나이브 베이즈 분류

확률적 학습

학습내용

- 확률의 기본 원리
- R로 텍스트 데이터를 분석하는 데 필요한 특화된 방법과 데이터 구조
- 나이브 베이즈를 이용한 SMS 스팸 메시지 필터의 구축 방법

나이브 베이즈의 이해

- 베이지안 기법 기반의 분류기
  - 훈련 데이터를 활용해 특징 값이 제공하는 증거를 기반으로 결과가 관측될 확률을 계산.
  - 나중에 분류기가 레이블이 없는 데이터에 적용될 때 결과가 관측될 확률을 이용해서 새로운 특징에 가장 유력한 클래스를 예측.
  - 사례
    - 스팸 이메일 필터링과 같은 텍스트 분류
    - 컴퓨터 네트워크에서 침입이나 비정상 행위 탐지
    - 일련의 관찰된 증상에 대한 의학적 질병 진단
  - 결과에 대한 전체 확률을 추정하고자 동시에 여러 속성 정보를 고려해야만 하는 문제에 가장 적합.
  - 영향력이 약한 개별의 속성들을 모두 결합하면 꽤 큰 영향을 미칠수도 있다.

베이즈 분류

- 베이지안 분류자(Bayesian Classifier)
  - 베이지안 분류자는 통계적인 분류자로, 주어진 인스턴스가 특정 클래스에 들어갈 가능성을 계산해서 클래스 배정 확률을 예측한다.
    - 베이즈 이론에 바탕을 두고 있다.
  - 클래스 분류 알고리즘을 비교하는 연구 과정에서 나이브 베이지안 분류자라고 불리는 단순 베이지안 분류자로 결정 트리와 지정한 신경망 분류자를 비교할 수 있다는 점을 발견.
  - 나이브 베이지안 분류자는 하나의 속성 값을 기준으로, 다른 속성이 독립적이라 전제했을 때 해당 속성 값이 클래스 분류에 미치는 영향을 측정한다.
    - 나이브(naïve) : 단순한, 순수한, 고지식한
    - 조건부 독립성은 계산을 단순화하기 때문에 나이브라는 이름이 붙었다.
  - 대규모 데이터베이스에 적용했을 때 높은 정확성과 빠른 속도가 장점이다.

베이즈 이론

- 18세기 확률과 결정 이론의 토대를 마련한 영국의 장로파 성직자 토마스 베이즈(Thomas Bayes)의 이름을 따서 명명된 이론.

- 베이지안 기법(Bayesian methods)
  - 여러 속성 값을 포함하고 있는 훈련 데이터 인스턴스 X가 있다고 하자. 베이즈 이론 식으로 풀어보자면,
    - 이 X는 어떤 가설(가설 H : 데이터 인스턴스 X가 특정 클래스 C에 속한다) 의 "증거"다.
    - 이미 X라는 증거를 확인한 가설 H가 참일 확률 P(H|X), 즉 데이터 인스턴스 X와 그 속성을 알고 있을 때 해당 인스턴스가 클래스 C에 속할 확률을 구하려는 것이다.
    - 확률 P(H|X)는 X를 전제했을 때 H일 확률, 즉 조건부 확률(posterior probability)이다.
    - [교재 예제] X=비아그라, H=비아그라 단어가 포함된 메시지는 스팸 메일이라는 가설, 이 때 X의 속성 조건을 이미 알고 있을 때의 H일 확률이 P(H|X)이다.
  - P(H)는 사전확률(prior probability) : 조건에 상관없이 스팸 메일일 확률
  - P(H|X)는 사전확률 P(H)보다 독립 사건 X만큼 더 많은 정보를 바탕으로 판별한다.
  - P(X|H)는 H를 전제했을 때 X일 확률, 즉 스팸 메일이라고 가정할 때, 비아그라 단어가 있을 확률
  - P(X)는 X에 대한 사전 확률, 비아그라 단어가 포함될 확률
  - P(H|X)= P(X|H)P(H) / P(X)

햄(ham) : 대개 스팸 메시지와 연관이 있는 키워드를 하나 이상 포함하고 있어서 스팸이 아님에도 차단 또는 여과된 이메일 메시지

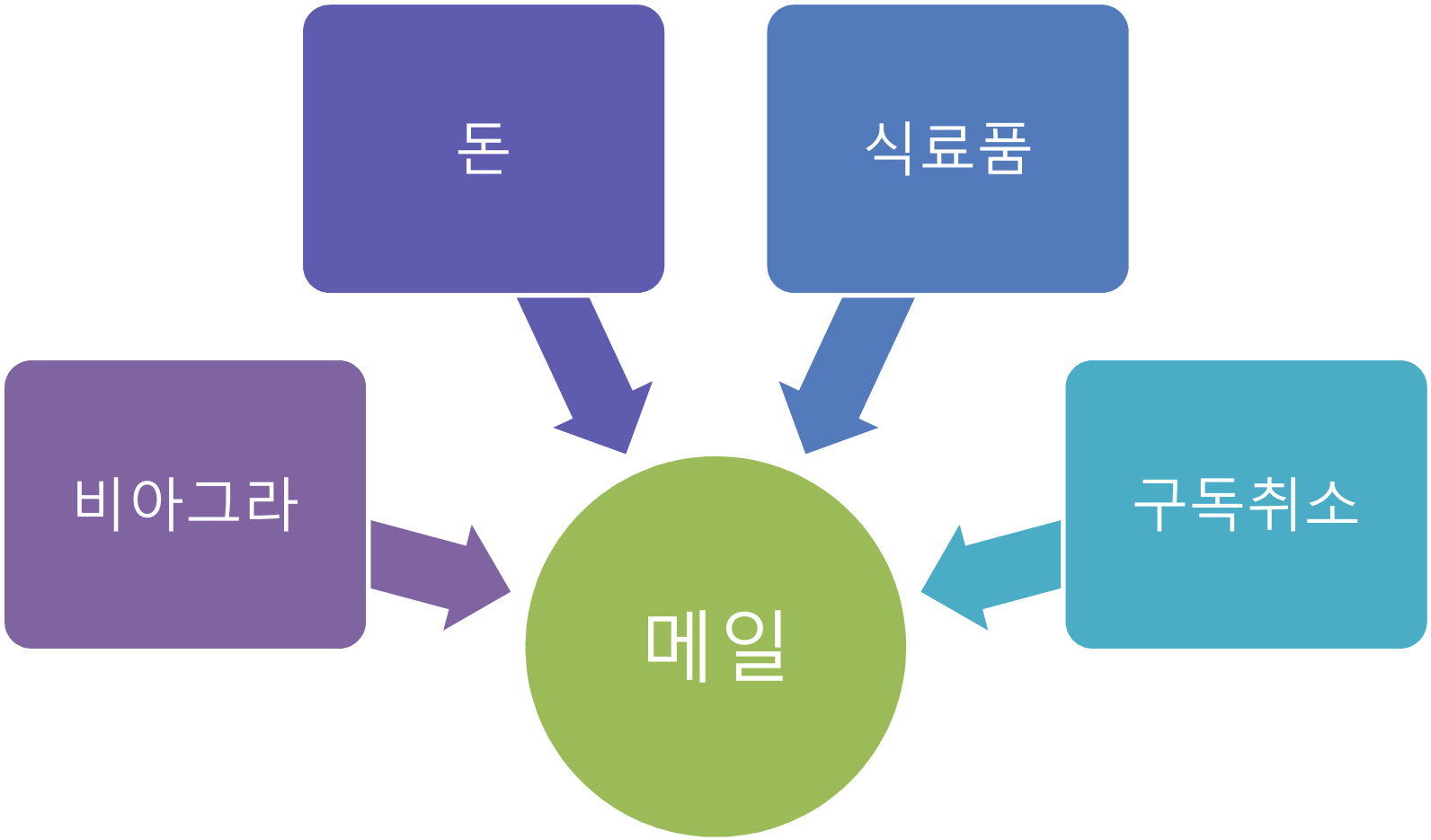
속성 →
증거 →

비아그라(W_1)	돈(W_2)	식료품(W_3)	구독 취소(W_4)	메일
예	아니오	아니오	예	스팸
아니오	예	예	아니오	햄
예	아니오	아니오	예	햄
아니오	아니오	아니오	예	햄
아니오	예	아니오	아니오	스팸
아니오	아니오	아니오	예	스팸
아니오	아니오	아니오	아니오	스팸
아니오	아니오	아니오	아니오	햄

50개의 e-mail 중 10개가 spam mail 이라면 받은 메시지가 스팸일 확률

P(스팸) = 0.2

P(햄) = 1 - 0.2 = 0.8


돈
식료품
비아그라
메일
구독취소

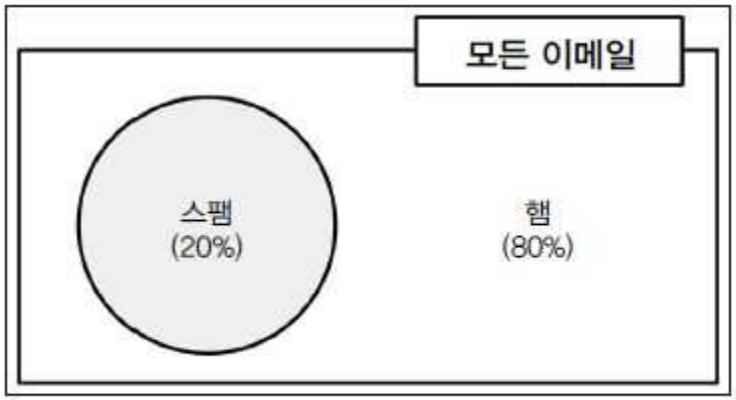


그림 4.1: 모든 이메일 확률 공간은 스팸과 햄의 분할로 시각화할 수 있다.

메시지에 '비아그라' 포함

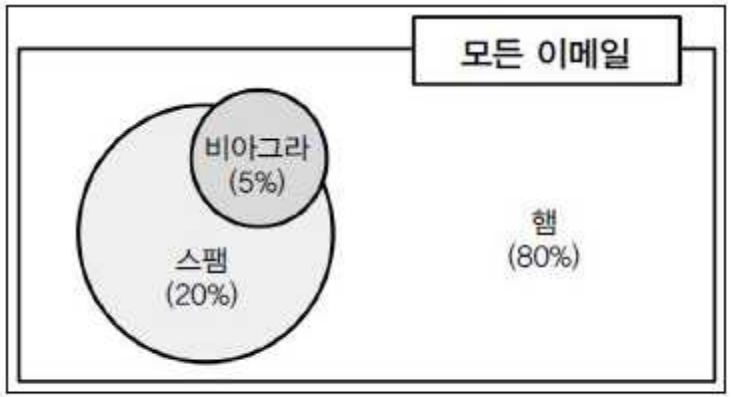


그림 4.2: 상호 배타적이지 않은 사건은 겹쳐진 부분으로 나타낼 수 있다.

'비아그라'가 포함된 사건;
모든 스팸 메시지가 단어 '비아그라'를 포함하는 것도 아니고
단어 '비아그라'가 있는 모든 이메일이 스팸인 것도 아니다.

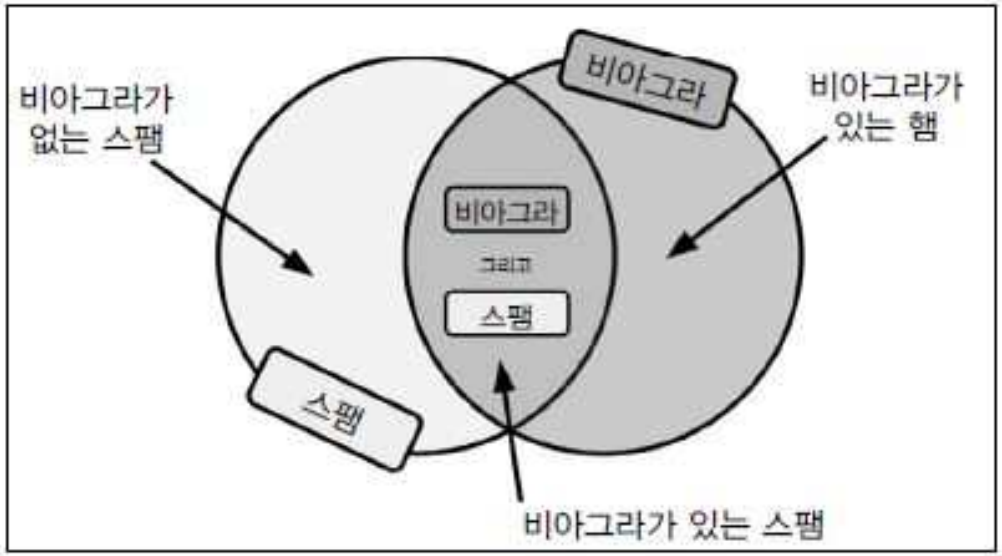


그림 4.3: 벤 다이어그램으로 스팸과 비아그라 사건의 중첩을 보여준다.

두 사건의 결합확률 : P(스팸 ∩ 비아그라)

독립사건 : P(스팸 ∩ 비아그라) = P(스팸) × P(비아그라)

종속사건 : ?

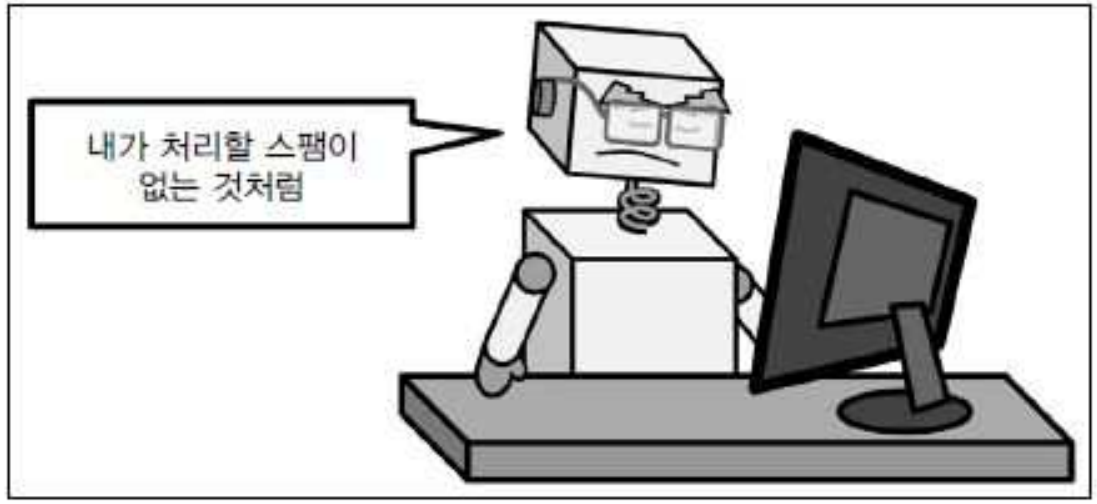


그림 4.4: 기계가 유용한 패턴을 식별하는 방법을 학습하려면 종속된 사건이 필요하다.

베이즈 정리 : 종속사건 간의 관계는 베이즈 정리를 이용해 설명 가능.
다른 사건이 제공하는 증거를 고려해 한 사건에 대한 확률 추정을 어떻게 바꿀지에 대해 사고하는 방식을 알려준다.

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

P(A|B) : 사건 B가 발생한 경우 사건 A의 확률

조건부 확률 : 사건 A의 확률이 사건 B가 발생한 경우에 종속적이다.

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

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

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

사전확률(Prior Probability) : 사건 발생 전 확률

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

$P(B)$ → 상수

사후확률(Posterior Probability) : 사건 발생 후 확률

우도(likelihood)

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

주변 우도(marginal likelihood)

각 증거에 메시지가 스팸이 될 확률을 측정한 사후 확률이 50%보다 크면 햄보다 스팸이 될 가능성이 크다

	비아그라		
빈도	예	아니오	총계
스팸	4	16	20
햄	1	79	80
총계	5	95	100

	비아그라		
우도	예	아니오	총계
스팸	4 / 20	16 / 20	20
햄	1 / 80	79 / 80	80
총계	5 / 100	95 / 100	100

그림 4.5: 스팸 사후 확률 계산에서 빈도와 우도표는 기반이 된다.

P(비아그라=예 | 스팸) = 4/20 = 0.2

P(스팸 ∩ 비아그라) = P(비아그라 | 스팸) × P(스팸) = (4/20) × (20/100) = 0.04

P(스팸 | 비아그라) = P(비아그라 | 스팸) × P(스팸) / P(비아그라)

= (4/20) × (20/100) / (5/100) = 0.8

따라서, 메시지가 단어 비아그라를 포함했을 때 메시지가 스팸일 확률은 80%이다.

비아그라를 포함하는 어떤 메시지도 필터링 되어야 한다는 것을 생각할 수 있다.

나이브 베이지안 분류자

나이브 베이즈 알고리즘

- 클래스 라벨이 매겨진 훈련 인스턴스 세트 D가 있다. 인스턴스마다 n개의 속성에 대한 n차원의 속성 벡터 $X=(X_1, X_2, ..., X_n)$을 담고 있다.

- m개의 클래스 라벨 $C_1, C_2, ..., C_m$이 있다고 하자. 나이브 베이지안 분류자는 주어진 인스턴스 X에 대해 X가 매겨져 있는 현재의 클래스에 포함될 가능성이 가장 높다는 전제하에 조건부 확률을 예측한다.

$$P(C_i|X) = P(X|C_i)P(C_i) / P(X)$$

- P(X)는 모든 클래스에 대해 상수이므로 $P(X|C_i)P(C_i)$가 최대라면 $P(C_i|X)$도 최대가 된다. 클래스의 사전 확률을 알 수 없는 상태라면 모든 클래스가 동일한 가능성을 갖는다고 가정한다.

나이브 베이즈안 분류자

- 훈련 데이터에는 딸린 속성이 많아 모든 인스턴스에 대해 $P(X|C_i)$를 계산하려면 연산량 부담이 크다. 따라서 연산 부담을 경감하기 위해 클래스 조건 독립성이란 간단 명료한 기본 가정을 둔다. 클래스 조건 독립성 전제는 하나의 속성값이 인스턴스의 클래스 라벨 결정에 대해 다른 속성의 영향을 받지 않는다는 것(=속성 사이에 의존 관계가 없다)을 말한다.
  - 속성값이 범주형 속성이라면(불연속/순서없음), 클래스 라벨이 C_i인 인스턴스 개수
  - 속성값이 연속형 속성이라면, 평균 μ, 표준편차 σ인 표준 정규분포를 따른다고 가정한다.

$$P(x_k|C_i)=g(x_k, \mu_{ci}, \sigma_{ci})$$

- 훈련 인스턴스 X의 클래스 라벨을 예측하려면 모든 클래스 C_i에 대한 $P(X|C_i)P(C_i)$를 계산해야 한다. $P(X|C_i)P(C_i)$이 최대인 클래스 C_i를 찾아 해당 클래스로 라벨을 매긴다.

	비아그라(W_1)		돈(W_2)		식료품(W_3)		구독 취소(W_4)		
우도	예	아니오	예	아니오	예	아니오	예	아니오	총계
스팸	4 / 20	16 / 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
총계	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

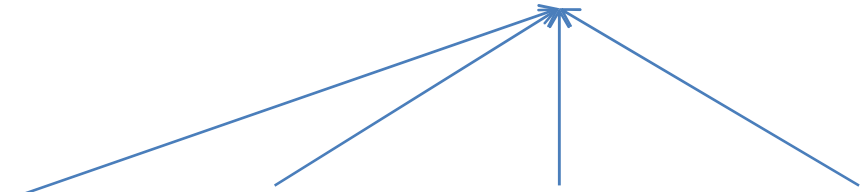
그림 4.6: 스팸과 햄 메시지에 추가 항목으로 우도를 첨가한 확장표

P(스팸 | 비아그라=예, 돈=아니오, 식료품=아니오, 구독취소=예)

$$= \frac{P(\text{비아그라} = \text{예}, \text{돈} = \text{아니오}, \text{식료품} = \text{아니오}, \text{구독취소} = \text{예} \mid \text{스팸}) \times P(\text{스팸})}{P(\text{비아그라} = \text{예}, \text{돈} = \text{아니오}, \text{식료품} = \text{아니오}, \text{구독취소} = \text{예})}$$

- 특징이 추가되면 가능한 모든 사건 교집합에 대해 확률을 저장해야 하므로 엄청난 양의 메모리가 필요. -> $P(A \cap B) = P(A) \times P(B)$
- 과거 데이터에서는 전혀 관측한 적도 없을 경우, 결합 확률은 0 -> 보정 필요.

클래스 조건부 독립(class conditional independence)



비아그라(W_1)	돈(W_2)	식료품(W_3)	구독 취소(W_4)	메일
예	아니오	아니오	예	스팸
아니오	예	예	아니오	햄
예	아니오	아니오	예	햄
아니오	아니오	아니오	예	햄
아니오	예	아니오	아니오	스팸
아니오	아니오	아니오	예	스팸
아니오	아니오	아니오	아니오	스팸
아니오	아니오	아니오	아니오	햄

P(스팸 | 비아그라=예, 돈=아니오, 식료품=아니오, 구독취소=예)

$$= \frac{P(\text{비아그라} = \text{예}, \text{돈} = \text{아니오}, \text{식료품} = \text{아니오}, \text{구독취소} = \text{예} \mid \text{스팸}) \times P(\text{스팸})}{P(\text{비아그라} = \text{예}, \text{돈} = \text{아니오}, \text{식료품} = \text{아니오}, \text{구독취소} = \text{예})}$$

상수

P(스팸 | 비아그라=예, 돈=아니오, 식료품=아니오, 구독취소=예)

$\propto$

P(비아그라=예|스팸) × P(돈=아니오|스팸) × P(식료품=아니오|스팸) × P(구독취소=예|스팸) × P(스팸)

= (4/20) × (10/20) × (20/20) × (12/20) × (20/100) = 0.012

$$P(\text{햄} \mid \text{비아그라=예, 돈=아니오, 식료품=아니오, 구독취소=예})$$

$$\propto$$

$$P(\text{비아그라=예}|\text{햄}) \times P(\text{돈=아니오}|\text{햄}) \times P(\text{식료품=아니오}|\text{햄}) \times P(\text{구독취소=예}|\text{햄}) \times P(\text{햄})$$

$$= (1/80) \times (66/80) \times (71/80) \times (23/80) \times (80/100) = 0.002$$

0.012 / 0.002 = 6 이므로 이 메시지가 스팸일 가능성이 햄일 가능성보다 6배 높다고 할 수 있다.

스팸일 확률 : 0.012 / (0.012+0.002) = 0.857

햄일 확률 : 0.002 / (0.012+0.002) = 0.143

➔ 확률이 큰 쪽으로 판별

라플라시안 보정

- 확률이 0인 경우에는 어떻게 계산해야 할까?
  - 라플라시안 보정(라플라스 추정) 확률계산기법
    - 프랑스 수학자 피에르 라플라스
    - 훈련 데이터베이스가 충분히 커서 인스턴스 한 개 정도를 추가해도 계산 확률에 미치는 영향이 미미하다고 하면, 모든 속성값에 인스턴스를 단 하나씩만 더한다면 확률 0문제를 피할 수 있다.
    - 이 때, 전부 q개의 가상 인스턴스를 추가했다면 분모에도 q를 더해서 전체 확률 계산을 보정해야 한다.

P(스팸 or 햄 | 비아그라=예, 돈=예, 식료품=예, 구독취소=예)

스팸일 확률

(4/20) × (10/20) × (0/20) × (12/20) × (20/100) = 0

라플라스 보정 후

(5/24) × (11/24) × (1/24) × (13/24) × (20/100) = 0.0004

햄일 확률

(1/80) × (14/80) × (8/80) × (23/80) × (80/100) = 0.00005

라플라스 보정 후

(2/84) × (15/84) × (9/84) × (24/84) × (80/100) = 0.0001

스팸일 확률 : 0.0004 / (0.0004+0.0001) = 0.8

햄일 확률 : 0.0001 / (0.0004+0.0001) = 0.2

수치 특성 이용

수치 특성을 이산화하는 방법 – 절단점(cut points)을 이용.

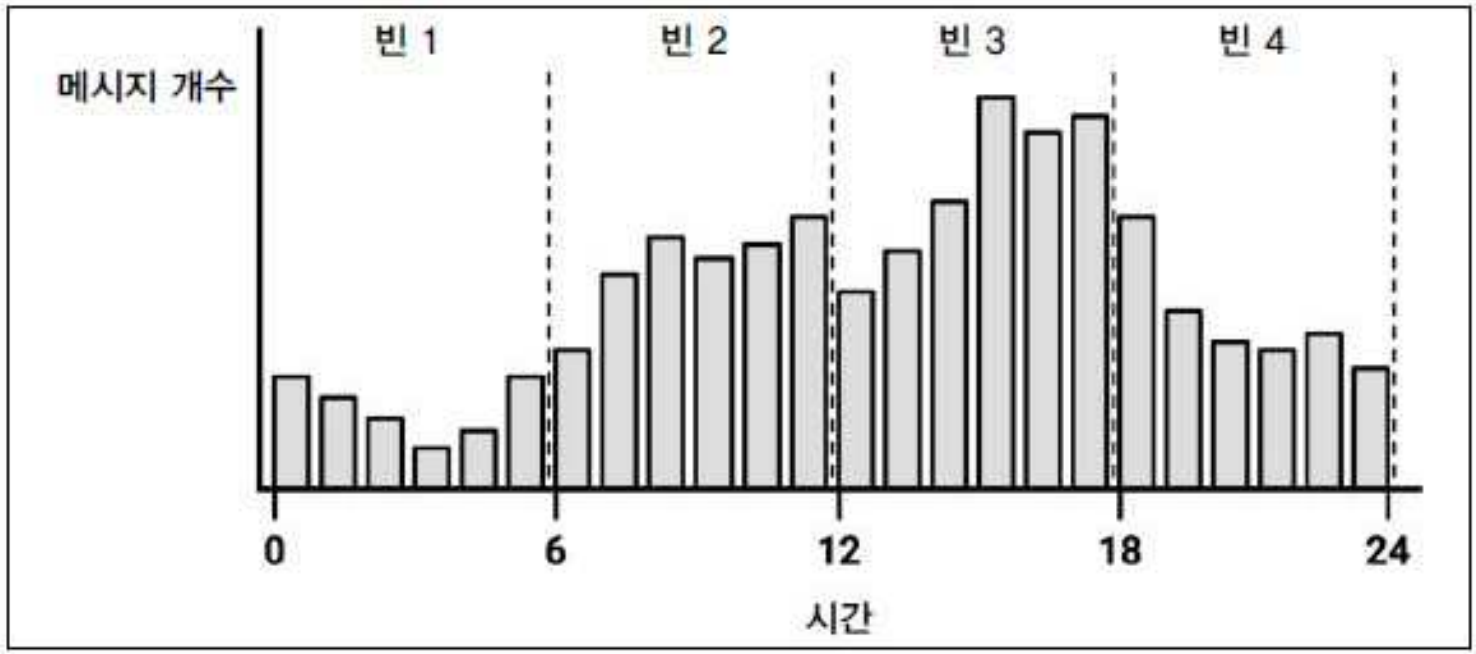


그림 4.7: 이메일 수신 시간에 따른 분포를 시각화한 히스토그램