인공지능수학 개론

(저자 김종락)

(임시적인) 목록

Chapter 8. 조건부 확률과 베이스 정리

확률은 우리 일상 생활에서 자주 등장한다. 유한한 표본 공간 S가 주어졌을 때 S의 부분집합 A (사건이라고 한다)가 일어날 확률은 $P(A)=\frac{|A|}{|S|}$가 된다. 예를 들어 서로 다른 동전 2개를 던졌을 때 일어날 수 있는 표본 공간은 $S=\{(H,H),(H,T),(T,H),(TT)\}$가 된다. 이때 두 동전 중 적어도 한 H가 나오는 사건 A의 확률은 $A=\{(H,H),(H,T),(T,H)\}$이므로 $P(A)=\frac{3}{4}$가 된다.

조건부 확률은 주어진 사건에 대하여 다른 사건이 일어날 확률을 말한다. 위의 예제에서 A라는 사건이 주어졌을 때, 두 동전 모두 H가 나오는 사건 B에 대한 확률을 $P(B|A)$라고 쓰고 $P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{|A\cap B|/|S|}{|A|/|S|}=\frac{1/4}{3/4}=1/3$가 된다.
조건부 확률과 밀접한 관계가 있는 것이 베이즈 정리(Bayes' theorem)이다.

이 장에서는 확률과 관련된 개념들을 소개하고 예제를 살펴본다.

■ 조건부 확률

● (정의) 사건 A가 주어졌을 때 사건 B가 일어날 확률을 **조건부 확률**이라고 하고, $P(B|A)$로 표시한다. 이는 아래와 같이 구해진다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (1)$$

만일 $P(B|A) = P(B)$일 때 즉 사건 A가 일어났을 때 사건 B가 일어날 확률이 사건 가 일어나는 확률과 같을 때 두 사건 A, B를 **독립**이라고 한다. 즉

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

을 만족한다.

● (베이스 정리) $P(B|A) = \frac{P(A \cap B)}{P(A)}$을 다시 쓰면

$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$가 된다. 따라서

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ (단 } P(B) \neq 0)$$

를 얻게 된다. 이것을 **베이스 정리** 또는 **베이스 룰**이라고 한다.

- $P(A|B)$를 B가 주어졌을 때 A의 사후확률 posterior probability이라고 한다.
- $P(B|A)$를 A의 우도likelihood라고 한다.
- $P(A)$를 사전 확률prior probability이라고 하고 $P(B)$를 주변우도marginal likelihood라고 한다.
- 따라서 posterior는 likelihood × prior에 비례한다.

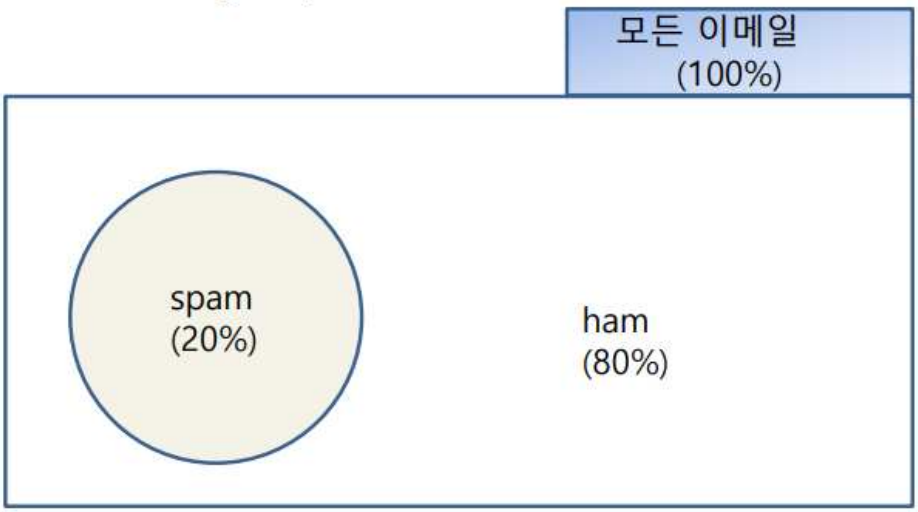
베이지언 룰의 기본적인 개념

- 베이지언 확률이론: 한 증거를 기반으로 한 사건의 유사성을 추정하는 개념

- 이메일이 스팸spam인지 햄ham(스팸의 반대말)인지 구별하는 데 사용된다.

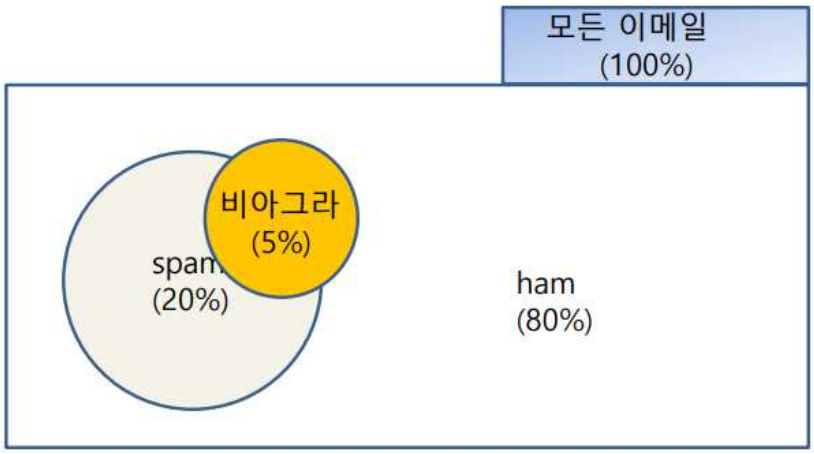
- 어떤 이메일이 왔다고 하자. 이것은 스팸이거나 햄이 된다.

- 따라서 Probability(ham)= 1-Probability(spam)

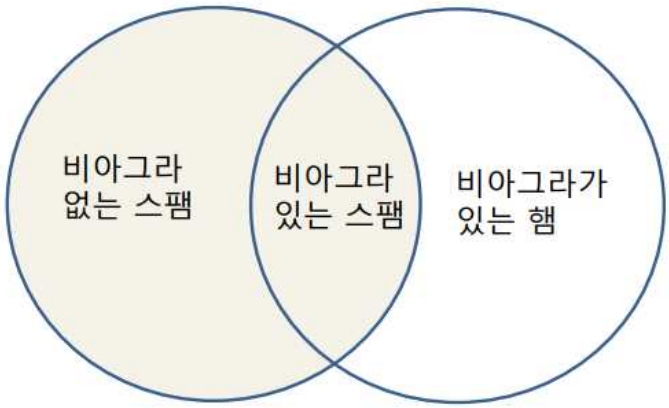


결합 확률

- '비아그라'라는 낯선 단어가 존재할 때^^



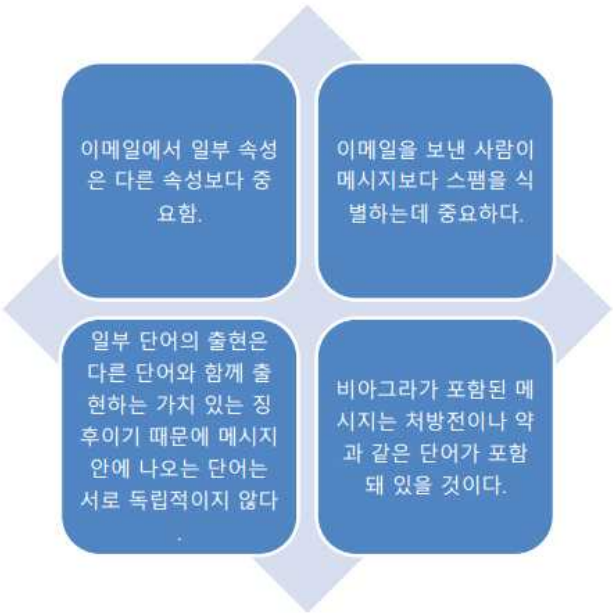
- 벤 다이어그램으로 표현된 세 가지 경우



나이브 베이즈 알고리즘

장점	단점
• 단순하고 빠르며 매우 효과적임	• 모든 속성은 동등하게 중요하고 독립적이라는 결함 가정(often-faulty assumption) 의존
• 노이즈와 결측 데이터가 있어도 잘 수행함	• 수치 속성으로 구성된 많은 데이터 셋에 대해 이상적이지 않음
• 훈련에 대한 상대적으로 적은 예제가 필요하지만 매우 많은 예제도 잘 수행함	• 추정된 확률은 예측된 범주보다 덜 신뢰적임.
• 예측에 대한 추정된 확률을 얻기 쉬움	

often-faulty assumption부가 설명



우도표likelihood table

우도	비아그라(W1)		돈(W2)		식료품(W3)		주소 삭제(W4)		총합
포함여부	Yes	No	Yes	No	Yes	No	Yes	No	
스팸	4/20	6/20	10/20	10/20	0/20	20/20	12/20	8/20	20
햄	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
총합	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

이메일 내용에서 W1=Yes, W2=No, W3=No, W4=Yes를 받았을 경우 이것이 스팸일 확률은?

$$P(\text{spam}|W1 \cap \sim W2 \cap \sim W3 \cap W4) = \frac{P(W1 \cap \sim W2 \cap \sim W3 \cap W4|\text{spam})P(\text{spam})}{P(W1 \cap \sim W2 \cap \sim W3 \cap W4)}$$

각각의 사후 확률 계산

- $P(\text{spam}|W1 \cap \sim W2 \cap \sim W3 \cap W4) = \frac{P(W1 \cap \sim W2 \cap \sim W3 \cap W4|\text{spam})P(\text{spam})}{P(W1 \cap \sim W2 \cap \sim W3 \cap W4)}$

  $= \frac{P(W1|spam)P(\sim W2|spam)P(\sim W3|spam)P(W4|spam)P(spam)}{P(w1)P(\sim W2)P(\sim W3)P(W4)}$

  $= \frac{4/20 \times 10/20 \times 20/20 \times 12/20 \times 20/100}{P(w1)P(\sim W2)P(\sim W3)P(W4)} = \frac{0.012}{P(w1)P(\sim W2)P(\sim W3)P(W4)}$

- $P(\text{ham}|W1 \cap \sim W2 \cap \sim W3 \cap W4) = \frac{P(W1 \cap \sim W2 \cap \sim W3 \cap W4|\text{ham})P(\text{ham})}{P(W1 \cap \sim W2 \cap \sim W3 \cap W4)}$

  $= \frac{1/80 \times 66/80 \times 71/80 \times 23/80 \times 80/100}{P(w1)P(\sim W2)P(\sim W3)P(W4)} = \frac{0.002}{P(w1)P(\sim W2)P(\sim W3)P(W4)}$

질문: 이 값들을 보고 이메일은 스팸일까 햄일까?

- 이전 슬라이드에서 분모는 공통이므로 분자 값의 비율을 살펴보자.
- 스팸의 비율: 0.012/(0.012+0.002)=0.857
- 햄의 비율: 0.002/(0.012+0.002)=0.143

- 즉 이 이메일은 85.7%의 확률로 스팸이며 14.3%의 확률로 햄이라고 할 수 있다. 따라서 스팸이라고 예측한다.

일반적인 베이즈 룰

- 일반적으로, 속성 F_1에서 F_n으로 주어진 증거를 고려해 범주 C에 대한 레벨 L, 즉 C_L 의 확률은:
- 범주 L로 조건 지어진 증거들$(F_i|C_L)$의 확률을 곱하고, 범주 L의 사전 확률 $P(C_L)$, 그리고 확률 갑을 변환하는 scaling factor $1/Z$를 전부 곱하는 것과 같다. 즉

- $P(C_L|F_1F_2,\dots,F_n) = \frac{1}{Z}P(C_L)\prod_{i=1}^{n}P(F_i|C_L)$

라플라스 추정기

- 다른 예제를 살펴보자.
- 비아그라=yes, 식료품=yes, 주소 삭제=yes 일 때 P(spam|~)을 계산하면
- 4/20 * 10/20 * 0/20 * 12/20 * 20/100=0
- 이것은 우리의 직관과 배치된다(왜?)
- 이를 방지하기 위하여 라플라스 추정기를 1로 설정한다. 이 경우
- P(spam|~)=5/24 * 11/24 * 1/24* 13/24* 20/100=0.0004
- P(ham|~)=2/84 * 15/84 * 9/84* 24/84* 80/100=0.0001
- 따라서 이 경우 스팸일 확률이 80%, 햄일 확률은 20%이다.

■ (연관규칙)

연관 규칙이란?

- {땅콩버터, 젤리} -> {빵}
  땅콩버터와 젤리를 사면 빵도 구매한다.
- 이런 것을 연관 규칙이라고 한다.
- 연관 규칙은 자율학습unsupervised이다.

연관 규칙의 적용사례

- 암 데이터 분석에서 단백질 서열과 자주 발견되는 흥미로운 DNA의 패턴 찾기
- 구매 패턴, 사기 신용카드나 보험과 복합해 발생하는 의료 청구 발견하기
- 휴대폰 서비스를 정지하거나 케이블 TV 패키지를 업그레이드하려는 행위 예측하기

연관 규칙 학습을 위한 아프리오리 알고리즘

- k개의 물건에 대한 모든 가능한 연관 관계는 2^k개 있다. 즉 모든 subsets의 개수와 같다 (why?)

- 좀 더 영리한 방법으로 조합의 수를 줄일 필요가 있다.

아프리오리Apriori 방법

- R. Agrawal & R. Srikant가 소개

'Fast algorithms for mining association rule, in Proceedings of the 20th International Conference on Very Large Databases' (1994), 487-499, available on google

장점	단점
• 매우 대량의 거래 데이터와 작동이 이상적으로 적합하다.	• 작은 데이터셋에 유용하지 않다.
• 쉽게 이해할 수 있는 결과를 내놓는다.	• 상식과 통찰력을 구별해야 한다.
• 데이터 마이닝에 유용하고 데이터베이스안에 예상하지 못한 지식을 발견한다.	• 무작위 패턴에서 거짓된 결과를 끄집어 내기 쉽다.

간단한 상황 예제

거래번호	구매 물품
1	{꽃, 병문안 카드, 소다}
2	{플러시 곰 인형, 꽃, 풍선, 사탕}
3	{병문안 카드, 사탕, 꽃}
4	{플러시 곰 인형, 풍선, 소다}
5	{꽃, 병문안 카드, 소다}

1. 아픈 친구나 가족을 방문하는 사람은 병문안 카드나 꽃을 사려는 경향
2. 새엄마를 방문하는 방문객은 플러시 곰 인형이나 풍선을 사는 경향

규칙 흥미 측정: 지지도와 신뢰도

- 아이템셋 {병문안 카드, 꽃}은 3/5=0.6의 지지도를 갖는다(5번 구매중 3번 나타나기 때문)
- {병문안 카드} -> {꽃}에 대한 지지지도 0.6
- {사탕}에 대한 지지도는 2/5=0.4이다.

- 지지도 정의

$$\text{support(X)} = \frac{\text{count}(X)}{N}$$

(N은 데이터베이스 거래 건수, count(X)는 그 중 아이템셋의 거래 개수)

규칙 흥미 측정: 지지도와 신뢰도

- 신뢰도 정의

$$\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X,Y)}{\text{support}(X)} = P(Y|X)$$

즉 신뢰도는 아이템셋 X의 물품으로부터 아이템셋 Y의 물품을 만들어 내는 거래의 비율을 말한다.

- {꽃}->{병문안 카드}의 신뢰도는 0.6/0.8=0.75
- {병문안 카드}->{꽃}의 신뢰도는 0.6/0.6=1이고 지지도는 06이므로 이런 것을 강한 규칙 strong rule이라고 한다.

아프리오리 원칙과 규칙 집합 생성

- 아프리오리Apriori 원칙은 빈번한 아이템셋의 모든 부분집합은 빈번해야 한다는 점
- 즉 {A,B}가 빈번하다면 {A}, {B} 둘 다 빈번해야 한다는 점이다.
- {A}가 원하는 지지도 경계 값을 충족하지 못함을 알면 {A,B}나 {A}를 포함한 아이템셋도 고려할 여지가 없다.
- {A}, {B}, {C}는 반복 1에서 빈번한 반면, {D}는 빈번하지 않았다고 하자. 그러면 반복 2에서 {A,B}, {A,C}, {B,C}만 고려한다. 이런식으로 고려할 가지수를 줄여 나간다.

예제: 연관 규칙과 자주 구매하는 식료품 식별

- 장바구니 분석: 구매자들이 아침에 패스트리와 오렌지 주스, 커피를 함께 자주 구매한다는 것을 알 수 있다면 이들을 같이 두는 것이 마케팅의 전략이 될 수 있다.
- 이런 기법은 영화 추천, 데이트 장소 추천, 의료품 중의 위험스러운 상호 작용 같은 여러 문제에 적용할 수 있다.