확률론적 관점에서 바라본 머신러닝

- 참고
  - 김기현의 자연어처리 딥러닝캠프 교재
  - 남기현님 블로그
  - PRML 정리 사이트

배경

머신러닝의 목표는 미지의 데이터에 대해 좋은 예측을 하는 것이다. 즉, 데이터를 잘 설명할 수 있거나, 주어진 데이터로부터 결괏값을 잘 예측하는 것들이 이에 속한다. 다시 말하자면 입력값을 넣었을 때 해당 데이터를 가장 잘 설명해주는 결괏값을 내놓는 함수라고 할 수 있다. 이 함수를 **확률론적인 관점**으로 해석하자면, 실제 확률 분포로부터 나온 데이터를 수집하고, 수집된 데이터를 가장 잘 설명하는 확률 분포 모델을 추정함으로써, 알고자 하는 실제 확률 분포를 근사하는 것이다.

필수 개념

1. 확률 변수

확률 변수는 확률을 이야기할 때 랜덤하게 발생하는 어떤 사건을 정의한다. 좀 더 엄밀하게 표현하자면 발생가능한 모든 사건들의 집합인 표본공간 안에서 특정한 확률을 가지고 발생하는 사건을 특정 수치에 대응(mapping) 시키는 **함수**이다. 그래서 확률 변수를 대문자 X로 표기하고, 확률 변수에서 나온 값은 소문자 x로 표기한다. 예를 들어 주사위를 던졌을 때(사건 X) 주사위의 숫자가 3이 나왔다면, 다음과 같이 표현할 수 있다.

$$P(X = 3) = 1/6$$

즉 이 수식에서, P는 확률을 의미하며, 확률변수 X가 특정 값을 가질 때, 확률값을 반환하는 함수라고 볼 수 있다. 여기서 확률변수는 주사위가 나오는 사건(X)을 특정 실수($x = \{1, 2, 3, 4, 5, 6\}$)로 매핑하는 함수이다.

2. 확률 변수와 확률 분포

이산 확률 변수와 확률 분포

보통 우리가 접하는 확률 변수는 불연속적인 이산값인 경우가 많다. 위에서 설명한 주사위가 이산 확률 변수의 한 예이다.

불연속적인 이산 확률 변수에 대한 확률 함수를 확률 질량 함수(pmf)라 하고 어떤 값에 대한 확률 값을 바로 알 수 있다. 이산 확률 분포로 대표적으로는 베르누이 분포와 멀티눌리 분포가 있다. 이 확률 분포를 일반화(시행 횟수 n 을 늘리면)하면 각각 이항 분포와 다항 분포가 된다.

연속 확률 변수와 확률 분포

불연속적인 이산 값을 다루는 분포가 있다면 연속적인 값을 다루는 분포 역시 생각해볼 수 있다. 그 분포를 연속 확률 분포라고 부른다.

예를 들어 1~6 사이의 모든 실숫값이 정육면체의 주사위가 아니라 '구' 위에 배치되어 있다고 가정하자. 그렇다면 '구' 에서 특정 위치(점)가 가장 위로 올라올 확률은 0에 가깝다. 따라서 우리는 점보다 어떤 영역(구간)에 대해 확률값을 구하는 편이 더 좋다.

이산 확률 변수와 마찬가지로 연속 확률 변수에 대한 확률 함수는 확률 밀도 함수(pdf)라고 한다. 확률 밀도 함수는 다음과 같이 정의할 수 있다.

$$\forall x \in X, p(x) \geq 0$$
$$It\ is\ not\ necessary\ that\ p(x) \leq 1$$
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

확률 질량 함수와 달리 확률 밀도 함수는 어떤 값에 대한 확률 밀도 함수가 꼭 1보다 작을 필요는 없으며, $p(x)$를 적분한 값은 항상 1이다.

3. 결합 확률과 조건부 확률

결합 확률

결합 확률 이란 두 개 이상의 사건이 동시에 일어날 확률을 말한다. 따라서 두 개 이상의 확률 변수를 가진다. 만약 각각의 사건이 독립이라면 다음의 조건을 만족한다.

$$P(A, B) = P(A)P(B)$$

조건부 확률

조건부 확률은 머신러닝과 딥러닝에서 아주 중요하다. 실제로 다루는 대부분의 문제가 이에 기반을 두기 때문이다. 조건부 확률은 두 사건에 대한 확률 분포이다. 다만 독립과는 달리, 하나의 확률 변수가 주어졌을 때 다른 확률 변수에 대한 확률 분포이다.

주사위를 두 개를 던졌을 때의 사건 A, B를 가정해보자.

$$P(A = 3|B = 2)$$

이 수식은 주사위 B가 2가 나온 상황이 주어졌을 때, 주사위 A의 값이 3이 나올 확률값을 말한다.

$$P(A|B = 2)$$

이 수식은 주사위 B가 2가 나온 상황이 주어졌을 때 주사위 A에서 얻을 수 있는 값의 분포를 말한다. 즉 A의 값이 주어지지 않았기 때문에 값을 반환하는 것이 아니라 확률 분포 함수를 반환하게 된다.

4. 베이즈 정리

확률의 법칙

- 합의 법칙

$$P(X) = \sum_{Y} P(X, Y)$$
$$P(Y) = \sum_{X} P(X, Y)$$

- 곱의 법칙

$$P(X, Y) = P(Y|X)P(X)$$

- 베이즈 정리

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$
$$P(X) = \sum_{Y} P(X|Y)P(Y)$$

5. 주변 확률 분포

두 개 이상의 확률 변수의 결합 확률 분포가 있을 때, 하나의 확률 변수에 대해서 적분을 수행한 결과를 말한다. 수식으로 나타내면 다음과 같다.

$$P(x) = \sum_{y \in Y} P(x, y) = \sum_{y \in Y} P(x|y)P(y)$$

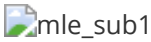
MLE(Maximum Likelihood Estimation)

하나의 관찰 데이터 집합 $X = (x_1, x_2, x_3, \ldots, x_n)^T$ 이 주어졌다고 가정해보자. 이 데이터 집합 하나가 관찰될 수 있는 확률은 과연 어떻게 될까? 각각의 데이터가 발현되는 가능성은 서로 독립적이므로(i.i.d) 이 확률값들은 모두 독립 사건으로 처리할 수 있다. 즉 확률의 곱으로써 표현이 가능해진다.

하나의 데이터는 동일한 분포로부터 발현되었을 것이므로 이 확률을 다음과 같이 표기할 수 있다.

$$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} N(x_n|\mu, \sigma^2)$$

이것을 그림으로 나타내면 다음과 같다.

 mle_sub1

그렇다면 어떤 관찰 데이터가 하나의 가우시안 분포(정규분포)를 따른다고 가정해보자. 우리가 얻는 것은 관찰 데이터 집합이고 이를 이용하여 원래의 정규분포를 결정하는 문제가 보통 우리에게 주어지는 문제이다. 쉽게 말해서 주어진 샘플이 어떻게 생겨먹은 정규분포에서 나왔는지를 맞추라는 얘기이다.

즉, 이제 $p(X|\mu, \sigma^2)$ 을 이용하여 이러한 관찰 결과를 만들어낼 만한다고 생각하는 가장 타당한 μ와 σ를 찾으면 되는 것이다. 다시 말하면 주어진 데이터를 이용하여 선택한 모델에 대한 파라미터를 결정하는 문제이다. 이를 **파라미터 추정**이라 부른다.

계산을 편하게 하기 위해 log를 붙여 식을 전개하면 다음과 같은 식이 전개된다.

$$J(\theta) = logp(X|\mu, \sigma^2) = -\frac{1}{2}log2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$$

이 함수의 값을 최대로 만드는 파라미터의 값을 편미분을 사용해 구할 수 있다.

$$\mu = \frac{1}{N}\sum x_n$$
$$\sigma^2 = \frac{1}{N}\sum (x_n - \mu)^2$$

이 가능도 함수를 가장 크게 만드는 모수 값을 추정하므로 최우추정법(Maximum Likelihood Estimation), MLE 라고 한다.