시각지능A – 이미지 분류 심화

손실 함수

목차 손실 함수

1. 손실 함수

손실 함수

- 손실함수(Loss function) 정의
  - 머신러닝 모델의 성능을 평가하고, 예측된 값과 실제 값 사이의 차이를 측정하기 위한 함수
  - Loss funcion, Cost function, Objective function 등으로 불림
  - Loss(y_pred, y_target)
- 해결하려는 문제의 특성에 맞게 선택
  - 지도학습 / 비지도학습
  - 회귀 / 분류
  - 텍스트 생성 / 이미지 생성

손실 함수

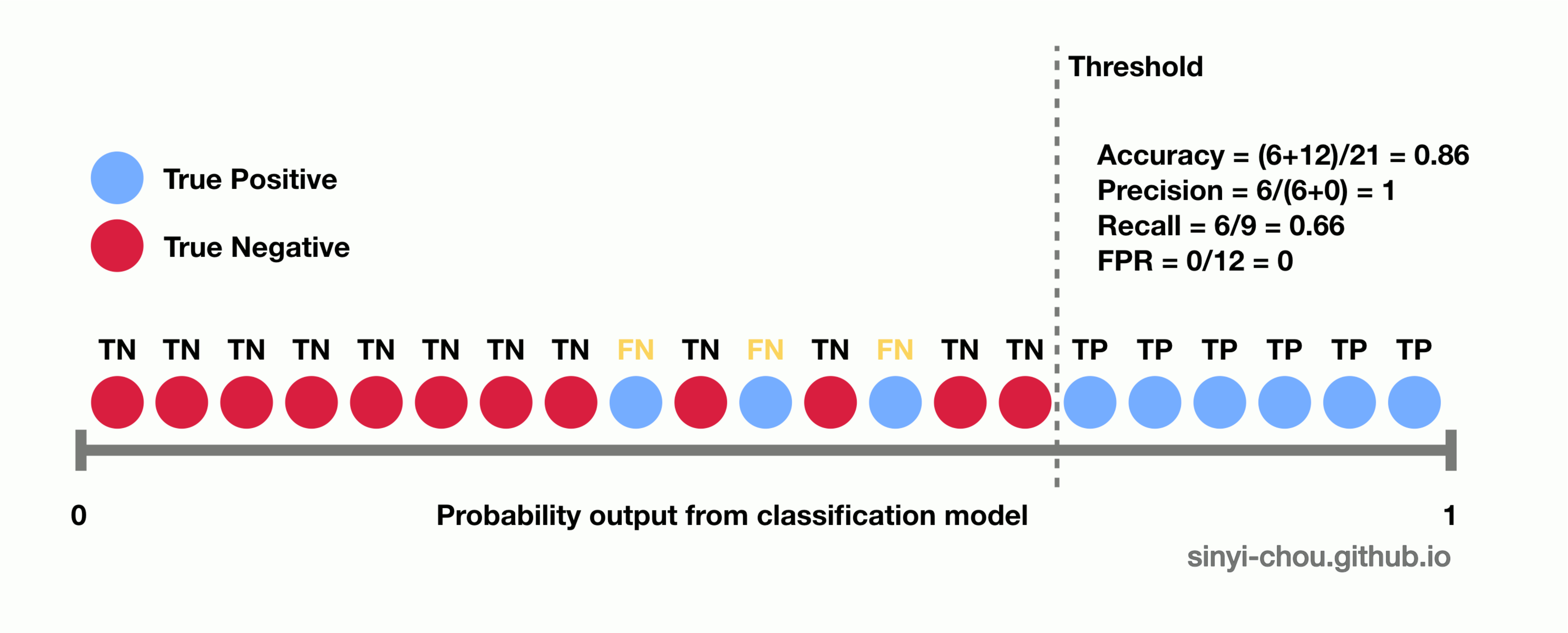
- 손실 함수는 모델이 가야 할 방향을 제시
  - 모델의 파라미터를 조정하면서 손실함수의 결과를 최소화하는 방향으로 학습이 진행
  - 반대로 모델은 손실 값이 작아지는 방향으로만 개선됨
  - 같은 모델이라 하더라도 손실함수에 따라 모델이 최적화되는 방식도 달라짐
- 모델의 틀린 정도를 정량화
  - 경우에 따라서는, 모델이 맞는 정도를 표현할 수 있음

손실함수와 평가 지표

- 손실 함수
  - 모델의 학습 과정에서 최적화(Optimization) 성능을 개선하기 위한 정량적 지표
- 평가 지표(Metrics)
  - 학습된 모델의 일반화 성능을 대변
  - 모델의 최종 성능을 평가하며, 이를 설명
- 경우에 따라서는 두 지표가 일치할 수 있음, 그러나

손실 함수가 평가지표가 될 수 없는 이유

- 분류 모델 평가 지표의 경우 Threshold에 의존하며, 불연속적임
  - 결과값이 연속형 변수가 아닌 이산형 변수(TP, FP, TN, FN)
  - Threshold 값은 얼마든지 변할 수 있음
  - Threshold가 변한다면, TP, FP, TN, FN의 갯수는 바뀜
  - 반면 분류 모델 평가지표에서 손실 값은 같은 Epoch 내에서 바뀌지 않음



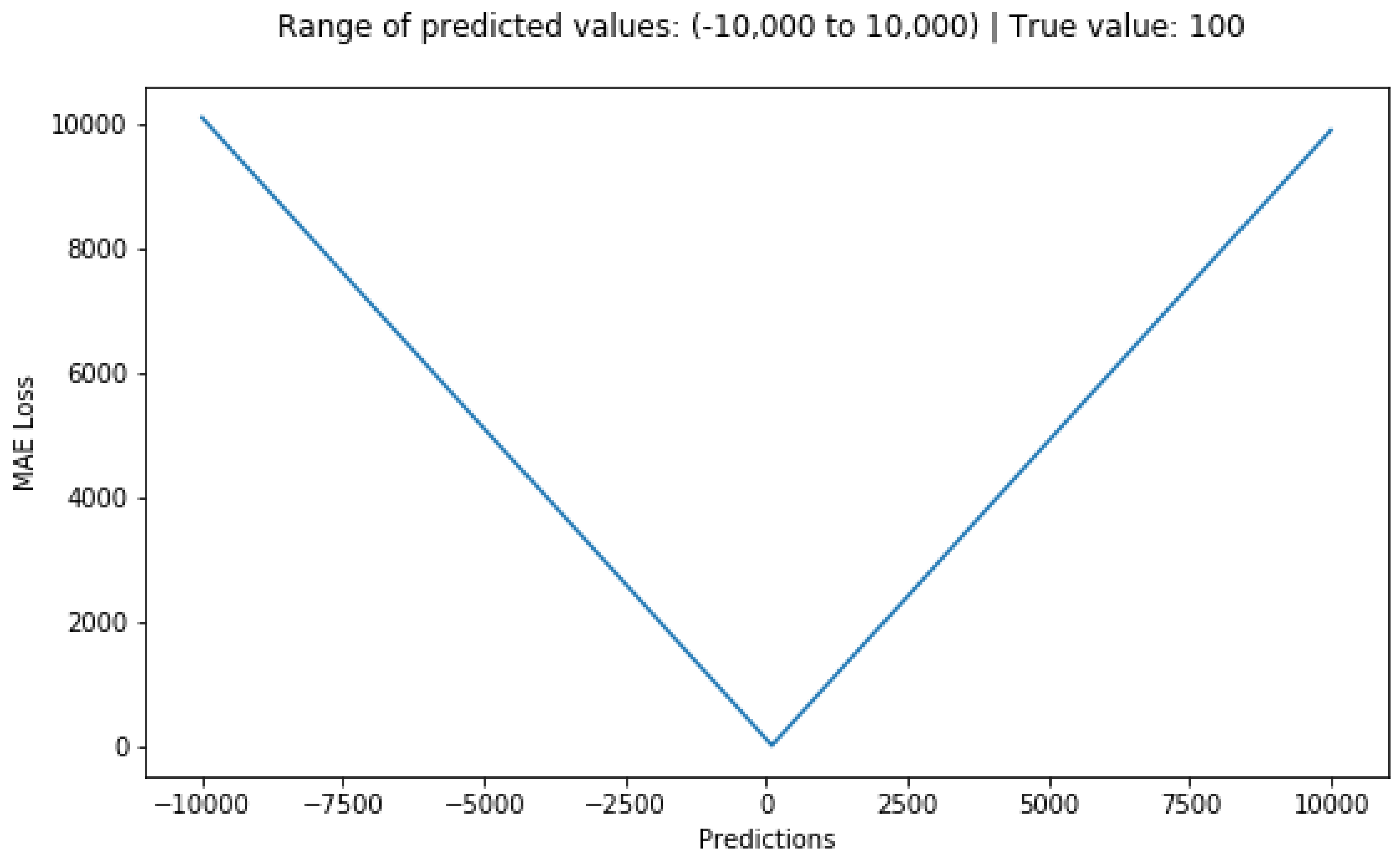
https://sinyi-chou.github.io/classification-auc/

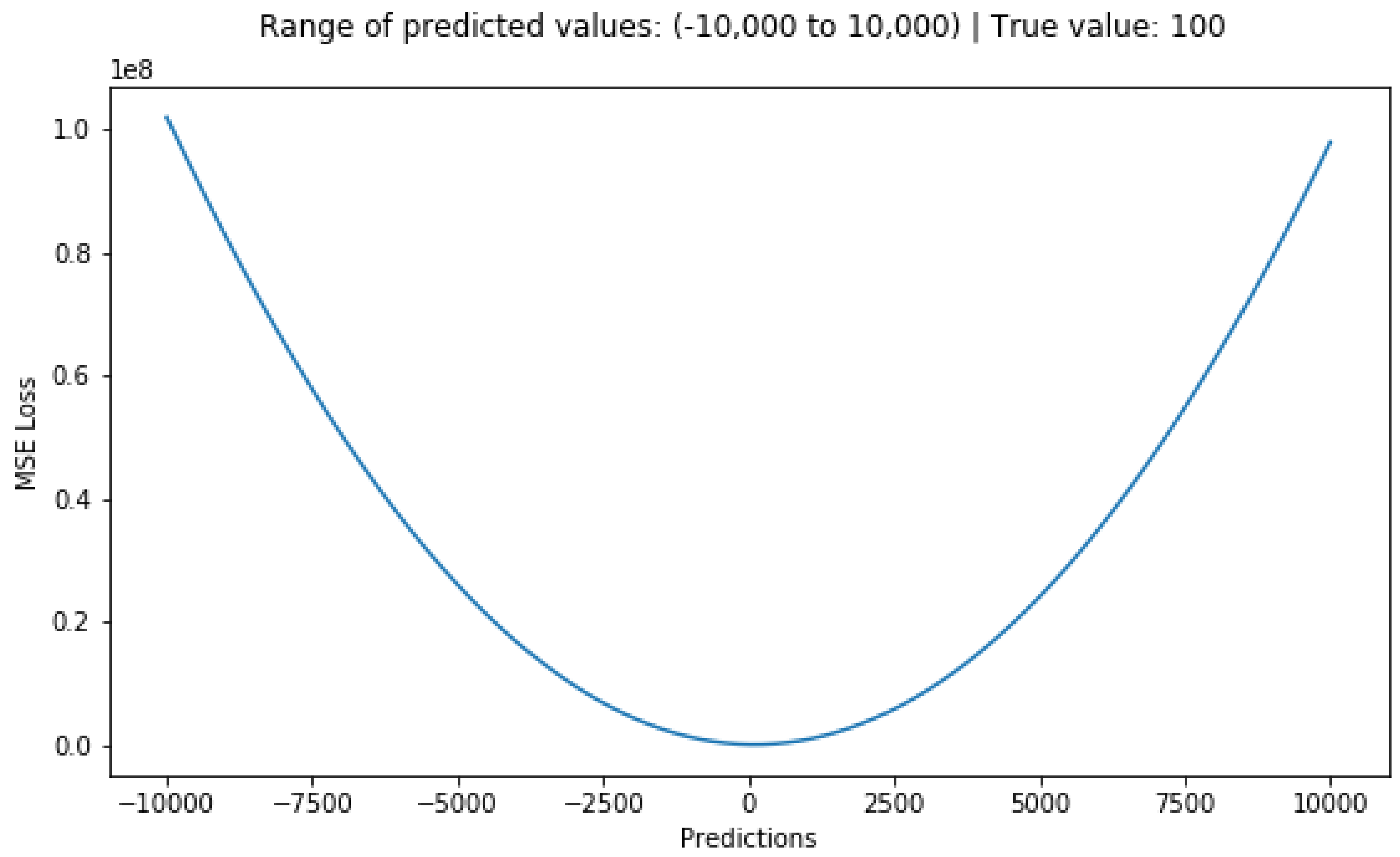
손실 함수가 평가지표가 될 수 없는 이유

- 손실 함수는 미분이 가능해야 함
  - 손실 값의 기울기(그래디언트)를 통해 모델의 파라미터를 업데이트
  - 다수의 평가지표는 미분이 불가능
  - 정확도(if x> 0.5, 1 else 0)
  - e.g. MAE v.s. MSE

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2$$





https://bo-10000.tistory.com/44

손실 함수가 평가지표가 될 수 없는 이유

- 평가지표는 손실함수로써 필요한 정보를 제공하지 않을 수 있음
  - 정확도: 모델이 얼마나 많은 샘플을 정확히 분류(TP, TN)했는지에 대한 정보를 제공
    - 잘못된 예측이 얼마나 틀렸는지 모름
    - 이 틀림이 어떤 방향으로 수정되어야 하는지에 대한 세부 정보를 제공하지 않음
  - 이를 개선하기 위한 평가지표들 또한 오답의 방향성을 전부 제시하지는 못함
    - Precision
    - Recall

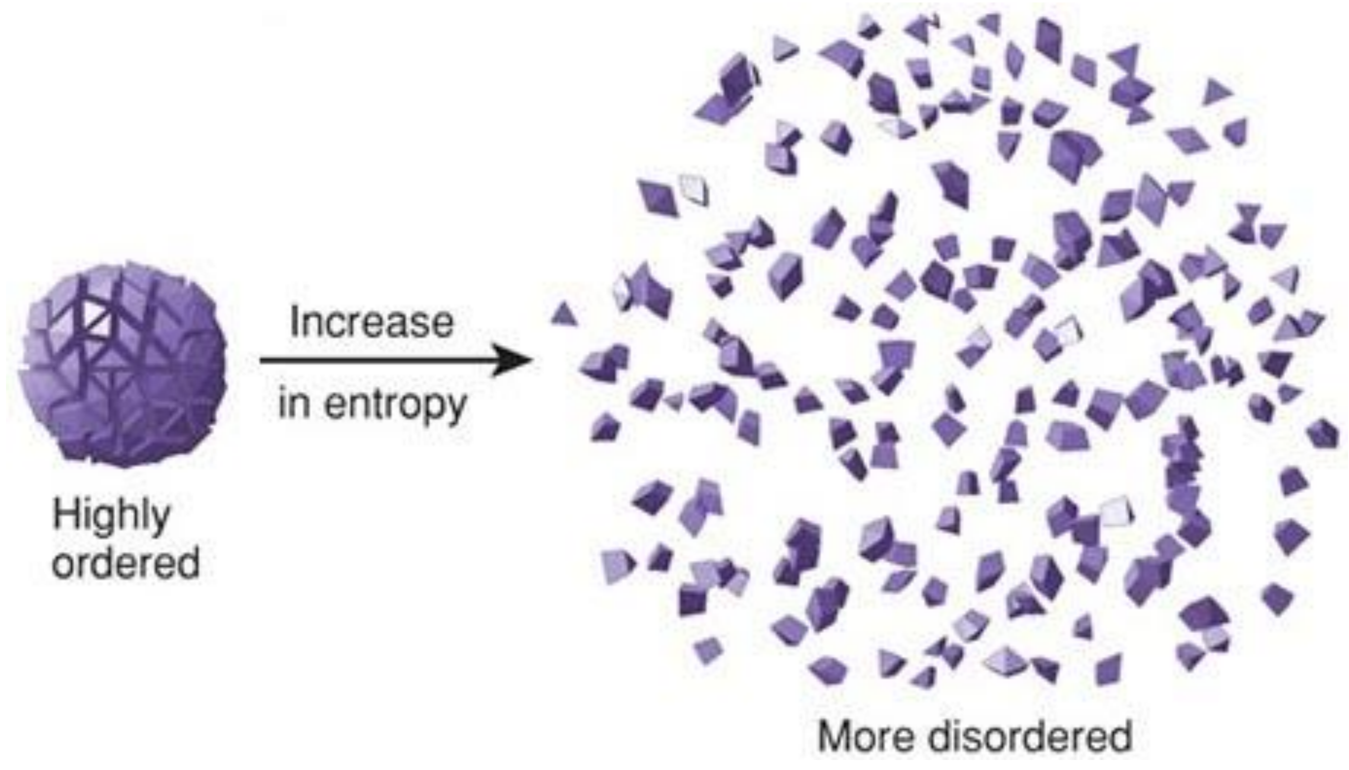
2. 정보 이론과 Entropy

정보 이론

- 정보 이론(Information theory)
  - 정보의 생성, 전달, 저장, 처리 및 해석과 관련된 수학적 원리와 개념을 연구하는 학문 분야
  - 20세기 초, 라디오, 전화, 전신 등 새로운 통신 기술이 급격히 발전하며 부각됨
  - 통신 시스템에서의 효율성과 신뢰성을 향상시키기 위한 이론적 기반
  - 제한된 대역폭에서 최대한 많은 정보를 전송하기 위한 방법
  - Entropy 또한 여기에서 유래된 개념

엔트로피

- 엔트로피(Entropy)
  - 물리학 중 열역학에서 출발한 개념
  - 열역학적 정의
    - 시스템의 무질서도 또는 에너지의 분산 정도를 나타내는 지표
    - 질서가 없고, 에너지가 고르게 분산될 수록 엔트로피가 높음
  - 통계역학적 정의
    - 특정 상태를 구성할 수 있는 마이크로 상태(microstate)의 수
    - 시스템의 불확실성과도 관련



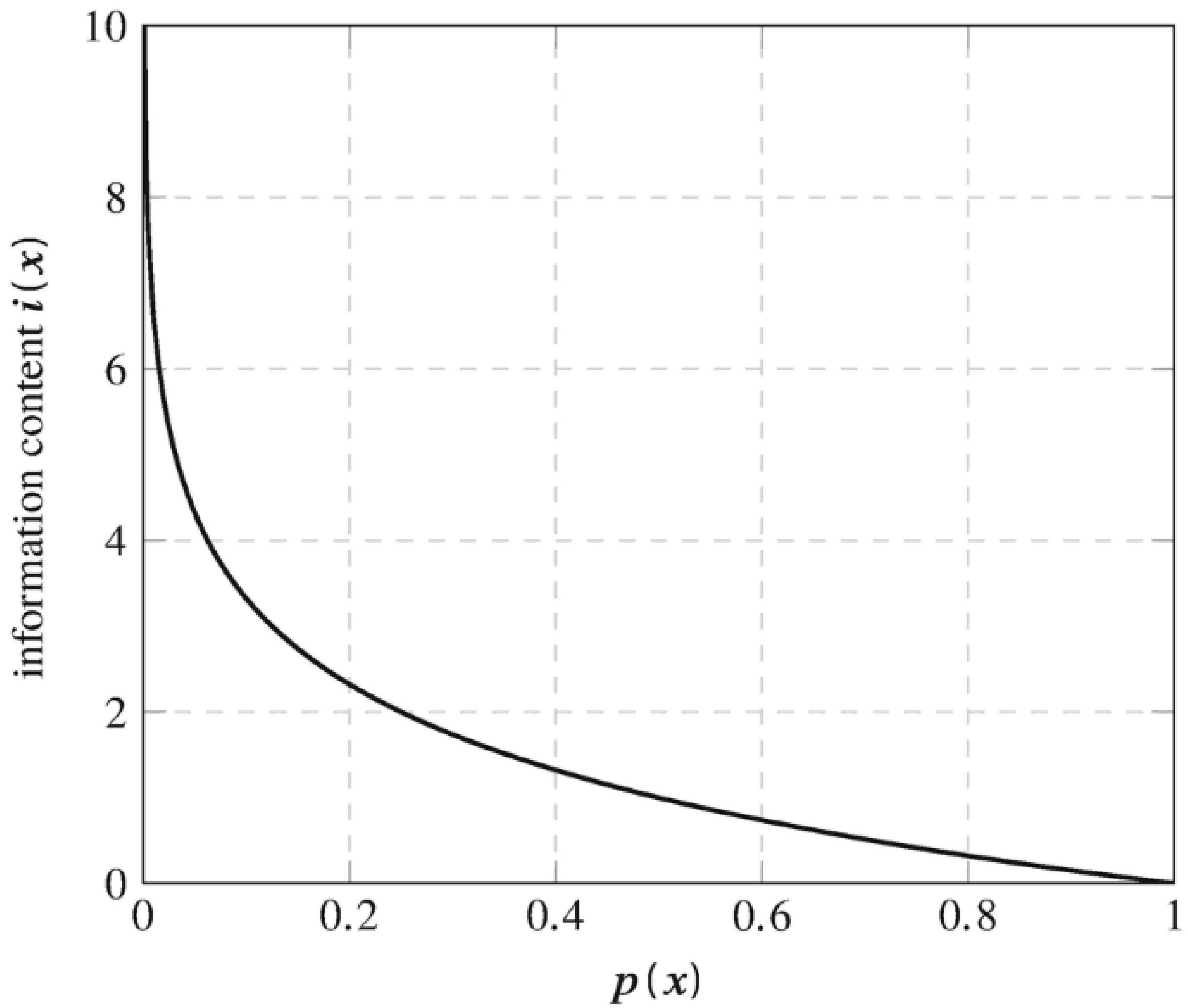
엔트로피

- 정보 이론에서의 엔트로피
  - 주어진 메시지나 데이터 집합의 불확실성 정도를 측정한 지표
  - 어떤 사건이 발생할 확률의 분포에 따라 그 사건이 가진 정보의 양을 측정
  - 예측 가능한 사건: 정보량 적음
    - 해는 동쪽에서 뜬다
  - 예측하기 어려운 사건: 정보량 많음
    - 흰 색 까마귀를 발견하는 경우

https://carto.com/blog/short-term-rental-data-airbnb-area-51

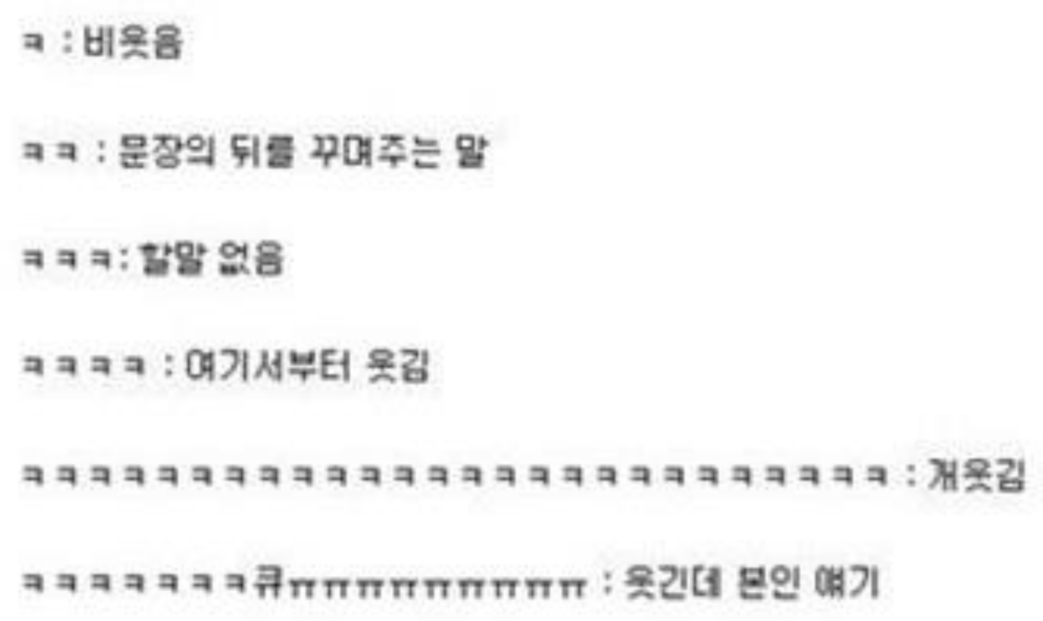
정보 이론과 Entropy

- 정보의 발생 확률(P)과 양(Quantity)의 관계
  - 정보의 양(IC, Information content)와 발생 확률은 반비례 관계
  - 발생 확률이 높은 정보는 영양가가 없음
    - 즉, 큰 리소스 투자 없이 얻을 수 있는 개연적인 정보
    - 정보의 양이 적음
  - 반면, 발생 확률이 낮은 정보는 비싸며, 정보량이 많음
    - 예측하기 어려운 사건일 수록 얻을 수 있는 내용이 많음



정보 엔트로피

- 통신 상에서, 사용할 수 있는 전파 양이 제한되어 있다고 가정
- 가장 효율적으로 전파를 분배하여 사용하는 방법은
  - 중요하고 드물게 발생하는 긴 전파로
  - 중요하지 않고 자주 등장하는 정보를 작고 짧은 전파로 나누는 것이 가장 효율적
- 이러한 정보의 화폐단위를 Entropy라 함
  - 사건이 발생하였을 때, 얻을 수 있는 예상된 정보량


ㅋ : 비웃음

ㅋㅋ : 문장의 뒤를 꾸며주는 말

ㅋㅋㅋ: 할말 없음

ㅋㅋㅋㅋ : 여기서부터 웃김

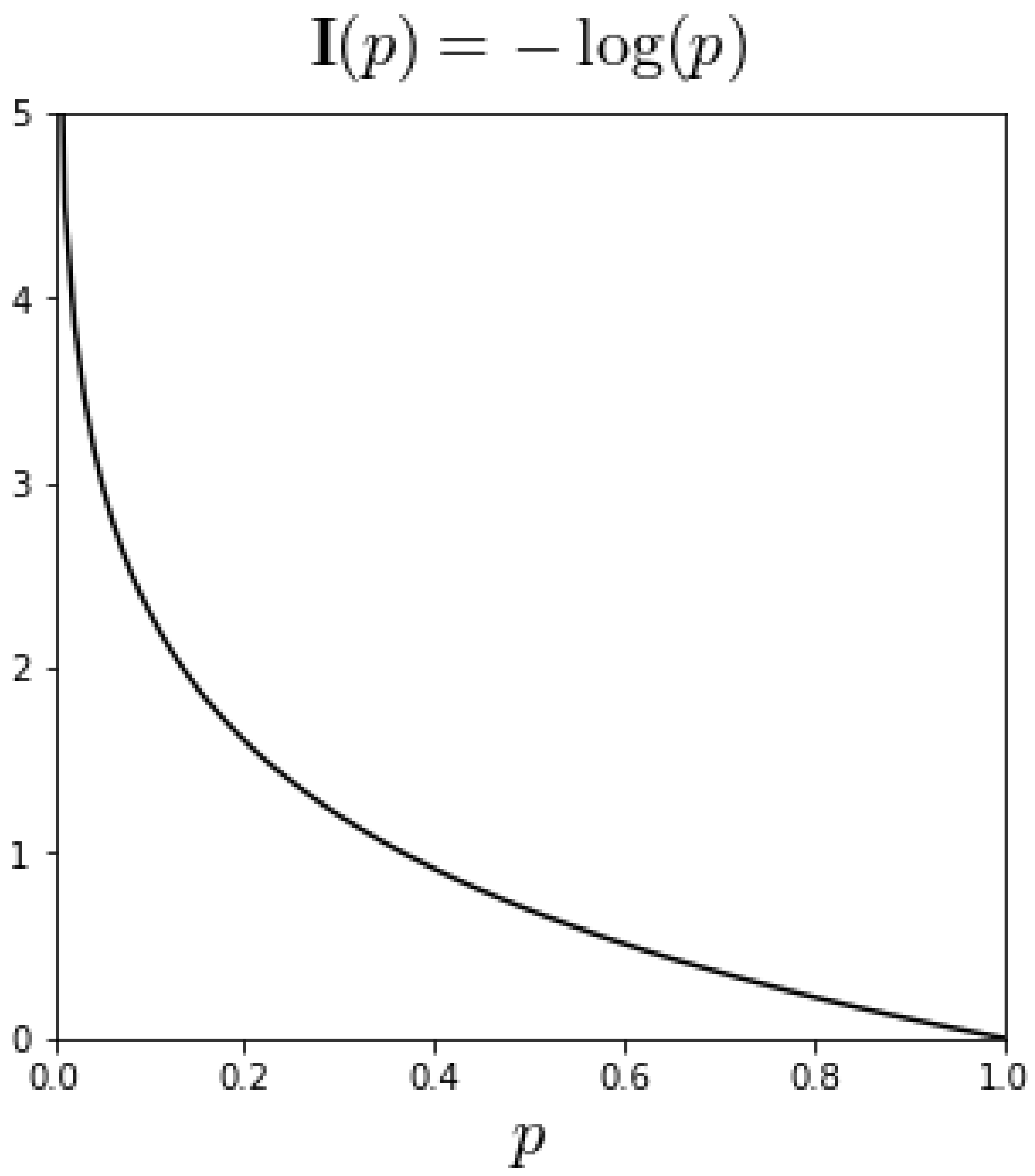
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ : 개웃김

ㅋㅋㅋㅋㅋㅋㅋ큐ㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠ : 웃긴데 본인 얘기

https://www.busan.com/view/busan/view.php?code=20120514000199

정보 엔트로피

- 정보의 확률과 양 사이의 관계는 아래 수식으로 표현 가능
- $I(x) = \log\left(\frac{1}{p(x)}\right) = -\log(p(x))$
  - $p(x)$: 정보의 발생 확률
  - $I(x)$: 사건이 발생하였을 때 얻을 수 있는 정보의 양



정보 엔트로피

- Entropy: 사건이 발생하였을 때, 얻을 수 있는 예상된 정보량
  - 사건이 발생하였을 때, 얻을 수 있는 정보량(IC)의 기댓값(Expectation)
  - 기댓값 = 사건의 확률 * 사건의 값 $= p(x) \times I(x)$
  - $H(x) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$
  - 해당 데이터로부터 얻을 수 있는 정보의 예측값
  - 낮을 수록 당연한 정보
  - 높을 수록 불확실한 / 뜻밖의 정보

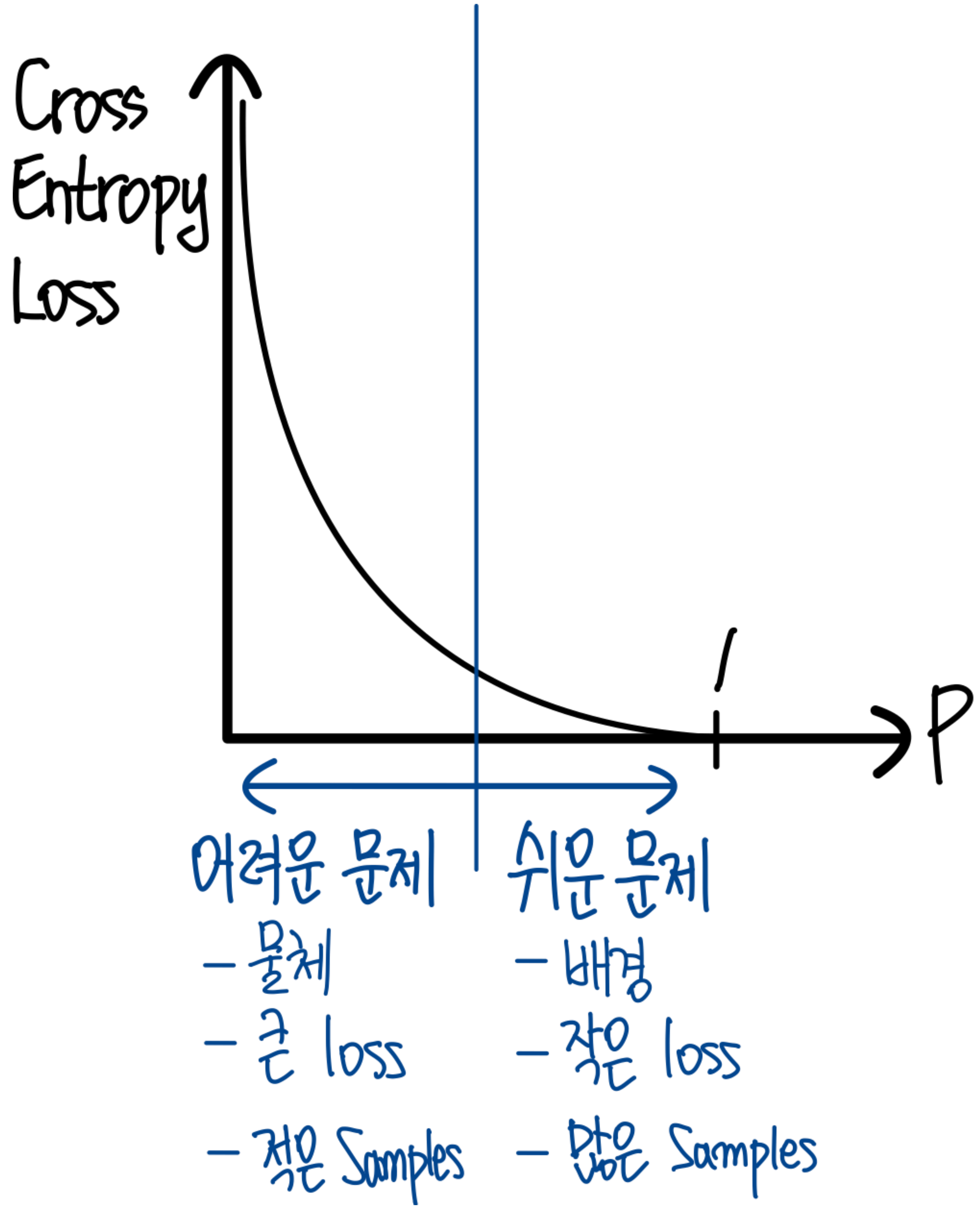
3. Cross Entropy Loss

Cross Entropy

- Cross Entropy loss
  - 서로 다른 확률 분포 집단 간 차이를 측정하는 함수
  - 모델이 예측한 값들의 분포와 정답 값들의 분포의 차이
  - $H(p,q) = -\sum_x p(x)\log\big(q(x)\big)$
    - $p(x)$: 실제 확률 분포(y_true)
    - $q(x)$: 예측 확률 분포(y_pred)

$$= \sum_x p(x) \times \big(-\log\big(q(x)\big)\big)$$

$$= \sum_x p(x) I_q(x)$$



https://ffighting.net/deep-learning-paper-review/object-detection/focal-loss/

Cross Entropy

$$H(y_{True}, y_{Pred}) = -\sum_{class} y_{True} \log(y_{Pred}) = -\log(p_t)$$

- x의 실제 분포p(x)에 따라, 모델의 예측 분포 q(x)의 정보량을 비교하는 식
  - p_t: 모델의 예측값(y_pred)
- 모델이 좋을 수록, p(x)로부터 q(x)를 잘 예측함
  - 즉 두 값이 비슷해지며, Cross entropy 값이 줄어듦
  - 반대로 두 값의 차이가 클 수록 Cross entropy가 증가하며, 모델로부터 예측된 결과가 불확실함을 의미

Cross Entropy

- 계산 예시
  - 데이터 x1에 대한 Target label의 One-hot encoding
    - None:0, Offensive:1, Hate: 0
  - 데이터 x1에 대한 모델의 prediction
    - None:0.1, Offensive:0.7, Hate: 0.2
  - $H(target, pred) = -[0 \times \log(0.1) + 1 \times \log(0.7) + 0 \times \log(0.2)] \approx 0.357$
  - 만약 prediction이 [0.33, 0.33, 0.33]라면?
  - 만약 prediction이 [0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125]라면?

4. Imbalanced dataset 손실 함수

불균형 데이터셋

- 모델은 불균형한 데이터셋에서 학습되기 어려움
  - 기본적인 교차 엔트로피 손실 함수는 모델이 잘못 예측한 경우와 정확히 예측한 경우를 동일하게 다룸
- 예시
  - None, Offensive, Hate 의 데이터 비율이 80:10:10 일 경우
  - 손실 함수는 다수 클래스에 대한 손실을 줄이면 전체 손실이 줄어듦
    - 자주 나타나는 클래스에 대해 높은 확률을 예측
    - 드물게 나타나는 클래스에 대해 낮은 확률을 예측
  - 전체 손실에 대하여
    - $L_{Total} = 0.8 \times L_{None} + 0.1 \times L_{Offensive} + 0.1 \times L_{Hate}$

불균형 데이터셋

- 모델은 불균형한 데이터셋에서 학습되기 어려움
  - 모델은 손실 함수를 따라 학습하며, 손실 값이 적은 방향으로 나아감
    - 데이터 셋이 불균형할 경우, 모델은 다수의 클래스를 예측하는 것을 선호
    - 전반적인 손실 값이 개선되더라도, 소수 클래스를 구분하는 능력이 저하됨
  - 이를 막기 위해 아래 방법들을 사용
    - Resampling
    - Weighted Cross Entropy
    - Focal loss 등 변형 손실 함수

Weighted Cross Entropy

- Weighted Cross Entropy
  - Cross Entropy 손실 함수에 클래스별 가중치를 추가하여 특정 클래스의 중요성을 조정
  - 클래스 불균형 문제를 해결하기 위한 가장 단순한 방법
  - $L_{Total} = w_A \times L_A + w_B \times L_B + w_C \times L_C$

```
# 클래스 불균형을 고려한 샘플링 가중치 계산
train_labels = [full_dataset.labels[i] for i in train_indices]  # train 데이터셋에 대한 레이블
normal_count = train_labels.count(0)
pneumonia_count = train_labels.count(1)
class_weights = 1. / torch.tensor([normal_count, pneumonia_count], dtype=torch.float)
sample_weights = [class_weights[label] for label in train_labels]
sampler = WeightedRandomSampler(weights=sample_weights, num_samples=len(sample_weights), replacement=True)
```

Weighted Cross Entropy

- 가중치 계산 방법
  - 1. 역빈도 가중치(Inverse frequency weighting)
    - 클래스의 빈도가 낮을수록 높은 가중치를 부여
    - 클래스의 수로 전체 샘플 수를 나눔
    - $w_i = \frac{N}{n_i}$ (또는 α)
    - $L = -\alpha \log(p_t)$

Weighted Cross Entropy

- 가중치 계산 방법
  - 2. 로그 확률 가중치(Logistic probability weighting)
    - 클래스 빈도의 로그 값을 이용
    - 로그는 큰 값을 줄여주므로, 빈도 차이를 어느 정도 완만하게 표현
    - 지나치게 큰 가중치 값을 부여하는 과정을 해소
    - $w_i = \log(\frac{N}{n_i})$ (또는 α)
  - 3. 사용자 정의

Focal Loss

- Focal loss
  - 불균형한 데이터셋을 다루는 작업에서 주로 사용되는 손실 함수
    - 객체 탐지 분야에서 주로 사용
  - Cross Entropy Loss의 변형(이진 분류 기준)
    - $CE(p_t) = -\log(p_t)$
    - $FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$
    - p_t: 모델의 예측값
    - $1-p_t$: 모델의 예측과 다른 클래스

Focal Loss

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

- 두 하이퍼파라미터를 추가하여 문제를 해결
  - 조정 요소(γ, gamma)
    - 어려운 예제(모델이 잘못 예측한 예제)에 더 큰 가중치를 부여
    - 쉬운 예제(모델이 잘 예측한 예제)의 손실 기여도를 줄이는 계수
  - 클래스별 가중치(α, alpha)
    - 드문 클래스에 대한 가중치를 높여 모델이 이러한 클래스에 더 신경 쓰도록 유도하는 계수
    - [0, 1]구간의 값

Focal Loss

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t)$$

- 작동 방식
  - 쉬운 문제
    - 모델이 확실하게 맞춘 예제들에 대해 p_t 가 1에 가까우면
    - $(1 - p_t)^{\gamma}$가 0에 수렴
    - 해당 예제의 손실이 감소
  - 어려운 문제
    - 모델이 잘못 예측한 예제들에 대해 p_t 가 0에 가까우면
    - $(1 - p_t)^{\gamma}$가 커짐
    - 손실이 증가하여, 학습이 추가적으로 필요한 신호를 보냄