Based on Information Theory

당연하다고 생각되는 사건은 그 확률이 높다는 얘기가 된다. 당연한만큼 자주 사용되므로 그 사건의 정보량은 작다

즉, 자주 발생하는 사건의 확률이 높으면 그 사건의 정보량은 작다.

h의 첫 번째 조건

- 확률 변수(Random variable) X
  1. X는 East, West 두 가지 값을 가질 수 있음.
- X 의 정보량 h(x) 는 p(x)에 대한 함수.   즉, h=f(p)
- $p(east) = 0.99999999,\ p(west) = 0.00000001$
- $h(west) > h(east)$ 여야 함.
- p(x)와 h(x)는 monotonic한 관계여야 한다. 즉, f는 단조 감소 함수.

확률 변수 X가 2개의 값을 가진다고 했을 때, 그 확률 변수 X의 확률은 p(X)이고, 그 정보량은 h(x)이다. 이때 정보량은 확률에 대한 함수이므로, h = f(p)라고 표현.

확률이 높으면 정보량은 작으므로 p(x)와 h(x)는 반비례 관계이다. 따라서 f는 단조 감소 함수.

h의 두 번째 조건

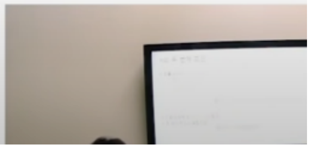
- 확률 변수(Random variable) X, Y
  1. X는 East, West 두 가지 값
  2. Y는 Rain, Not Rain 두 가지 값
  3. X,Y 는 독립

- $h(x, y) = h(x) + h(y)$
- $p(x, y) = p(x) * p(y)$

즉, f(p(x,y)) = f(p(x) * p(y)) = f(p(x)) + f(p(y))

- 이를 만족하는 f는 => log 함수.
- 첫 번째 조건과 결합하면

$$\boldsymbol{h(x) = -log_2 p(x)}$$



확률 변수가 X,Y 2개이고 서로 독립일 때, 정보량과 확률은 다음과 같이 표현가능하다.

h가 p(x)의 함수, 즉 h = f(p(x))이므로 대입하면,

h(x,y) = f(p(x,y)) = f(p(x)*p(y)) = f(p(x)) + f(p(y))로 나타낼 수 있다. 따라서 이 조건을 만족시키는 함수인 log 사용. 단조 감소 함수이므로 -를 붙이지만 p(X)는 0~1사이의 값이기 때문에 h(x)는 양수이다. 이때 밑이 2이면 bit단위이다.

h(x)의 평균을 내면 (확률을 곱해주면) 평균 정보량이 된다.

얼마나 자주 사용되는지에 대한 확률 X 인코딩 길이

$$\boldsymbol{h(x) = -log_2 p(x)}$$

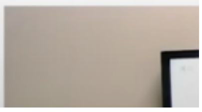
- $h(east) = -log_2 p(east) = -log_2(0.99999999) = 0.000000014$
- $h(west) = -log_2 p(west) = -log_2(0.00000001) = 26.5754247591$

그러면 평균적인 정보량은 ..?

- $p(east) * h(east) + p(west) * h(west) = 0.99999999 * 0.000000014 + 0.00000001 * 26.6$
- 보다 일반적으로는,

$$H[X] = -\sum_x p(x) log_2 p(x) = E_p[-log_2 p(x)]$$

- 이 값이 바로 ENTROPY !!



Entropy 의 몇 가지 특징

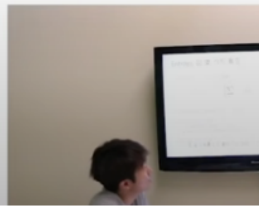
- Continuous Variable 인 경우

$$\mathrm{H}[\mathbf{x}] = \lim_{\Delta \to 0} \left\{ \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) \, \mathrm{d}x$$

- Entropy 는 Average Coding Length의 Lower Bound!

- Entropy Maximize ?
  1. Discrete variable : Uniform
  2. Continuous variable : Gaussian

- Entropy Minimize ?
  1. 한 점에 확률이 다 몰려 있는 경우! -> 0이 된다.

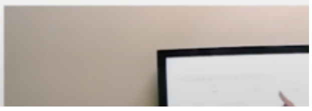


즉, 모델링 오류때문에 발생한 추가 비용은

$$\left(-\sum_x p(x) log_2 q(x)\right) - \left(-\sum_x p(x) log_2 p(x)\right) = -\sum_x p(x) log_2 \frac{q(x)}{p(x)}$$

Continuous variable의 경우는

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned}$$



즉, 모델링 오류때문에 발생한 추가 비용은

$$\left(-\sum_x p(x)log_2 q(x)\right)-\left(-\sum_x p(x)log_2 p(x)\right) = -\sum_x p(x)log_2\frac{q(x)}{p(x)}$$

Continuous variable의 경우는

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}.\end{aligned}$$

Cross-Entropy

$$\begin{aligned}\mathrm{KL}(p\|q) &= \underbrace{\boxed{-\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x}}}_{\text{Cross Entropy}} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}.\end{aligned}$$

- $KL(p|q) = H(p,q) - H(p)$

Classification 할때 왜 Loss 함수는 Cross Entropy 일까?

- p(모분포, 정답)를 근사하기 위해 q(뉴럴넷)를 만들었음.
- H(p)는 q와 무관함. 즉, q의 parameter로 미분하면 사라짐.
- 그래서 H(p,q)를 loss 함수로 씀. -> KL 을 쓰는 것과 마찬가지.

P|Q : 내가 모델링 한 것과 실제 모델의 평균 정보량의 차이. 내가 모델링 했다는 의미는 인코딩, 즉 정보량을 모델링 한 것이지 확률변수의 확률을 조작할 순 없다.

실제 실험을 통해, 모분포를 추정하여 정보량을 추정한 것이라서 확률에는 영향을 미치지 않는다. (확률은 절대적임)

따라서 실제 모델의 값은 Lower Bound 값이다. (인코딩 = 정보량의 길이가 최소이기 때문에)

KL Divergence는 (내가 모델링한 평균 정보량) - (실제 모델의 평균 정보량) or 두 모델의 유사도 (거리)

평균 정보량은 항상 양수이다.

값이 작을수록 내가 모델링을 잘 했다는 얘기가 된다.

모든 확률변수 X에 대해서, q(x)와 p(x)가 같으면 됨.

Mutual Information

- x와 y가 Independent 면 p(x,y) = p(x) * p(y)
- 만일 Independent가 아니라면, KL Divergence를 이용하여 p(x)*p(y) 가 p(x,y)에 얼마나 가까운에 대한 idea를 얻을 수 있다.

$$
\begin{aligned}
\mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\
&= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}
\end{aligned}
$$

- $I(x,y) \geq 0$, x,y가 독립일 때 등호 성립