인공지능 수학
(인공지능관련 수학개념들)

Some Math & Probability

- Expected value
- Covariance
- Differentiation
- Information Theory

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition A" is a also a random variable (the percentage will be slightly differently every time you poll).
- Roughly, probability is how frequently we expect different outcomes to occur if we repeat the experiment over and over

Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Expected value

- Weighted average of possible **values** of a random variable

Discrete case:

$$E(X) = \sum_{\text{all x}} x_i \, p(x_i)$$

Continuous case:

$$E(X) = \int_{\text{all x}} x_i \, p(x_i) dx$$

Expected value

- If X is a random integer between 1 and 10, what's the expected value of X?

Expected value

- If X is a random integer between 1 and 10, what's the expected value of X?

$$\mu = E(x) = \sum_{i=1}^{10} i(\frac{1}{10}) = \frac{1}{10}\sum_{i}^{10} i = (.1)\frac{10(10+1)}{2} = 55(.1) = 5.5$$

Average vs. Expected Value

- E.g. Random integer between 1 and 10

  – Average (a.k.a. arithmetic mean)

    Given a list of (4, 6, 9, 1, 10)

      Average: (4+6+9+1+10) / 5

  – Expected Value : 5.5

Variance/standard deviation

"The expected squared distance (or deviation) from the mean"

i.e. spread of a data set around its mean value

$$\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Variance

Discrete case:

$$Var(X) = \sigma^2 = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Continuous case:

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 \, p(x_i) dx$$

$$\text{Var}(x) = E(x-\mu)^2 = E(x^2) - [E(x)]^2$$

$$Var(X) = \sum_{\text{all x}} (x_i - \mu)^2 p(x_i) = \sum_{\text{all x}} x_i^2 p(x_i) - (\mu)^2$$

$$= E(x^2) - [E(x)]^2$$

Proofs (optional):

$E(x-\mu)^2 = E(x^2 - 2\mu x + \mu^2)$

$= E(x^2) - E(2\mu x) + E(\mu^2)$ expected value: $E(X+Y) = E(X) + E(Y)$

$= E(x^2) - 2\mu E(x) + \mu^2$ $E(c) = c$

$= E(x^2) - 2\mu\mu + \mu^2$ $E(x) = \mu$

$= E(x^2) - \mu^2$

$= E(x^2) - [E(x)]^2$

Variance

Find the variance and standard deviation for the number of ships to arrive at the harbor

x	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1

(the mean is 11.3).

Variance and std dev

x^2	100	121	144	169	196
$P(x)$	.4	.2	.2	.1	.1

$$E(x^2) = \sum_{i=1}^{5} x_i^{\,2} p(x_i) = (100)(.4) + (121)(.2) + 144(.2) + 169(.1) + 196(.1) = 129.5$$

$$Var(x) = E(x^2) - [E(x)]^2 = 129.5 - 11.3^2 = 1.81$$

$$stddev(x) = \sqrt{1.81} = 1.35$$

Interpretation: On an average day, we expect 11.3 ships to arrive in the harbor, plus or minus 1.35. This gives you a feel for what would be considered a usual day.

Gaussian (Normal)

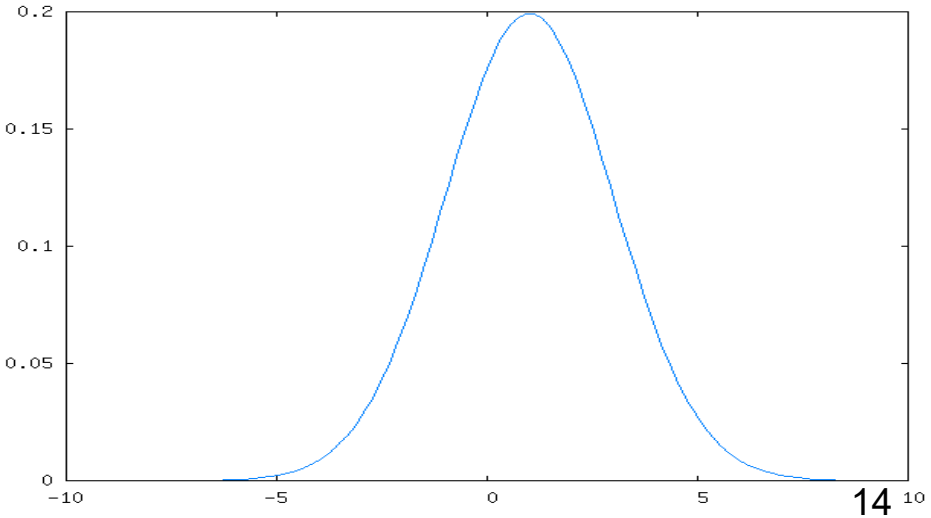
- If I look at the height of women in country xx, it will look approximately Gaussian
- Small random noise errors, look Gaussian/Normal

- Distribution:

$$x \sim N(\mu, \sigma^2) \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Mean/var

$$E[x] = \mu$$
$$Var(x) = \sigma^2$$

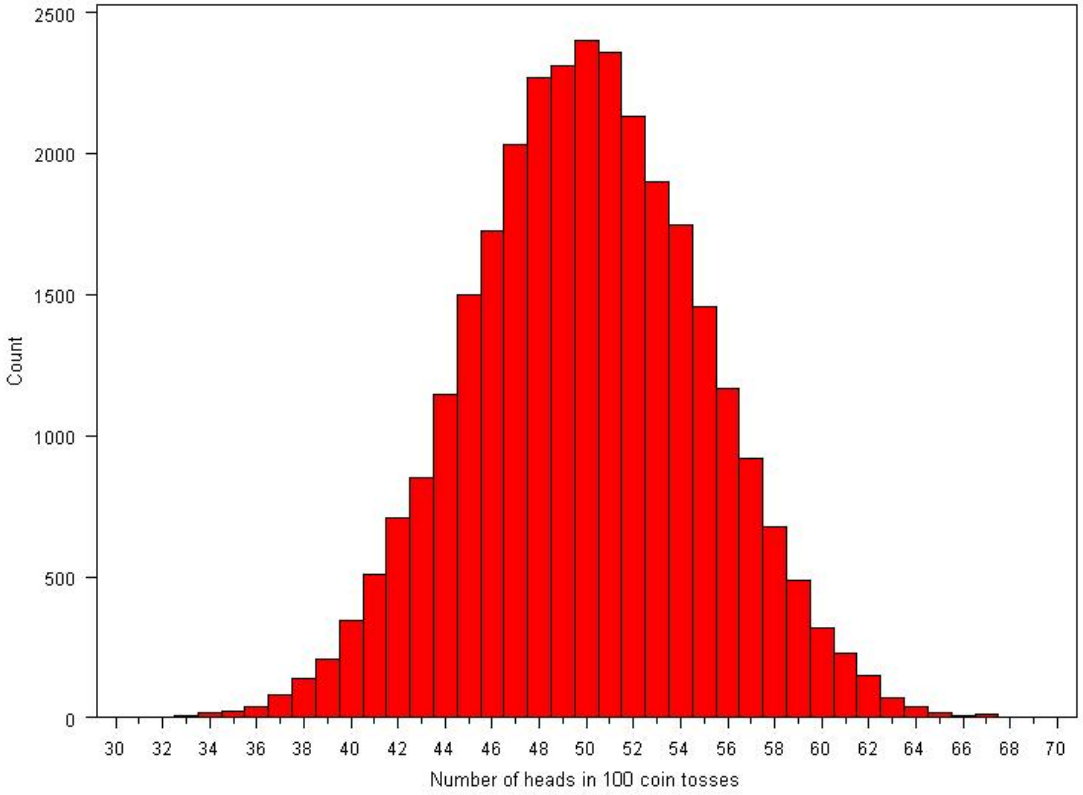


Coin tosses

- Number of heads in 100 tosses
  - Flip coins virtually
  - Flip a virtual coin 100 times; count # of heads
  - Repeat this over and over again a large number of times (e.g. 30,000)
  - Plot the results

Coin tosses

- Number of heads in 100 tosses
  - 30,000 trials



Mean = 50

Std. dev = 5

Follows a normal distribution

∴95% of the time, we get between 40 and 60 heads…

Covariance: joint probability

- The covariance measures the strength of the linear relationship between two variables

  e.g. A positive covariance means both investments' returns tend to move upward or downward in value at the same time.

$$E[(x-\mu_x)(y-\mu_y)]$$

$$\sigma_{xy} = \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)\, P(x_i, y_i)$$

Interpreting Covariance

- **Covariance** between two random variables:

$$E[(x-\mu_x)(y-\mu_y)]$$

cov(X,Y) > 0 X and Y are positively correlated

cov(X,Y) < 0 X and Y are inversely correlated

cov(X,Y) = 0 X and Y are independent

공분산 계산

$$X = D - mean(D)$$

$$X = \begin{pmatrix} | & | & | & & | \\ X_1 & X_2 & X_3 & \cdots & X_d \\ | & | & | & & | \end{pmatrix} \in \mathbb{R}^{n \times d}$$

$$X^T X = \begin{pmatrix} — & X_1 & — \\ — & X_2 & — \\ & \cdots & \\ — & X_d & — \end{pmatrix} \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

$$\frac{X^T X}{n} = \frac{1}{n} \begin{pmatrix} dot(X_1, X_1) & dot(X_1, X_2) & \cdots & dot(X_1, X_d) \\ dot(X_2, X_1) & dot(X_2, X_2) & \cdots & dot(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ dot(X_d, X_1) & dot(X_d, X_2) & \cdots & dot(X_d, X_d) \end{pmatrix}$$

공분산 계산

$$D = \begin{bmatrix} 170 & 70 \\ 150 & 45 \\ 160 & 55 \\ 180 & 60 \\ 170 & 80 \end{bmatrix}$$

$$X = D - mean(D) = \begin{bmatrix} 170 & 70 \\ 150 & 45 \\ 160 & 55 \\ 180 & 60 \\ 172 & 80 \end{bmatrix} - \begin{bmatrix} 166 & 62 \\ 166 & 62 \\ 166 & 62 \\ 166 & 62 \\ 166 & 62 \end{bmatrix} = \begin{bmatrix} 4 & 8 \\ -16 & -17 \\ -6 & -7 \\ 14 & -2 \\ 6 & 18 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & -16 & -6 & 14 & 6 \\ 8 & -17 & -7 & -2 & 18 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ -16 & -17 \\ -6 & -7 \\ 14 & -2 \\ 6 & 18 \end{bmatrix} = \begin{bmatrix} 540 & 426 \\ 426 & 730 \end{bmatrix}$$

$$\Sigma = \frac{1}{5} X^T X = \frac{1}{5} \begin{bmatrix} 540 & 426 \\ 426 & 730 \end{bmatrix} = \begin{bmatrix} 108 & 85.2 \\ 85.2 & 146 \end{bmatrix}$$

Bayes Rule

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)}$$

- Which tells us:
  - how often A happens given that B happens, written P(A|B),
- When we know:
  - how often B happens given that A happens, written P(B|A)
  - how likely A is on its own, written P(A)
  - how likely B is on its own, written P(B)
- E.g.
  - P(Fire|Smoke) means how often there is fire when we can see smoke
  - P(Smoke|Fire) means how often we can see smoke when there is fire

Bayes' Rule

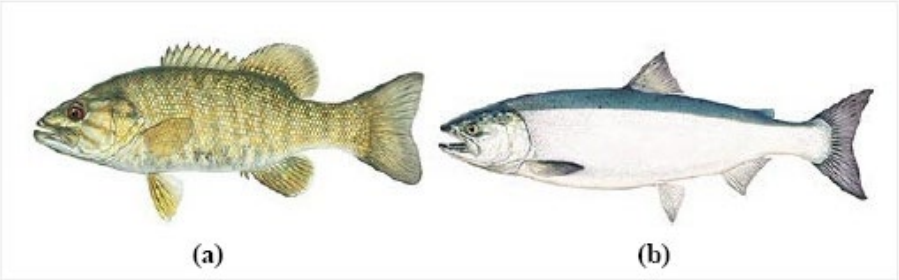


FIGURE 1: 농어(a)와 연어(b)

- **Posterior($P(w_i|x)$)**: 피부 밝기(x)가 주어졌을 때 그 물고기가 농어일 확률 또는 연어일 확률. 즉 단서가 주어졌을 때, 대상이 특정 클래스에 속할 확률. 우리가 최종적으로 구해야 하는 값이다.
- **Likelihood($P(x|w_i)$)**: 농어 또는 연어의 피부 밝기(x)가 어느 정도로 분포되어 있는지의 정보. 즉 각 클래스에서 우리가 활용할 단서가 어떤 형태로 분포 돼 있는지를 알려준다. Posterior를 구하는 데 있어서 매우 중요한 단서가 된다.
- **Prior($P(w_i)$)**: 피부 밝기(x)에 관계 없이 농어와 연어의 비율이 얼마나 되는지의 값. 보통 사전 정보로 주어지거나, 주어지지 않는다면 연구자의 사전 지식을 통해 정해줘야 하는 값이다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)} \qquad P(w_i|x) = \frac{P(x|w_i)P(w_i)}{\sum_j P(x|w_j)P(w_j)}$$

https://hyeongminlee.github.io/post/bnn001_bayes_rule/

Differentiation

Differentiation is all about measuring change.
E.g. Measuring change in a linear function:

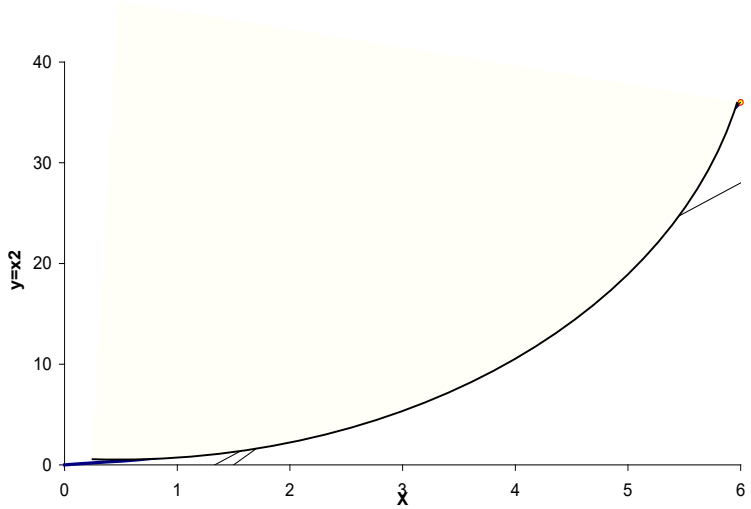
$$y = a + bx$$

a = intercept

b = constant slope i.e. the impact of a unit change in x on the level of y

$$\mathbf{b} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Example: A firms cost function is $Y = X^2$



X	ΔX	Y	ΔY
0		0	
1	+1	1	+1
2	+1	4	+3
3	+1	9	+5
4	+1	16	+7

$$Y = X^2$$

$$Y+\Delta Y = (X+\Delta X)^2$$

$$Y+\Delta Y = X^2+2X.\Delta X+\Delta X^2$$

$$\Delta Y = X^2+2X.\Delta X+\Delta X^2 - Y$$

since $Y = X^2 \quad \Rightarrow \quad \Delta Y = \quad 2X.\Delta X+\Delta X^2$

$$\frac{\Delta Y}{\Delta X} = 2X+\Delta X$$

The slope depends on X and ΔX

The slope of the graph of a function is called the **derivative** of the function

$$f'(x) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

- The process of differentiation involves letting the change in x become arbitrarily small, i.e. letting $\Delta x \to 0$

e.g if = 2X+ΔX and $\Delta X \to 0$

$\Rightarrow$ = 2X in the limit as $\Delta X \to 0$

Differentiation

- Examples:

  $f(x) = 3$

  $f(x) = 4x$

  $f(x) = 4x^2$

Differentiation

- Product Rule

If y = $u.v$ where u and v are functions of x,
($u = f(x)$ *and* $v = g(x)$) Then

$$\frac{dy}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$$

- Examples

i) $y = (x+2)(ax^2+bx)$

$$\frac{dy}{dx} = (x+2)(2ax+b) + (ax^2+bx)$$

ii) $y = (4x^3-3x+2)(2x^2+4x)$

$$\frac{dy}{dx} = (4x^3-3x+2)(4x+4) + (2x^2+4x)(12x^2-3)$$

Differentiation

- **The Chain Rule**

  If y is a function of v, and v is a function of x, then y is a function of x and

$$\frac{dy}{dx} = \frac{dy}{dv} \cdot \frac{dv}{dx}$$

- Example:

  i) $y = (ax^2 + bx)^{½}$

  let $v = (ax^2 + bx)$, so $y = v^{½}$

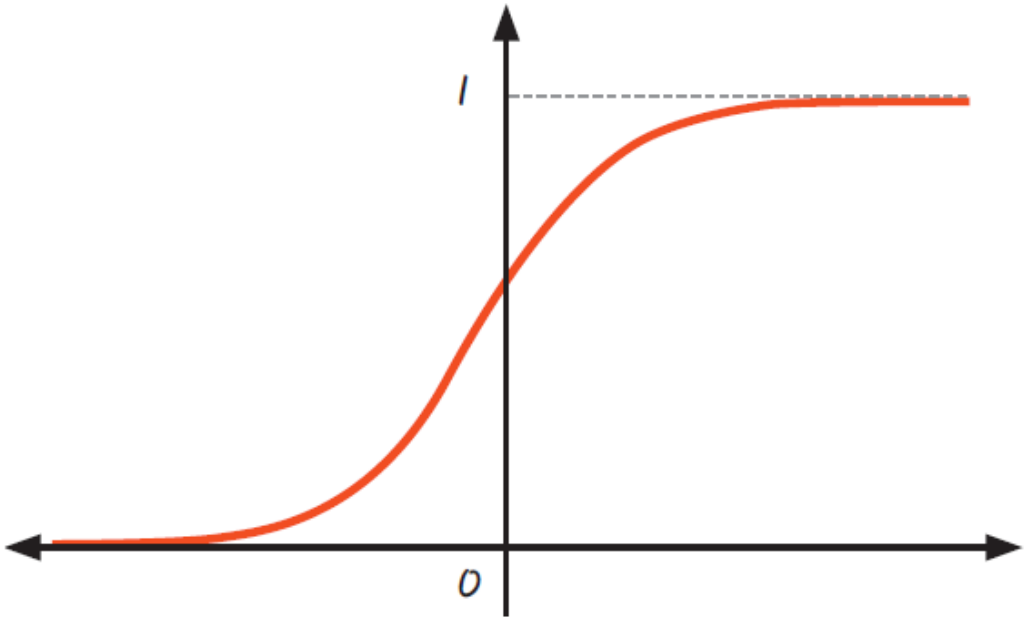
$$\frac{dy}{dx} = \frac{1}{2}\left(ax^2 + bx\right)^{-\frac{1}{2}} .(2ax + b)$$

Sigmoid Function

- 시그모이드 함수 :
  - 지수 함수에서 밑의 값이 자연 상수e인 함수를 말함
- 자연 상수 e :
  - '자연 로그의 밑', '오일러의 수' 등 여러 이름으로 불림
- 파이(π)처럼 수학에서 중요하게 사용되는 무리수임
- 그 값은 대략 2.718281828…임
- 자연 상수 e가 지수 함수에 포함되어 분모에 들어가면 시그모이드 함수가 됨

$$f(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid Function



Information Theory

- 핵심 아이디어: 잘 일어나지 않는 사건(unlikely event)은 자주 발생하는 사건보다 정보량이 많다(informative)는 것
  - 자주 발생하는 사건은 낮은 정보량을 가진다. 발생이 보장된 사건은 그 내용에 상관없이 전혀 정보가 없다는 걸 뜻한다.
  - 덜 자주 발생하는 사건은 더 높은 정보량을 가진다.
  - 독립사건(independent event)은 추가적인 정보량(additive information)을 가진다. 예컨대 동전을 던져 앞면이 두번 나오는 사건에 대한 정보량은 동전을 던져 앞면이 한번 나오는 정보량의 두 배이다.

https://ratsgo.github.io/statistics/2017/09/22/information/

Shannon's Entropy

- 확률변수 X의 값이 x인 사건의 정보량:

$$I(x) = -\log P(x)$$

  - 예: 동전을 던져 앞면이 나오는 사건: $-\log_2 0.5 = 1$,
  - 예: 주사위를 던져 눈이 1이 나오는 사건: $-\log_2 1/6 = 2.5849$

  % 밑이 2인 경우 정보량의 단위를 섀년(shannon) 또는 비트(bit)라고 함.

- Shannon's Entropy: 모든 사건 정보량의 기대값

$$H(P) = H(x) = E_{X\sim P}[I(x)] = E_{X\sim P}[-\log P(x)]$$

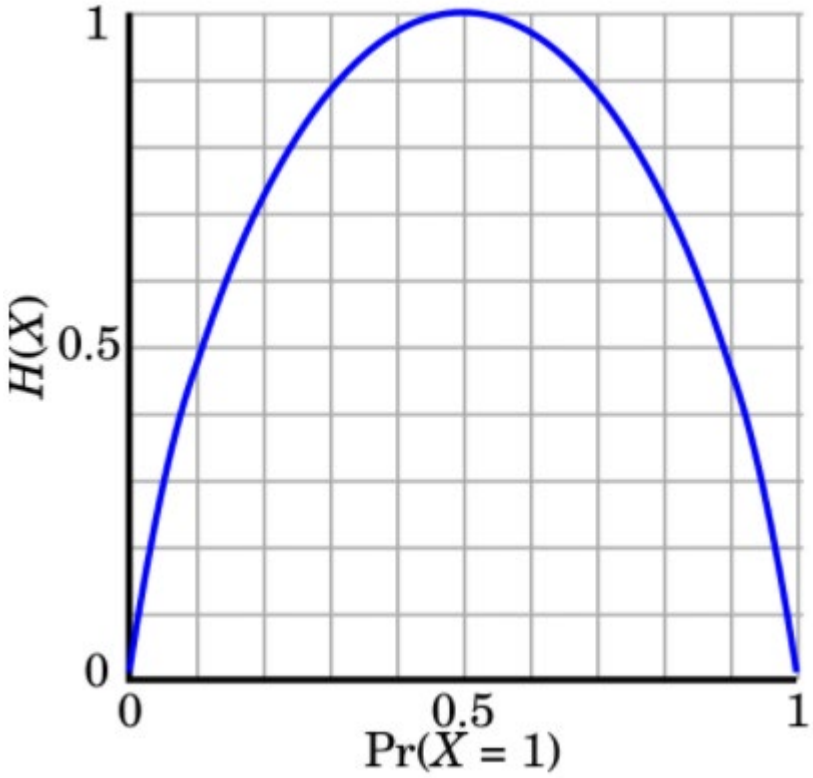
- 앞면, 뒷면이 나올 확률이 동일한, 공평한 동전을 1번 던지면

$$\begin{aligned} H(P) = H(x) &= -\sum_x P(x)\log P(x) \\ &= -(0.5\times\log_2 0.5 + 0.5\times\log_2 0.5) \\ &= -\log_2 0.5 \\ &= -(-1) \end{aligned}$$

https://ratsgo.github.io/statistics/2017/09/22/information/

$$H(P) = H(x) = -\sum_{x} P(x) \log P(x)$$

Entropy



- x 축은 동전의 공정한 정도(1, 즉 앞면이 나올 확률)

https://ratsgo.github.io/statistics/2017/09/22/information/

$$H(P) = H(x) = -\sum_{x} P(x) \log P(x)$$

Cross Entropy

$$H(P,Q) = E_{X \sim P}[-\log Q(x)] = -\sum_{x} P(x) \log Q(x)$$

- H(P,Q) 는 P의 엔트로피인 H(P)와 유사한 꼴이나 로그 바깥에 곱해지는 확률이 P(x)이고, 로그 안에 들어가는 것이 Q(x). 엔트로피는 엔트로피이되 두 확률분포가 교차로 곱해진다

https://ratsgo.github.io/statistics/2017/09/22/information/

KL Divergence

- 두 확률분포의 차이를 계산
- 실제 데이터의 분포 P(x)와 모델이 추정한 데이터의 분포 Q(x) 간에 차이를 KLD를 활용해 구할 수 있음

$$D_{KL}(P||Q) = E_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = E_{X\sim P}\left[-\log \frac{Q(x)}{P(x)}\right]$$

$$\begin{aligned} D_{KL}(P||Q) &= -\sum_x P(x)\log\left(\frac{Q(x)}{P(x)}\right) \\ &= -\sum_x P(x)\{\log Q(x) - \log P(x)\} \\ &= -\sum_x \{P(x)\log Q(x) - P(x)\log P(x)\} \\ &= -\sum_x P(x)\log Q(x) + \sum_x P(x)\log P(x) \\ &= H(P,Q) - H(P) \end{aligned}$$

$$H(P,Q) = H(P) + D_{KL}(P||Q)$$

https://ratsgo.github.io/statistics/2017/09/22/information/

Softmax Function

- k 차원의 벡터에서 i번째 원소를 z_i, i번째 클래스가 정답일 확률을 p_i로 나타낸다고 하였을 때 소프트맥스 함수는 p_i를 다음과 같이 정의됨

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad for\ i = 1, 2, \ldots k$$

- 3차원인 경우

$$softmax(z) = [\frac{e^{z_1}}{\sum_{j=1}^{3} e^{z_j}} \ \frac{e^{z_2}}{\sum_{j=1}^{3} e^{z_j}} \ \frac{e^{z_3}}{\sum_{j=1}^{3} e^{z_j}}] = [p_1, p_2, p_3]$$