2. 확률의 기초와 머신러닝 응용

필요성

- 머신러닝 모델은 모두 확률기반!
- 베이즈 정리와 마르코프 체인은 현대 딥러닝 기술의 기반이 되는 이론들!

확률이론

기본 개념 정리

- 확률실험 : 어떤 사건의 확률을 확인하기 위해 실험을 수행하는 것.
- 표본공간 : 확률실험에서 나올 수 있는 모든 경우를 포함하는 것.
- 사건 : 표본공간의 부분집합.
- $A \cap B$: 곱사건 / $A \cup B$: 합사건 / $S - A = A^C$: 여사건
- $A \cap B = \varnothing$: A와 B는 배반사건
- 확률 : 사건이 발생할 0~1 사이의 확률값.

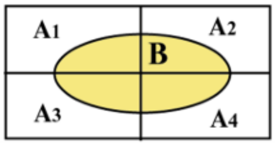
조건부 확률

- 모든 머신러닝 문제는 조건부 확률을 계산하는 문제.
    - ex) 신체의 여러 수치가 이러이러 할 때, (조건) 당뇨병에 걸릴 확률은?

- **조건부 확률** : A라는 사건(부분집합)을 표본공간이라고 생각하고 그 안에서 확률을 생각하는 것.

  $P(B|A) = \frac{P(B\cap A)}{P(A)}$: 사건 A가 발생했을 때의 사건 B의 확률

- **전확률공식 :** 사건 B의 전체 확률은 가능한 원인 (A_i) 별로 쪼개서 계산한 확률의 총합이다.



사건 $A_1, A_2, ..., A_n$이 표본공간 S의 분할이고 $P(A_i) > 0, (i = 1, 2, ..., n)$이면 임의의 사건 B에 대하여 다음 확률공식이 성립한다.

$$P(B) = \Sigma_{i=1}^{n} P(A_i)P(B|A_i)$$

ex) 인형 뽑기 기계 3개 중에서 인형 A를 뽑을 확률은 (기계1 에서 인형A를 뽑을 확률) + (기계2 에서 인형A를 뽑을 확률) + (기계3 에서 인형A를 뽑을 확률)

- **베이즈 정리 :** 어떤 사관이 관측됐을 때, 그 정보를 바탕으로 다른 사건의 확률을 계산하는 방법.

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

일반화 형태 : 사건 B가 일어났을 때, 해당 사건의 여러 원인 사건 A 중 A_k가 실제 원인일 확률을 구하는 것.

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\Sigma_{i=1}^{n} P(A_i)P(B|A_i)}$$

MLE와 MAP

- **MLE** : 관측된 결과가 가장 잘 설명될 수 있는 확률을 계산.

  ex) 동전을 10번 던져서 7번 앞면이 나왔다면 앞면이 나올 확률은 70%이다.

  - 단점 : 동전을 10번 던져서 7번 나오던, 100번 던져서 70번 나오던 똑같이 추정한다는 것.
  - 관측 결과를 D, 앞면이 나올 확률을 θ라고 할 때,

    $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$ 를 최대화 하는, θ를 찾는 것!

- **호에프딩(Hoeffding) 부등식** : 예측된 확률이 신빈성이 있는 지 판단하는 척도. (0에 가까울수록 신빈성 높음)
- **MAP** : 사전 정보를 반영해 가장 그럴듯한 확률을 계산하는 방법.

  ex) 동전을 10번 던져서 7번 앞면이 나온 결과가 관찰됨. 하지만 사전 지식으로는 동전이 앞면이 나올 확률은 50%임. 이 사전 지식을 고려해 앞면이 나올 확률을 다시 계산하면 70%보다 작은 확률로 예측하게 됨. (60% 언저리)

확률변수

- **표본공간** : 통계실험에서 발생가능한 모든 결과를 모은 공간.
- **결합분포** : 표본공간에서 발생 가능한 무수히 많은 확률 변수들 중 **여러 개의 확률 변수를 동시에 다루는 것**을 말함.
  - 확률변수의 표기 : $X, Y, Z...$ 또는 $X_1, X_2, X_3, ...$
- **확률 벡터** : 확률변수들의 순서쌍.
- **기댓값** : 각 사건이 벌어졌을 때의 **이득 값과 그 사건이 벌어질 확률을 곱한 것**을 **전체** 사건에 대해 합한 값.

  X가 확률분포 $f(x)$를 가지는 확률변수라 하자.

  X가 이산형인 경우 X의 평균 또는 기댓값은

  $$\mu = E[X] = \Sigma_x xf(x) = \Sigma_x xp(X = x)$$

  X가 연속인 경우 X의 평균 또는 기댓값은

  $$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

- **분산** : 확률 변수 X의 분포가 평균을 중심으로 밀집된 정도.

  X가 확률분포 $f(x)$를 가지는 확률변수라 하자. X의 기댓값 $\mu = E[X]$에 대하여

  X가 이산형인 경우 X의 분산은

  $$E[(x-\mu)^2] = \Sigma_x (x-\mu)^2 f(x)$$

  X가 연속인 경우 X의 분산은

  $$E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

  - 분산의 변형

    $$\sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2$$

- **표준편차** : 분산을 보기 편한 단위로 변환한 척도로, 분산의 제곱근으로 계산.

$\sqrt{E[(x-\mu)^2]}$

결합 확률 분포

- **결합 확률 분포** : 두 개의 분포를 하나로 묶어 생각하는 것.
    - 예시 : 키 + 몸무게, 동전 2개 던지기

- **결합 확률 질량 함수** : 이산형 데이터에서 결합 확률 분포가 가질 수 있는 **모든 조합의 확률.**
- **결합 확률 밀도 함수** : 연속형 데이터에서 결합 확률 분포가 가질 수 있는 **모든 조합의 확률.**

- **주변 확률 질량 함수** : 이산형 데이터에서 합쳐진 **두 개의 확률 분포 중 하나만 고려하는 질량 함수.**
- **주변 확률 밀도 함수** : 연속형 데이터에서 합쳐진 **두 개의 확률 분포 중 하나만 고려하는 질량 함수.**

나이브 베이즈 분류기와 마르코프 결정과정

- 두 사건이 독립인 경우의 표현
    - $p(x,y) = p(x)p(y)$
    - $P(A \cap B) = P(A)P(B)$
- 두 사건이 독립인 경우, 베이즈 정리는 의미가 없어짐!

  $p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x)$ → (y에 전혀 영향 받지 않음.)

- 확률변수의 조건부 확률

  $p(x|y) = \frac{p(x,y)}{p(y)}$
    - 체인룰
        - $P(A \cap B) = P(A)P(B)$
        - $p(x,y) = p(x|y)p(y)$
- 조건부 독립식 : 조건 z가 주어졌을 때, x와 y가 서로 독립임.

  $p(x,y|z) = p(x|z)p(y|z)$

→ x와 y는 z에 대해 따로 볼 수 있음.

ex) x : 우산을 들고 나갔는가?

y : 길에 물이 고여 있는가?

z : 비가 왔는가?

→ 비가 왔을 때 우산을 들고 나갈 확률 / 비가 왔을 때 길에 물이 고일 확률 로 분리해서 볼 수 있음.

$$p(x|y,z) = p(x|z)$$

→ x가 y,z에 의해 결정된다면, y는 무시 가능하다.

ex) x : 시험 점수

y : 공부 시간

z : 시험 난이도

→ 시험 점수(x)와 공부 시간(y) 사이의 상관관계는 분명히 존재하지만 시험 난이도(z)가 너무 쉽다면 그 상관관계가 무시될 수도 있다.

나이브 베이즈 분류기

- 각 특징이 독립이라고 가정하고, 확률 계산으로 가장 가능성 높은 클래스를 고르는 분류기

  $$p(y|x_1, x_2, ..., x_n) \propto p(x_1, x_2, ..., x_n|y)p(y)$$

  따라서,

  $$p(y|x_1, x_2, ..., x_n) \propto p(y)p(x_1|y)p(x_2|y)...p(x_n|y)$$

마르코프 결정과정

- 연속된 사건이 있을 때 현재 시점으로부터 바로 직전 시점의 사건까지만 고려하자.
  - 1차 마르코프 체인 : $p(w_n|w_{n-1}, w_{n-2}, ..., w_1) = p(w_n|w_{n-1})$
  - 2차 마르코프 체인 : $p(w_n|w_{n-1}, w_{n-2}, ..., w_1) = p(w_n|w_{n-1})p(w_{n-1}|w_{n-2})$
- 마르코프 체인을 이용하면 결합확률 분포를 계산할 수 있다.
  - 원래는 결합사건 w_i 사이의 의존 관계를 모두 계산해야 하지만, 마르코프 체인을 사용해 결합확률 분포를 아래와 같이 간단하게 계산 할 수 있다.

$$p(w_1, w_2, ..., w_n) = p(w_1|w_2, ...w_n)p(w_2, ..., w_n) = ... = p(w_1|w_2)p(w_2|w_3)...p(w_{n-1}|w_n)$$

$$= \Pi_{i=1}^{n} p(w_i|w_{i-1})$$