

그림 데이터와 엔트로피

1. 정보이론에서 엔트로피가 갖는 의미

물리학에서 엔트로피는 어떠한 물리계의 무질서한 정도를 의미한다. 정보이론에서도 엔트로피는 무질서한 정도를 뜻하며, 더욱 엄밀히 말하자면 “불확실성”을 의미한다. 위의 왼쪽 사진은 모든 영역이 검은색이므로 불확실성이 낮다. 즉, 데이터 중에 어떤 부분을 선택하더라도 검은색이 추출된다.

오른쪽 사진의 데이터에서는 흰색과 검은색이 무작위하게 분포되어 있기 때문에 불확실성이 높다.

* 정보이론에서 엔트로피가 높다는 것은 정보의 양이 많다는 것을 의미한다.

1. 엔트로피에 대한 확률적 정의

정보이론에서는 엔트로피를 확률적으로 정의한다. 먼저, 엔트로피를 정의하기 위해 전체 데이터에서 어떠한 기호가 등장할 확률 (Symbole Probability)를 [식 1]과 같이 정의한다.

(1)

은 데이터의 길이이며, 는 데이터에서 어떠한 기호 가 등장하는 횟수이다.

어떠한 기호의 정보량은 아래의 [식 2]와 같이 정의된다. 로그 함수의 정의에 따라 정보량은 가 작아질수록 커지며, 가 1에 가까울수록 0으로 수렴한다. 어떠한 기호가 나타날 확률이 아주 낮다면, 이 기호가 나타난다는 것을 알기 위해서는 많은 정보가 필요할 것이다. 그러나 어떠한 기호가 나타날 확률이 1에 수렴한다면, 적은 양의 정보만으로도 기호가 나타날 것을 쉽게 유추해 낼 수 있을 것이다.

(2)

마지막으로, 엔트로피는 아래의 [식 3] 과 같이 정의된다.

(3)

위 [식 3]의 정의와 같이 엔트로피는 정보량의 기댓값(쉽게 말해서 평균값) 이라고 생각할 수 있다. 즉, 정보이론에서 엔트로피는 어떠한 데이터를 표현하기 위한 평균 정보량을 의미한다.

[식 3]에서 알 수 있듯이 엔트로피는 각각의 기호가 등장할 확률이 비슷해질수록 증가하며, 각각의 기호가 등장할 확률이 모두 동일할 때 가장 크다.

엔트로피란 무질서의 척도로써, 완전한 질서 상태일시 엔트로피 S = 0

엔트로피가 낮다 -> 확정적인 정보가 많음. 특정 심볼이 발생 확률이 높음. 예측성이 있음.

엔트로피가 높다 -> 예측이 어려움. 각 심볼들의 발생확률이 동일함/무작위성, 랜덤성이 높음, 중복성이 거의 없음. 평균 정보량이 높음.

엔트로피의 상하한값은 [식 4]와 같다.

(4)

1. 하한값
   1. 모든 심볼 중 오직 하나만이 발생확률이 1이고, 나머지 심볼들의 발생확률이 0일 때.
   2. 불확실성이 없음
2. 상한값
   1. 소스 내의 모든 심볼들이 동일한 발생확률을 가질 때 ( )
   2. 불확실성이 가장 높음을 의미

엔트로피를 DDOS/DrDOS 공격 탐지에 쓰는 방법

1. 카메라를 통해서 DDOS/DRDOS 공격을 가했을 시 -> DDOS 공격을 효율적으로 하기 위해 타겟을 한정시켜서 패킷을 보내므로, 카메라에서 나가는 패킷의 도착지 주소의 다양성이 매우 적어질 것이다 -> 도착지 주소의 엔트로피가 작을 것이다!
2. 카메라에 DDOS 공격이 가해졌을 경우 -> 다수의 좀비 혹은 IP 변조를 통해 공격할 것이므로 카메라에 패킷을 보내는 출발지 주소의 다양성이 매우 많아질 것이다 -> 출발지 주소의 엔트로피가 클 것이다!