Probability Density Function

2016314726 정영준

확률 밀도 함수(Probability Density Function)는 확률 질량 함수(Probability Mass Function)가 이산형 분포의 확률함수라고 하면 연속형 분포에서의 확률함수라고 할 수 있습니다. 먼저 확률 밀도 함수에 대한 이해는 확률 질량 함수에서 이산형 분포가 연속형 분포로 바뀐 것이기 때문에 확률 질량 함수에 대한 설명을 간단하게 하겠습니다. 확률 질량 함수는 이산형 확률변수 X에 대해 확률 P(X=x)를 말하는데 보통 f(x)로 함수형으로 표기합니다. 이산형 확률 질량 함수는 셀 수 있는 경우에 대한 확률 함수로 동전의 앞면, 뒷면 출현 횟수, 주사위의 눈 수, 교통사고 발생건수, 특정 시의 인구 등이 이산형 확률 질량 함수로 표현 가능합니다. 확률 밀도 함수는 이와 다르게 셀 수 없는 경우, 즉 연속형 변수에 대해 사용하는 확률 함수로 키, 몸무게, 온도, 습도 등이 이에 해당됩니다. 여기서 저는 확률 밀도 함수가 왜 필요한지에 대해 의문이 들어서 이를 찾아보았습니다. 연속확률변수의 경우 확률변수가 취할 수 있는 값이 연속적이며 무한하기 때문에 분포를 표현하는 것이 불가능합니다. 만약 이산형처럼 특정한 확률변수 하나에 특정한 확률값에 대응된다면, 어떤 확률변수가 특정한 구간 안에 포함될 확률은 무한대가 되어 버립니다. 확률변수가 취할 수 있는 값이 무한개이기 때문입니다. 그러나 특정한 확률변수의 값에 대한 확률값을 0으로 설정할 수도 없습니다. 이렇게 되면 특정 구간 안에 확률변수가 포함될 확률은 어느 구간에서든 0이 되어버리기 때문입니다. 연속형 확률변수의 이러한 난점을 돌파하기 위해 확률 밀도 함수(probability density function, PDF)가 필요하게 되었다고 합니다. 이처럼 확률 밀도 함수는 확률 질량 함수와 달리 특정 구간에서의 확률의 값을 0으로 설정해버리고 확률 질량 함수를 적용시키면 모든 값에 대한 확률값이 0이 되어버리기 때문입니다. 그래서 이를 다른 관점으로 나타낸 확률 밀도 함수가 필요한 것입니다. 확률 밀도 함수의 정의에 대해 알아보았고 이제 확률 밀도 함수의 성질에 대해 알아보겠습니다. 먼저 확률 밀도 함수는 확률 질량 함수와 마찬가지로 가능한 확률 구간에 대한 확률값의 합이 1이 됩니다. 확률 밀도 함수는 구간을 기준으로 하기 때문에 적분으로 이를 나타내면 f(x) (a<=x<=b)의 경우 a에서 b까지 f(x)를 정적분한 값은 1이 됩니다. 확률 밀도 함수에서 왜 Density라는 단어가 들어갔는지 생각해보았습니다. P(x1 <= X <= x2)는 확률 밀도 함수 f(x)를 x1에서 x2를 구간으로 정적분한 값이 됩니다. 여기서 적분식 마지막에 쓰이는 dx는 구간의 길이를 나타냅니다. f(x)의 y축은 확률을 나타내기 때문에 이는 확률 / 구간길이를 의미합니다. 밀도는 질량 / 부피로 나타낼 수 있습니다. 확률 구간의 길이를 일종의 부피로 생각해 본다면 질량은 얼마나 무거운가, 확률 밀도 함수에서는 얼마나 확률이 큰가로 나타낸다고 생각하면 이는 곧 밀도를 구하는 것과 같습니다. 확률 밀도 함수에서 밀도의 의미는 이와 같이 해석이 가능합니다. 마지막으로 많이 사용되는 정규분포의 확률 밀도 함수 식을 알아보았습니다. f(x) = 1/(2\*pi)^(1/2) \* exp(-x^2/2)로 표현이 되었으며 이는 평균이 0, 분산이 1일때의 식입니다. 확률 밀도 함수의 활용보다 정확한 의미와 확률 밀도 함수가 필요한 이유에 대해 조사를 중점적으로 하였습니다.