10.3 교차엔트로피와 쿨백-라이블러 발산

- 교차엔트로피 → 분류기 성능 평가 지표
- 쿨백-라이블러 발산 → 교차엔트로피를 응용한 것으로, 두 확률분포의 모양이 얼마나 유사한지 평가.

교차-엔트로피

두 확률분포 p, q의 교차-엔트로피 (cross entropy) $H[p, q]$는

이산확률분포의 경우,

$$H[p, q] = -\sum_{k=1}^{K} p(y_k) \log_2 q(y_k)$$

연속확률분포의 경우,

$$H[p, q] = -\int_y p(y) \log_2 q(y) \, dy$$

* 교차엔트로피를 자세히 공부한 엔트로피, 결합엔트로피, 조건부엔트로피와 다르게 확률변수가 아닌 확률분포를 인수로 받는다는 점에 유의!

예제 다음 두 분포의 교차엔트로피 = 0.25

$$p = [1/4, 1/4, 1/4, 1/4]$$
$$q = [1/2, 1/4, 1/8, 1/8]$$

$$= -\frac{1}{4} \cdot \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \cdot \log_2 \frac{1}{8} - \frac{1}{4} \cdot \log_2 \frac{1}{8}$$

$$= 1/4$$

교차엔트로피를 사용한 분류성능측정

"Y가 0 또는 1이라는 값만 갖는 이진분류문제"

- p는 X값이 정해졌을 때 정답인 Y의 확률분포다.
- 이진분류문제에서 Y는 0 또는 1이다.

따라서 p는

- 정답이 $Y=1$일 때,

$$p(Y=0)=0, \quad p(Y=1)=1$$

- 정답이 $Y=0$일 때,

$$p(Y=0)=1, \quad p(Y=1)=0$$

분포 q_θ는 X값이 정해졌을 때의 예측값의 확률분포다. 모수가 μ인 베르누이분포로 가정한다.

$$q_\theta(Y=0)=1-\mu$$
$$q_\theta(Y=1)=\mu$$

따라서 확률분포 p와 q_θ의 교차엔트로피는,

- 정답이 $Y=1$일 때, ($p(Y=0)=0$, $p(Y=1)=1$)

$$H[p,q_\theta] = -\cancel{p(Y=0)\log_2 q_\theta(Y=0)} - p(Y=1)\log_2 q_\theta(Y=1)$$
$$= -\log_2 \mu$$

- 정답이 $Y=0$일 때, ($p(Y=0)=1$, $p(Y=1)=0$)

$$H[p,q_\theta] = -p(Y=0)\log_2 q_\theta(Y=0) - \cancel{p(Y=1)\log_2 q_\theta(Y=1)}$$
$$= -\log_2(1-\mu)$$

교차엔트로피는 분류성능이 좋을수록 작아지고, 분류성능이 나쁠수록 커진다.

이유는 다음과 같다.

→ $Y=1$ 일 때는 μ가 작아질수록, 즉 예측이 틀릴수록 $-\log_2 \mu$의 값도 커진다.

→ $Y=0$일 때는 μ가 커질수록, 즉 예측이 틀릴수록 $-\log_2(1-\mu)$ 값도 커진다.

따라서, 교차엔트로피 값은 예측의 틀린 정도를 나타내는 오차함수의 역할을 할수 있다.
N개의 학습 데이터 전체에 대해, 교차엔트로피 평균을 구하면 다음식으로 표현할수 있다. 이 값을 로그손실 (log loss) 이라고도 한다.

(binary)

$$\text{log loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{y_i \log_2 \mu_i}_{\text{정답}(Y=1)} + \underbrace{(1-y_i)\log_2(1-\mu_i)}_{\text{정답}(Y=0)}\right)$$

같은 방법으로, 이진분류가 아닌 다중분류에서도 교차엔트로피를 오차 함수로 사용할수 있다.
다중분류문제의 교차엔트로피 손실함수를 카테고리 로그손실 (categorical log-loss)
이라고 한다.

$$\text{categorical log loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\underbrace{\mathbb{I}(y_i=k)}_{\text{지시함수}}\log_2 \underbrace{p(y_i=k)}_{\text{분류기가 계산한 } y_i=k \text{ 확률}}\right)$$

* 위 식에서 $\mathbb{I}(y_i=k)$는 $y_i=k$ 일 때만 1이고, 그렇지 않으면 0이 되는 지시함수 (indicator function) 다. $p(y_i=k)$는 분류모형이 계산한 $y_i=k$ 확률이다.

(사이킷런 (Scikit-learn) 패키지의 metrics 서브패키지는 로그손실을 계산하는 log_loss 함수를 제공. sklearn.metrics.log_loss)

쿨백-라이블러 발산 → 변분법적 추론 (Variational ~~difference~~ inference)

↳ Deeplearning → VAE (variational auto encoder) ?

GAN!

쿨백-라이블러 발산 (Kullback-Leibler divergence)은 두 확률분포 $p(y)$, $q_k(y)$의 분포모양이 얼마나 다른지를 측정하는 계산한 값이다.

$KL(p||q_k)$로 표기한다.

(이산확률분포)

$$KL(p||q_k) = H[p,q_k] - H[p]$$

$$= \sum_{i=1}^{K} p(y_i) \log_2 \left(\frac{p(y_i)}{q_k(y_i)} \right)$$

(연속확률분포)

$$KL(p||q) = H[p,q_k] - H[p]$$

$$= \int p(y) \log_2 \left(\frac{p(y)}{q_k(y)} \right) dy$$

쿨백-라이블러 발산 = 교차엔트로피에서 기준이 되는 p분포의 엔트로피값을 뺀 값

= 상대 엔트로피 (relative entropy)

(값이 항상 양수이며, $p(x)$, $q_k(x)$ 두 확률분포가 완전히 같을 경우에만 0이 된다.)

when $p(x) = q(x)$,

$$KL(p\|p) = H[p,p] - H[p]$$

$$= \int p(y) \log_2 \left(\frac{p(y)}{p(y)}\right) dy$$

$$= 0$$

$$KL(p\|q) = 0 \iff p = q$$

쿨백-라이블러 발산은 거리 (distance)가 아니라, 확률분포 q가 기준확률분포 p와 얼마나 다른지를 나타내는 값이고, 두 확률분포의 위치가 달라지면 일반적으로 값도 달라진다.

$$KL(p\|q) \neq KL(q\|p)$$

∵ 쿨백-라이블러 발산 = 상대 엔트로피.

예제

가변길이 인코딩과 쿨백-라이블러 발산

4개의 글자 A, B, C, D로 씌여진 다음과 같은 문서를 가변길이 인코딩할 경우를 생각하자.

'DBCADDAAAAA···'

이 문서를 구성하는 글자의 확률분포는 다음과 같다.

$$p(Y=A) = \frac{1}{2},\ p(Y=B) = \frac{1}{4},\ p(Y=C) = \frac{1}{8},\ p(Y=D) = \frac{1}{8}$$

이 때, 한 글자당 인코딩된 글자수는 분포 q의 엔트로피인 1.75가 된다.

$$\sum_{k=1}^{K} p(y_i) \log_2 p(y_i) = -\frac{1}{2} \cdot \log \frac{1}{2} + \left(-\frac{1}{4} \cdot \log_2 \frac{1}{4}\right) + \left(-\frac{1}{8} \cdot \log_2 \frac{1}{8}\right)$$
$$+ \left(-\frac{1}{8} \cdot \log_2 \frac{1}{8}\right) = \boxed{1.75}$$

그런데, 가변길이 인코딩을 사용하지않고 고정길이 인코딩을 사용한다는 것은 다음과 같은 분포를 가정한 것과 같다.

$$q(Y=A)=\frac{1}{4},\quad q(Y=B)=\frac{1}{4},\quad q(Y=C)=\frac{1}{4},\quad q(Y=D)=\frac{1}{4}$$

실제로 한 글자당 인코딩된 글자수는 다음과 같이 계산할 수 있다.

$$\sum_{i=1}^{K} p(y_i)\log_2 q(y_i) = -\frac{1}{2}\cdot\log_2\frac{1}{4} + \left(-\frac{1}{4}\cdot\log_2\frac{1}{4}\right) + \left(-\frac{1}{8}\cdot\log_2\frac{1}{8}\right) + \left(-\frac{1}{8}\cdot\log_2\frac{1}{4}\right)$$

$$= \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 2 + \frac{1}{8}\cdot 2$$

$$= 2$$

쿨백-라이블러 발산은 잘못된 분포 q로 인코딩되었을 때의 한 글자당 인코딩된 글자수와 원래의 분포 p를 사용하였을 때 한 글자당 인코딩된 글자수의 차이인 0.25와 같다.

$$KL(p\|q) = \sum_{i=1}^{K} p(y_i)\log_2\left(\frac{p(y_i)}{q(y_i)}\right)$$

$$= -\sum_{i=1}^{K} p(y_i)\log_2 q(y_i) - \left(-\sum_{i=1}^{K} p(y_i)\log_2 p(y_i)\right)$$

$$= H[p,q] - H[p]$$

$$= 2.0 - 1.75 = 0.25$$

즉, 확률분포 q의 모양이 확률분포 p와 다른 정도를 정량화한 값
= 쿨백-라이블러 발산

scipy.stats.entropy 함수 → 두 개의 확률분포를 인자로 넣으면 쿨백-라이블러 발산을 계산 (base=2)