 A와 B가 있을때
두 사상이
연속적으로 또는 동시에 일어나는
확률을 말합니다

예를들면
로또에 당첨될 사건과
벼락에 맞을 사건이
동시에 일어날 확률을 말합니다

표기는 이렇게 합니다

$$P(A \cap B)$$

로또에 당첨될 사건　　벼락에 맞을 사건

로또에 당첨되었는데 바로 벼락에 맞은겁니다

조건부 확률은
어떠한 상황이 주어졌을때
그 상황속에서 다른 상황이
일어날 확률을 말합니다

예를들면
비가 오는 사건이 일어나는 경우하에
우산이 팔리는 사건이
일어나는 경우를 말합니다

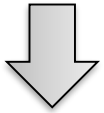
표기는 이렇게 합니다

$$P(A \mid B)$$

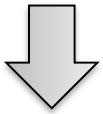
우산이 팔릴 사건　　비가 올 사건

비가온다는 조건하에
우산이 팔릴 확률입니다

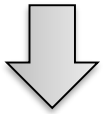
1. 확률의 시행과 사건



2. 결합 확률과 조건부 확률



3. 독립사건과 종속사건



4. 베이즈 정리

사건은
독립사건과 종속사건으로
나뉘는데

독립사건이란
두개의 사건이 일어났는데
두사건이 전혀 연관이 되지 않았다면
독립사건입니다

예를 들면
동전던지기의 결과와 화창한 날씨와는
서로 독립적입니다

동전 던지기의 결과가 앞면이 나올 사건을 A 라고 하면 P(A) 는 동전을 던졌을때 앞면이 나올 확률입니다

날씨가 화창할 사건을 B라고 하면
P(B)는 날씨가 화창할 확률입니다

그런데
이 두사건은 서로 독립적입니다
날씨가 화창하다고 동전 던지기의
결과가 꼭 앞면이 나오는것은 아닙니다

사건 B

사건 A

독립사건을 조건부 확률로 나타내면 아래와 같습니다

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

P( | ) = P()

사건B는 사건A에 전혀 영향을 주지 않습니다

종속사건은
사건 B 가 일어났을 경우와
일어나지 않았을 경우에 따라서
사건 A 가 일어날 확률이 다를때
A 는 B 의 종속사건이라고 합니다

비가오는 사건과
우산이 팔릴 사건으로

예를 들어 보면

비가오면 우산이 팔릴 확률이
높아지므로
두 사건의 관계는 종속 관계라고
할 수 있습니다

종속사건을 조건부 확률로 나타내면?

$$P(A \mid B)$$

우산이 팔릴 사건　　비가 올 사건

이렇게 나타냅니다

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

사건A와 사건B가
동시에 일어날 결합확률

사건 B가 일어날 확률

정리하면

화창한 날씨에 동전던지기

독립사건의 조건부 확률 :

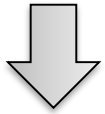
$$P(A|B) = P(A)$$

비가올때 우산 팔기

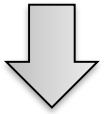
종속사건의 조건부 확률 :

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

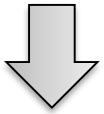
1. 확률의 시행과 사건



2. 결합 확률과 조건부 확률



3. 독립사건과 종속사건



4. 베이즈 이론

종속 사건의 조건부 확률 공식에 양쪽에 P(B) 를 곱해보겠습니다

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

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

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

$P(A|B) * P(B) = P(A \cap B)$ 는

$P(B|A) * P(A) = P(A \cap B)$ 로 나타낼 수 있습니다

그러면 다시 P(A|B) 를 기준으로 식을 정리하면

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

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

베이즈 분류의 조건부 확률 공식입니다

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

이 공식에 사건만 바꿔보겠습니다

$$P(\text{간암}|\text{흡연}) = \frac{P(\text{흡연}|\text{간암}) * P(\text{간암})}{P(\text{흡연})}$$

흡연을 하는 사람이 간암에 걸릴 확률을 유추해볼 수 있습니다

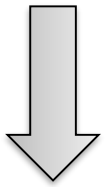
흡연을 하는 사람이 간암에 걸릴 확률을 구해보겠습니다

	흡연		
	YES	NO	총합
간암	4	16	20
정상	1	79	80
	5	95	

$$P(\text{간암}|\text{흡연}) = \frac{P(\text{흡연}|\text{간암}) \times P(\text{간암})}{P(\text{흡연})} = \frac{4/20 \times 20/100}{5/100} = 0.8$$

흡연을 하게 되면 간암에 걸릴 확률이 80% 입니다

그런데 나이브 베이즈의 장점이
간암의 요인을 흡연이외에 여러 개로
나열해서 추론할 수 있다는데 있습니다



P(간암 | 흡연, 음주, 직업, 성별,..)

나이브 베이즈 알고리즘은
단순하고 빠르며
매우 효과적이어서
1950년대 이후부터 현재까지
활발하게 연구되고 있습니다

문제

아래의 이원 교차표를 보고 손으로 계산해서 비아그라라는 단어가 포함되어져 있으면 스팸일 확률일 몇 % 인지 알아내시오 !

	비아그라		
	YES	NO	총합
스팸	4	16	20
햄	1	79	80
	5	95	

답

	비아그라		
	YES	NO	총합
스팸	4	16	20
햄	1	79	80
	5	95	

$$P(\text{스팸}|\text{비아그라}) = \frac{P(\text{비아그라}|\text{스팸}) * P(\text{스팸})}{P(\text{비아그라})}$$

$$= \frac{4/20 * 20/100}{5/100} = 0.8$$

비아그라라는 단어가 포함되어 있으면 스팸일 확률이 80% 나 되는구나 !

참고자료

1. 그림으로 설명하는 개념쏙쏙 통계학 성안당 - 와쿠이 요시유키 지음

2. R을 활용한 머신러닝 에이콘 - 블레트란츠 지음

3. EBS 고교강의 친절한 하영쌤의 수학

cafe.daum.net/oracleoracle

사랑하는 자여 네 영혼이 잘됨같이 네가 범사에 잘되고 강건하기를 내가 간구하노라

– 성경 요한삼서 1장 2절