크로스 엔트로피 (cross entrophy) $H[p,q]$

- 두 확률분포 $p(y)$, $q(y)$의 크로스 엔트로피 $H[p,q]$는 다음과 같이 정의한다.

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

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

↳ "크로스 엔트로피"는 주로 분류문제 (classification)의 "목표값 분포"와 "예측값 분포"를 비교하는데 사용된다."

↳ "크로스 엔트로피"는 언제 사용하는가?

⇒ p 하고 q가 얼마나 닮았는지 측정할 때 사용합니다.

⇒ 일종의 Error Function 입니다.

⇒ p 하고 q하고 비슷하면 크로스엔트리 값은 ↓

p 하고 q가 다르면 크로스엔트리 값은 ↑

쿨백-라이블러 발산 (Kullback-Leibler divergence) ⇒상대 엔트로피

↳ 두 확률분포 $p(y)$, $q(y)$의 차이를 정량화하는 방법의 하나. (relative entrophy)

$$KL(p\|q) = H[p,q] - H[p] = \int p(y) \log_2 \left(\frac{p(y)}{q(y)}\right) dy$$

⇒ 값은 항상 양수이며, 두 확률분포 $p(x)$, $q(x)$가 완전히 같을 때만 0이 됨.

↳ 크로스 엔트로피, 엔트로피, 쿨백 라이블러 발산 모두 $\log$ 값을 계산하므로 계산량이 많아서 다른 방식을 찾아보니 "지니 불순도" (Gini Impurity)사용.

지니 불순도 (Gini Impurity)

↳ 엔트로피처럼 확률분포가 얼마나 치우쳐있는가를 재는 척도이지만 로그를 사용하지 않으므로 계산량이 적어 엔트로피 대신 사용됨.

$$G[Y] = \sum_{k=1}^{K} p(y_k)(1 - p(y_k))$$