**1. 엔트로피의 개념**

엔트로피의 개념은 에너지와 함께 현대 물리학 및 관련 분야의 초석을 구성한다.

열역학은 거시적 시스템의 동작과 관련된 측면을 설명하는 경험적 물리 이론이다. 어떤 형태로든, 모든 큰 물리적 시스템이 이 이론을 충족하는 것으로 나타났습니다. 이는 에너지와 엔트로피라는 두 가지 가장 관련성이 높은 개념에 바탕을 두고 있다. 독일의 물리학자이자 수학자인 Rudolf Julius Emanuel Clausius (1822–1888)는 1865년에 엔트로피의 개념을 다소 추상적인 선에 따라 도입했다. 가역 과정의 경우, 정확한 미분량 는 을 통한 미분 열전달 와 관련이 있으며, T는 절대 온도이다. 의 양은 열의 전달(물리적 변환의 특정 경로에 따라 달라짐)을 엔트로피의 정확한 양(경로 독립)으로 변환하는 적분 인자의 역할을 한다. 이 관계는 Clausius에 의해 유명한 부등식 으로 일반화되었으며, 등식은 가역과정에 해당한다. 부등식은 불가역 과정에 해당하며, 시간의 화살에 대한 인간의 인식과 깊은 관련이 있는 소위 열역학의 제2법칙에 의해 직접적으로 암시된다.

10년 후, 오스트리아의 물리학자이자 철학자인 Ludwig Eduard Boltzmann(1844-1906)은 열역학 엔트로피 S와 미시 세계를 연결하는 중요한 발견을 했다. 이 유명한 공식인 에서 W는 시스템에 대한 우리의 정보와 양립할 수 있는 동등한 미시적 가능성의 총 수이다. Boltzmann은 이 관계에 알고 있었지만, 자신의 논문 중 하나에서도 이를 쓴 적이 없는 것으로 보인다. 미국의 물리학자, 화학자, 수학자 Josiah Willard Gibbs(1839~1903)는 이 연결의 물리적 의미를 더 논의하고 확장했다. 그들의 노력은 현재 통계역학으로 알려진 강력한 이론의 형성에 절정에 달했다. 이 이름이 당시에는 매우 논란이 많은 문제였다. 실제로, 이 이름은 뉴턴 역학의 완전한 결정론적 이해의 초석인 역학(*mechanics*)이라는 단어와 비결정론적 개념에 기반을 둔 확률론적 설명의 초석인 통계(*statistics*)라는 단어를 병치시킨다. 온도의 개념 자체에 대한 Boltzmann의 해석은 유체 자체를 구성하는 분자의 자발적인 시공간 변동과 직접적으로 관련이 있었다.

Max Planck, Paul and Tatyana Ehrenfest, and Albert Einstein의 공헌을 포함한 많은 중요한 공헌들이 뒤따랐다. 이후 엔트로피에 관한 중요한 다음 단계인 양자역학 시스템으로의 확장을 한다. 1927년 헝가리계 미국인 수학자, 물리학자, 컴퓨터 과학자인 János Lajos Neumann(John von Neumann, 1903-1957)에 의해 도입되었다.

1948년 미국의 전기공학자이자 수학자인 Claude Elwood Shannon(1916–2001)이 그의 “통신의 수학적 이론(Mathematical Theory of Communication)”의 엔트로피 개념을 기반으로 했다. 이것은 오늘날 정보 이론으로 널리 언급되는 것의 시작이었고, 그 안에서 미국의 물리학자 Edwin Thompson Jaynes (1922–1998)는 최대 엔트로피 원리를 도입하여 통계역학과의 연결을 확립했다. 이후 몇 가지 일반화가 도입되었으며, 그 중 첫 번째 일반화가 1961년 헝가리 수학자 Alfréd Rényi(1921-1970)에 의해 이후 언급됩니다. 이후 수십 년 동안 정보 이론, 사이버네틱스, 기타 컴퓨터 기반 프레임의 영역이 그 뒤를 이었다. 모든 경우에 엔트로피는 접근 가능한 시스템 상태 수의 측정값으로 나타나거나, 시스템에 대한 무지 또는 불확실성의 척도로 나타난다.

1988년, 그리스-아르헨티나-브라질 물리학자 Constantino Tsallis는 지수가 가 실수인 로 표기된 비가산 엔트로피(nonadditive entropy)를 기반으로 통계역학 자체의 일반화를 제안하였다. 이 이론은 현재 비확장 통계역학(*nonextensive statistical mechanics*)이라고 불린다. 이후 엔트로피 함수의 폭발적 증가가 있었다. 오늘날 이용 가능한 문헌에는 그러한 엔트로피가 50개가 넘게 있다. 그러나 그들 중 물리학과 다른 곳에서 깔끔한 응용을 발견한 사람은 거의 없다.

이산 변수에 대한 The Boltzmann–Gibbs–von Neumann–Shannon 엔트로피 함수(또는 단순성을 위한 엔트로피)는 다음과 같이 정의된다.

여기서 는 (물리학에서는 볼츠만 상수, 정보 이론에서는 k = 1로 간주됨) 일반적인 양의 상수이며, 는 의 가능한 상태에 해당하는 확률이다.

**2. 계산 사례**

데이터 과학에서 엔트로피는 열이 얼마나 혼합되어 있는지를 측정하는 방법으로 사용된다. 특히, 엔트로피는 무질서를 측정하는 데 사용됩니다. 예를 들어 서울에 사는지는 의부를 나타내는 열의 엔트로피를 찾는다고 가정한다.

|  |
| --- |
| 거주지: 서울 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 0 |
| 1 |

서울에 사는 사람은 11명, 그렇지 않은 사람은 4명이다. 만약 누군가가 이 열이 얼마나 혼합되어 있는지 물어보려고 했다면, 서울에 사람들인 대다수와 혼합되어 있다고 말할 수 있을 것이다. 엔트로피는 우리에게 “혼합된” 답을 수량화하는 방법을 제공한다. 열의 (1)과 (0)이 더 많이 혼합될수록 엔트로피는 더 높아진다. 만약 (1)과 (0)이 같은 양을 가지면 엔트로피는 1이 될 것이고, 만약 (1)로만 구성된다면 엔트로피는 0일 될 것이다.

다음 공식을 사용하여 엔트로피를 계산할 수 있습니다.

공식의 각 단계를 살펴보고 “거주지: 서울”의 열의 엔트로피를 계산해 봅시다.

1. 각각의 고윳값을 하나의 열에 반복하여 에 할당해야 합니다. 이 예제에서는 “거주지: 서울” 열에 (0) 또는 (1)의 두 가지 사례()가 있습니다.
2. 그런 다음 데이터에서 해당 값이 발생할 확률을 계산합니다. (1)의 경우 확률은 11/15입니다. (0)의 경우 확률은 4/15입니다.
3. 각각의 경우의 확률을 가지고 그것을 확률의 로그 밑수 2에 곱합니다. 엔트로피는 비트 단위로 측정되기 때문에 2가 가장 일반적이다. (1)의 경우 를 얻는다. (0)의 경우 가 나옵니다.
4. 다음으로 각각을 합산하면, 입니다.
5. 마지막으로, 위에서부터의 총합, 즉 를 부정합니다.

모든 단계를 종합하면 다음과 같은 결과를 얻을 수 있습니다.

엔트로피는 0.836641입니다.

이렇듯 엔트로피는 무질서 또는 불확실성의 척도이며 기계 학습 모델과 데이터 사이언티스트의 목표는 일반적으로 불확실성을 줄이는 것입니다.

무질서를 측정하고 나서는 그것에 대한 추가 정보(특징/독립 변수)가 주어진 목표 변수/클래스에서 이 장애의 감소를 측정하기 위한 메트릭이 필요합니다. 이것을 정보 이득이라고 부릅니다. 수학적으로 다음과 같이 쓸 수 있습니다.

Y에 대한 추가 정보 X가 주어진 Y의 엔트로피에서 Y의 엔트로피를 빼기 위해 Y에 대한 불확실성의 감소를 계산한다. 이것은 정보 이득이라고 불립니다. 이 불확실성의 감소가 클수록, X로부터 Y에 대한 더 많은 정보를 얻는다.

간단한 분할표를 사용한 예를 들어보면,

|  |  |  |  |
| --- | --- | --- | --- |
| **신용등급** | **부채** | | |
| 정상 | 높음 | **합계** |
| **우수** | 3 | 1 | 4 |
| **좋음** | 5 | 3 | 8 |
| **불량** | 0 | 4 | 4 |
| **합계** | 8 | 8 | 16 |

여기서 목표 변수는 “정상과” “높음” 두 가지 값을 가질 수 있는 “부채”이며 “우수”, “좋음” 및 “불량” 값을 가질 수 있는 신용 등급이라는 한 가지 특징만 있습니다. 관측치는 총 14개이며, 이 중 7개는 정상 부채 클래스, 7개는 높은 부채 클래스에 속합니다. 그래서 그 자체로 짝수 분할입니다. 맨 위 행에 걸쳐 요약하면 기능 신용 등급에 대해 “우수” 값을 가진 4개의 관측치가 있음을 알 수 있습니다. 또한 “우수” 신용 등급에 대한 목표 변수가 어떻게 분할되어 있는지도 알 수 있습니다. 신용 등급에 대해 “우수” 값을 갖는 관측치의 경우 “정상” 부채 클래스에 속하는 3개 및 “높은” 부채 클래스에 속하는 1개만 있습니다.

분할표를 사용하여 대상 변수의 엔트로피를 자체적으로 계산한 다음 특징인 신용 등급에 대한 추가 정보를 바탕으로 대상 변수의 엔트로피를 계산합니다. 이를 통해 “신용 등급”이 목표 변수 “부채”에 대해 얼마나 많은 추가 정보를 제공하는지 계산할 수 있습니다.

대상 변수의 엔트로피는 클래스 레이블 “정상”과 “높음” 사이의 짝수 분할로 인한 최대 무질서 상태에서 1이다. 다음 단계는 신용 점수에 대한 추가 정보가 주어진 목표 변수의 부채 엔트로피를 계산하는 것이다. 이를 위해 신용 점수의 각 값에 대한 책임에 대한 엔트로피를 계산하고 각 값으로 끝나는 관측치 비율의 가중 평균을 사용하여 추가한다.

특징인 신용 등급에 따라 목표 변수에 대한 엔트로피를 얻었습니다. 이제 신용 등급에서 부채에 대한 정보 이득을 계산하여 이 기능이 얼마나 유용한지 확인할 수 있습니다.

신용 등급을 아는 것은 목표 변수인 부채에 대한 불확실성을 줄이는 데 도움이 되었습니다. 의사결정 트리가 엔트로피와 정보 이득을 사용하여 노드를 분할할 특징을 결정함으로써 각 분할을 통해 대상 변수를 예측하고 트리 분할을 중지할 시점을 결정하는 방법과 이유를 알아보았습니다.