기초 이론부터 실무 실습까지

머신 러닝 익히기

programmers

Part 02. 머신러닝 기초와 배경

정 정 민

programmers

Chapter 06. 확률 이론

1. 기본 정의
2. 확률 분포
3. 확률론적 모델링과 추론

기본 정의

확률이란

- 특정한 사건이 일어날 가능성을 수치로 표현
- 0~1 사이의 값을 갖음

- 일반적으로 확률(Probability)의 P를 활용해 확률을 표시
- 또한, 어떠한 사건인지 사건을 알려주는 확률 변수(probability variable) x를 활용

- $P(x)$: 확률 변수 x가 특정 값을 가질 확률
- 만약 $P(x = 3)$ 이라 하면 확률 변수가 특정한 값인 3을 가질 확률을 의미

- 기본적인 확률 계산
  - 합의 법칙 : 두 사건 A와 B가 서로 베타적이라면, A 또는 B 확률 : $P(A) + P(B)$
  - 곱의 법칙 : 두 사건 A와 B가 서로 독립일 때, A와 B가 동시에 발생할 확률 : $P(A) \times P(B)$
  - 조건부 확률 : 사건 B가 일어난 상태에서 사건 A가 일어날 확률 : $P(A|B)$

확률 분포

확률 분포

- 확률 변수가 취할 수 있는 값들과
- 그 값들이 발생하는 확률을 설명하는 개념

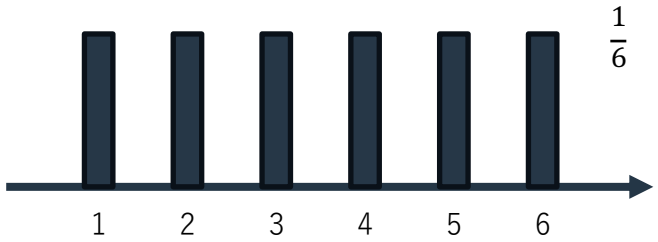
- 크게 아래의 분포들로 나눌 수 있음
  - 이산 확률 분포 (Discrete Probability Distribution)
  - 연속 확률 분포 (Continuous Probability Distribution)

이산 확률 분포 (Discrete Probability Distribution)

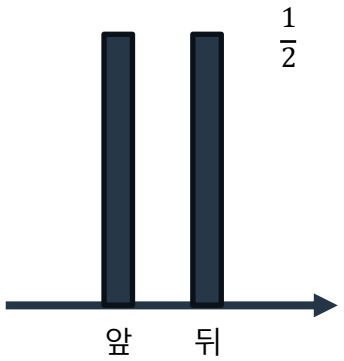
- 확률 변수가 취할 수 있는 값이 개별적이고 셀 수 있는 경우

- 확률 분포이므로
- 각 변수에 해당하는 확률 값의 총 합은 1

- 예로는,
  - 주사위를 던졌을 경우 확률 분포
    - 가능한 사건인 확률 변수가 1~6으로 개별적임
  - 동전 던지기
    - 역시 가능한 확률 변수가 앞 혹은 뒤로 개별적



주사위 던지기에 따른 확률 분포



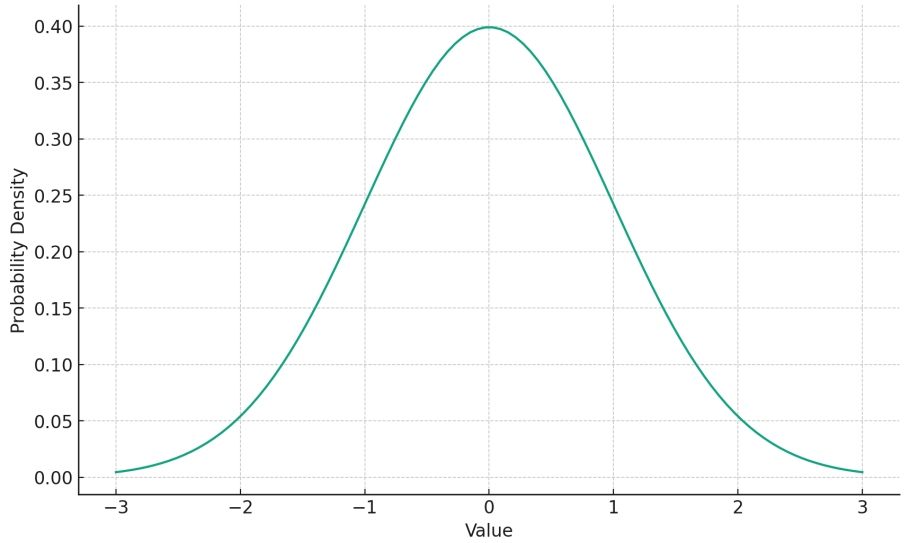
동전 던지기에 따른 확률 분포

연속 확률 분포 (Continuous Probability Distribution)

- 확률 변수가 연속적인 범위의 값(실수 범위의 값)을 취할 수 있을 때 적용

- 역시 확률 분포이므로
- 가능한 모든 확률 변수 전 구간의 적분 값은 1
- 변수의 범위가 실수이므로 딱 하나의 구체적인 값에 대한 확률은 0

- 예로는,
  - 전 국민의 키를 모은 데이터의 분포
    - 키는 소수 범위까지 포함해 실수 범위를 갖고 있음
  - 정규 분포
    - 자연 및 과학에서 많이 나타나는 대표적인 연속 분포
    - 평균이 0, 표준편차가 1인 정규 분포를 표준 정규 분포라고



분포와 확률 변수

- 확률 변수는 실험, 관찰, 또는 무작위 과정의 결과로 나타날 수 있는 수치적인 값

- 이러한 확률 변수는 확률 분포에 영향을 받음
- 만약 확률 분포를 알고 있다면 확률 변수를 임의로 생성할 수 있음
  - 이를 샘플링(Sampling) 과정이라 함

- 특정 분포 D를 따르는 확률 변수 X를 n개 샘플링 하면 아래와 같이 표현 가능
  - $X_1, X_2, \ldots, X_n \sim D$

- 예를 들어,
  - 동전 던지기 분포에서 하나의 데이터를 샘플링 하면 앞면이 나왔고
  - 정규 분포에서 하나의 데이터를 샘플링 해서 나온 값은 0.02421 나왔다

확률론적 모델링과 추론

programmers

확률론적 모델링

- 주어진 데이터를 확률 이론의 관점에서 해석하고 모델을 설계하는 과정을 의미함
  - 수학적인 모델을 통해 데이터를 분석 및 활용하는 과정

- 데이터가 특정 확률 분포를 따른다고 가정
  - 데이터에 존재하는 불확실성(noise)을 인정하면서!
- 이 분포는 데이터의 특성을 분석하거나 미래의 사건에 대한 예측에 활용됨

- 앞으로 수업에서 다룰 머신 러닝 모델은 모두 확률론적 모델링에 해당함

모델의 예측과 데이터

- 머신 러닝 모델의 출력은 확률론적 관점에서 예측된 결과물

- 따라서 실제 결과물과 차이가 있을 수 있음
- 일반적으로
  - 모델의 예측은 $\hat{y}$
  - 실제 정답은 y로 표시

- 사용하는 입력 데이터는 X로 표시
  - 입력 데이터는 일반적으로 오른쪽과 같은 형태로 생김

- X와 y를 포함해 일반적으로 전체 학습 데이터라고 함

개별 데이터 →

독립 변수

종속변수

출석 번호	혈액형	MBTI	키	중간 성적	기말 성적	정답
3	O	ENTP	163	92	96	1
4	A	INTP	168	89	88	0
…	…	…	…	…	…	…
10	B	ESTJ	155	88	82	1
11	AB	ENFP	170	90	91	1

X y

E.O.D