

2장

기초 수학

2.1 확률 변수와 확률 분포

2.1.1 확률 변수

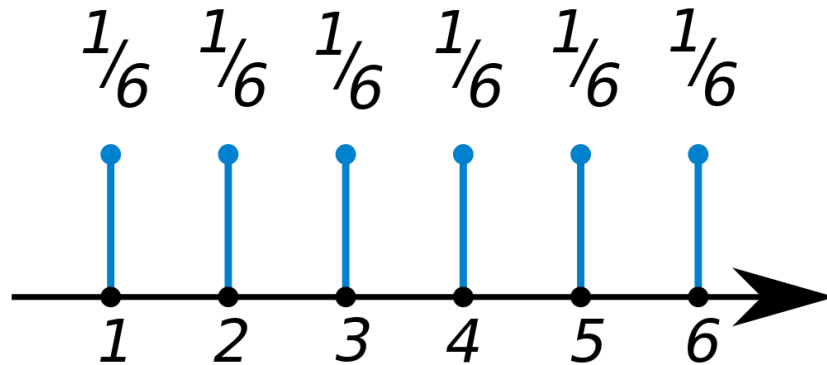
- 확률 변수

- 랜덤 변수 : 확률을 이야기할 때 랜덤하게 발생하는 어떤 사건을 정의
- 주사위를 던졌을 때 (사건 x) 주사위의 숫자가 3이 나왔다면 (값=3)
 - $p(x=3) = 1/6$
 - p : 확률 (probability), 확률 변수가 특정 값을 가질 때 확률값을 반환하는 함수
 - "주사위를 던져 3이 나올 확률은 $1/6$ "
- 확률 변수 x 가 값 x 가 나올 확률 값 p
 - $P(x = x) = P(x) = p$ where $0 \leq p \leq 1$

$$\sum_{i=1}^N P(x = x_i) = \sum_{i=1}^N P(x_i) = 1$$

2.1.1 확률 변수

- 이산 확률 변수
 - 랜덤 변수가 불연속적인 이산(discrete) 값인 경우가 많음 (예 : 주사위)
- 이산 확률 분포
 - 이산 확률 변수 (discrete random variable) 를 갖는 확률 분포
 - 확률 질량 함수(probability mass function)를 통하여 표현 가능
 - 주사위 : $f(x) = 1/6$



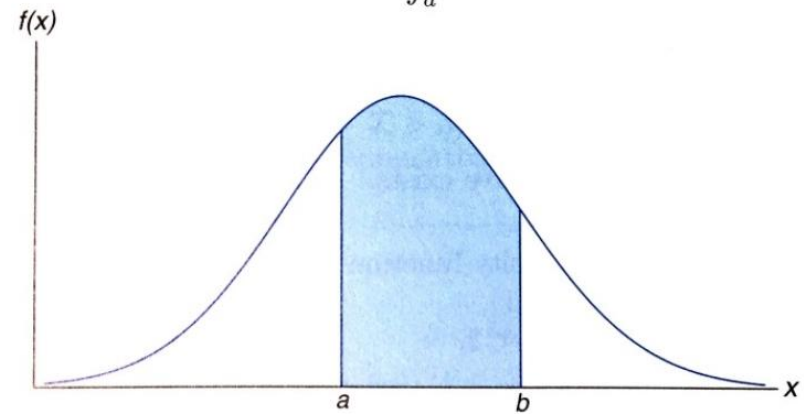
2.1.1 확률 변수

- 이산 확률 분포의 종류
 - 베르누이 분포 (Bernouli distribution)
 - 0과 1 두 개의 값만 보임
 - 이항 분포
 - 멀티눌리 분포 (Multinouli distribution)
 - 여러 개의 이산적인 값을 가짐 (예 : 주사위)
 - 다항 분포

2.1.1 확률 변수

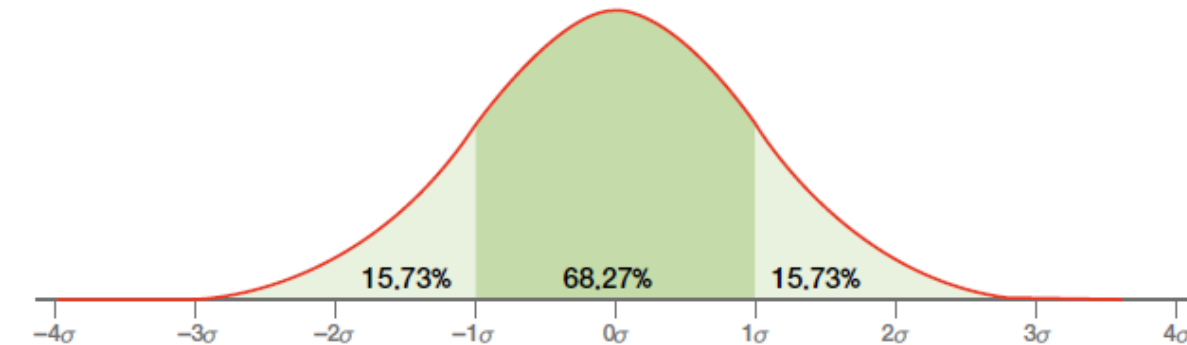
- 연속 확률 변수 (Continuous random variables)
 - 어떤 범위에 속하는 모든 실수 값을 취할 수 있는 확률 변수
- 연속 확률 분포 (Continuous probability distribution)
 - 연속 확률 변수를 갖는 확률 분포
 - 확률값이란 보통 어떠한 구간의 넓이를 의미
 - 확률 밀도 함수를 이용해 분포를 표현할 수 있는 경우를 의미
 - $p(x) \geq 0$
 - (not necessary that $p(x) \leq 1$)
 - $\int_{-\infty}^{\infty} p(x) dx = 1$ ($p(x)$ 를 적분한 값은 항상 1)

$$P(a < X < b) = \int_a^b f(x) dx.$$



2.1.1 확률 변수

- 확률 밀도 함수의 예
 - 정규 분포 (Normal distribution) – 가우시안 분포 (Gaussian distribution)



▶ 가우시안 분포의 확률 밀도 함수

- 구간의 넓이가 확률값
 - 특정 값 x 의 확률값은 구할 수 없음
 - 확률값을 구하기 위해서는 구간 (x_1, x_2) 이 주어져야 함

2.1.2 결합 확률

- 결합 확률 (joint probability)
 - 두 개 이상의 사건이 동시에 일어날 확률
 - 두 개 이상의 확률변수가 필요
 - 주사위 2개 (A와 B)를 던질 때의 확률
 - $P(A, B)$
 - A가 3, B가 2가 나올 확률
 - $P(A=3, B=2)$
 - 독립 (Independence) 관계 : 두 사건이 서로에게 영향을 끼치지 않을 때
 - $P(A, B) = P(A) * P(B)$

2.1.3 조건부 확률

- 조건부 확률 (conditional probability)
 - 두 사건에 대한 확률 분포
 - 독립과 달리, 조건부 확률은 하나의 확률 변수가 주어졌을 때 다른 확률 변수에 대한 확률 분포
 - $P(A|B) = P(A,B)/P(B)$ or $P(A, B) = P(A|B)P(B)$
 - 사건 B가 주어졌을때, 사건 A에 관한 확률 분포
 - $P(A = 3 \mid B = 2)$
 - 주사위 B가 2가 나온 상황에서 주사위 A 값이 3이 나올 확률 값
 - $P(A \mid B = 2)$
 - 주사위 b가 2가 나온 상황에서 주사위 A에서 얻을 수 있는 값의 확률 분포

2.1.3 조건부 확률

- 베이즈 정리 (Bayes theorem)
 - 사건 A와 B에 대한 관계를 반대로 만들 수 있음

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

2.1.4 주변 확률 분포

- 주변 확률 (marginal probability)
 - 개별 사건의 확률이지만, 결합사건들의 합으로 표시될 수 있는 확률

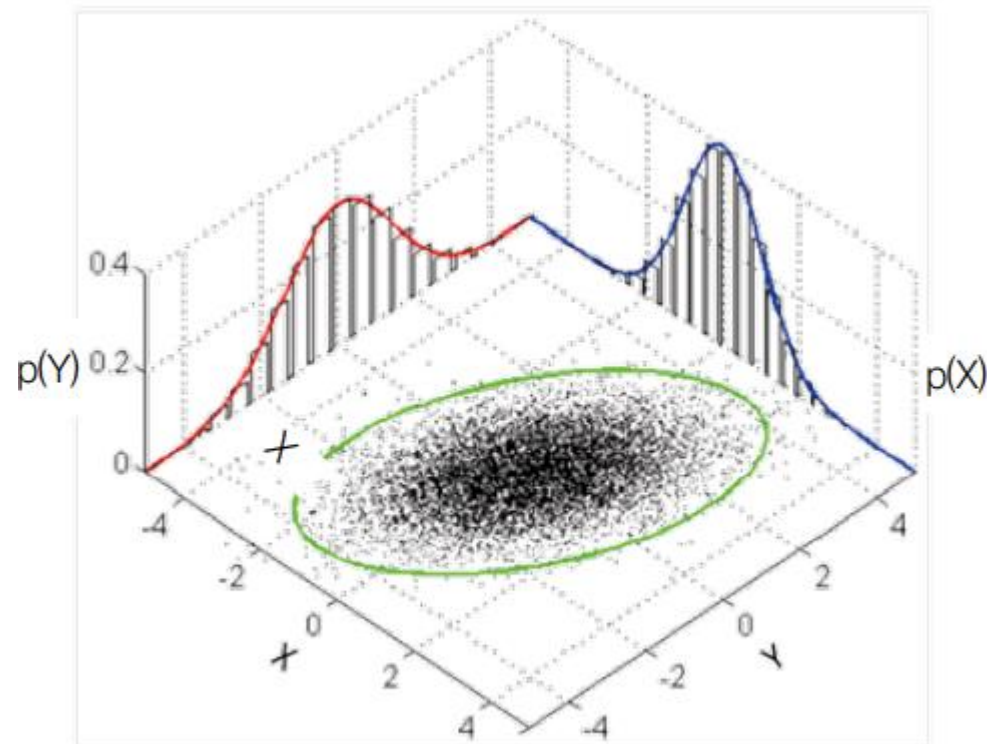
$x \backslash y$	0	1	$P(X=x)$
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$P(Y=y)$	$\frac{1}{2}$	$\frac{1}{2}$	1

주변 확률

(결합 확률)

2.1.4 주변 확률 분포

- 주변 확률 분포 (margin probability distribution)
 - 두 개 이상의 확률 변수의 결합 확률 분포 (joint probability distribution) 가 있을 때, 하나의 확률 변수에 대해서 적분을 수행한 결과를 말한다



▶ 주변 확률 분포의 개념(출처_ <http://bit.ly/2wQPZff>)

2.1.4 주변 확률 분포

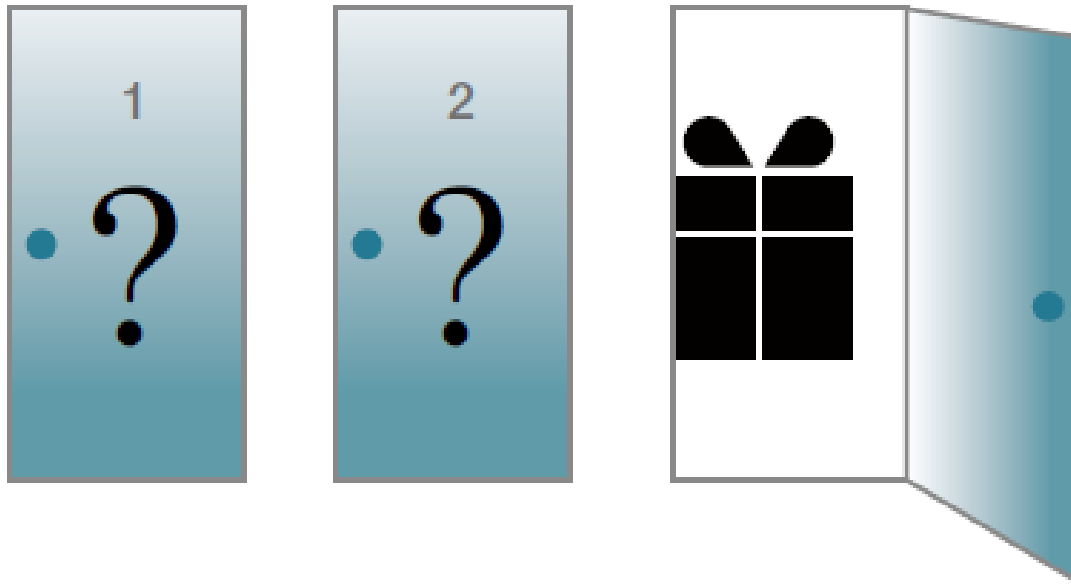
- 결합된 확률 변수들 중 단지 한 변수에 관한 확률분포 만을 고려
 - $P(x) = \sum_{y \in Y} P(x, y) = \sum_{y \in Y} P(x|y)P(y)$ | 이산 확률
 - $p(x) = \int p(x, y) dy = \int p(x|y)p(y) dy$ | 연속 확률

$x \backslash y$	0	1	$P(X=x)$
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$P(Y=y)$	$\frac{1}{2}$	$\frac{1}{2}$	1

주변확률

(결합 확률)

2.2 쉬어가기 : 몬티 홀 문제



▶ 몬티 홀 문제

3개의 문 : 0, 1, 2

A: 시청자가 처음 고른 문

B : 진행자가 열어주는 문

C: 상품이 숨겨진 문

3개의 문 : 0, 1, 2
A: 시청자가 처음 고른 문
B : 진행자가 열어주는 문
C: 상품이 숨겨진 문

2.2 쉬어가기 : 몬티 홀 문제

$$P(C = 2 | A = 0, B = 1) > P(C = 0 | A = 0, B = 1)$$

$$\begin{aligned} P(C = 2 | A = 0, B = 1) &= \frac{P(A = 0, B = 1, C = 2)}{P(A = 0, B = 1)} \\ &= \frac{P(B = 1 | A = 0, C = 2)P(A = 0, C = 2)}{P(A = 0, B = 1)} \\ &= \frac{P(B = 1 | A = 0, C = 2)P(A = 0)P(C = 2)}{P(B = 1 | A = 0)P(A = 0)} \\ &= \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}, \end{aligned}$$

where $P(B = 1 | A = 0) = \frac{1}{2}$, $P(C = 2) = \frac{1}{3}$, and $P(B = 1 | A = 0, C = 2) = 1$.

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

$$P(A, B) = P(A | B)P(B)$$

2.2 쉬어가기 : 몬티 홀 문제

$$P(C = 2 | A = 0, B = 1) > P(C = 0 | A = 0, B = 1)$$

$$\begin{aligned} P(C = 0 | A = 0, B = 1) &= \frac{P(A = 0, B = 1, C = 0)}{P(A = 0, B = 1)} \\ &= \frac{P(B = 1 | A = 0, C = 0)P(A = 0, C = 0)}{P(A = 0, B = 1)} \\ &= \frac{P(B = 1 | A = 0, C = 0)P(A = 0)P(C = 0)}{P(B = 1 | A = 0)P(A = 0)} \\ &= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}, \end{aligned}$$

$$\text{where } P(B = 1 | A = 0, C = 0) = \frac{1}{2}$$

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

$$P(A, B) = P(A | B)P(B)$$

2.3 기댓값과 샘플링

2.3.1 기댓값

- 기댓값
 - 보상과 그 보상을 받을 확률을 곱한 값의 총합 (보상에 대한 가중평균)
 - 이산 확률 변수

$$\text{expected reward from dice} = \sum_{x=1}^6 P(x=x) \times \text{reward}(x)$$

$$\text{where } P(x) = \frac{1}{6}, \forall x \text{ and } \text{reward}(x) = x \quad \frac{1}{6} \times (1+2+3+4+5+6) = 3.5$$

$$\mathbb{E}_{x \sim P(x)} [\text{reward}(x)] = \sum_{x=1}^6 P(x=x) \times \text{reward}(x) = 3.5$$

- 연속 확률 변수

$$\mathbb{E}_{x \sim p} [\text{reward}(x)] = \int p(x) \cdot \text{reward}(x) dx$$

2.3.2 몬테카를로 샘플링

- 샘플링
 - (통계학) 임의의 확률 분포 $p(x)$ 로부터 표본을 추출하는 작업
- 몬테카를로
 - 난수를 이용하여 함수의 값을 확률적으로 계산하는 알고리즘을 부르는 용어
- 몬테카를로 샘플링
 - 랜덤 성질을 이용하여 임의의 함수 적분을 근사하는 방법

2.3.2 몬테카를로 샘플링



▶ 한반도의 넓이를 근사하고 싶다면?

임의로 점을 흩뿌려, 한반도 안과 밖의 점의 비율로 전체 면적 계산

$$\mathbb{E}_{\mathbf{x} \sim P} [f(\mathbf{x})] = \sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}) \cdot f(\mathbf{x}) \approx \frac{1}{K} \sum_{i=1}^K f(x_i) \text{ where } x_i \sim P(\mathbf{x})$$

$$\mathbb{E}_{\mathbf{x} \sim p} [f(\mathbf{x})] = \int p(\mathbf{x}) \cdot f(\mathbf{x}) d\mathbf{x} \approx \frac{1}{K} \sum_{i=1}^K f(x_i) \text{ where } x_i \sim p(\mathbf{x})$$

2.4 MLE

- 일반화 (generalization)
 - 머신 러닝의 목표
 - 미지의 데이터에 대해 좋은 예측을 하는 것
 - 데이터를 잘 설명, 또는 주어진 데이터로부터 결괏값을 잘 예측
 - 좋은 일반화
 - 알고자 하는 실제 확률 분포로부터 데이터를 수집(샘플링)
 - 수집된 데이터를 가장 잘 설명하는 확률 분포 모델 추정
 - 실제 확률 분포 근사

2.4 MLE

- 게임
 - 압정의 납작한 면이 바닥에 떨어지면 +50원, 반대 -20원
 - 샘플링
 - 동일한 압정을 구매하여 100번 실험
 - 납작한 면이 바닥으로 30번 떨어짐

$$\begin{aligned}\mathbb{E}_{x \sim P}[\text{reward}(x)] &= P(x = \text{flat}) \times 50 + P(x = \text{sharp}) \times (-20) \\ &\approx \frac{30}{100} \times 50 - \left(1 - \frac{30}{100}\right) \times 20 = 15 - 14 = 1\end{aligned}$$

2.4 MLE

- 이항분포
 - 압정을 던지는 사건은 베르누이 분포를 따름

$$K \sim \mathcal{B}(n, \theta)$$

- Θ : 이항분포의 파라미터 (알려진 확률)
- N번 압정을 던져 납작한 면이 k번 바닥으로 떨어질 확률

$$\begin{aligned} P(K = k) &= \binom{n}{k} \theta^k (1 - \theta)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \cdot \theta^k (1 - \theta)^{n-k} \end{aligned}$$

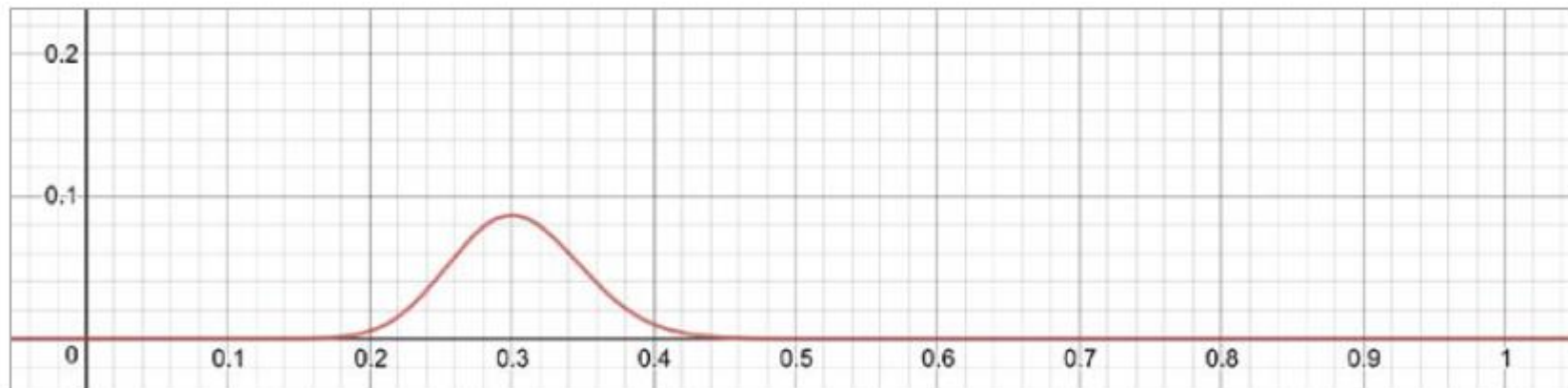
2.4 MLE

- 압정을 실제로 던지는 경우
 - $n = 100, k = 30$
 - 확률을 가장 잘 설명하는 함수 (가능도 함수)

$$J(\theta) = \frac{100!}{30!(100-30)!} \cdot \theta^{30} (1-\theta)^{100-30}$$

$$J(\theta) = P(n=100, k=30 | \theta)$$

- 가능도 (likelihood)
 - 함수에 θ 를 넣어 얻는 결과값



▶ 가능도 함수 곡선: x축은 θ , y축은 가능도를 나타냅니다.

2.4 MLE

- 최대 가능도 추정 (Maximum Likelihood Estimation: MLE)
 - 주어진 데이터를 잘 설명하기 위해 가능도를 최대화하도록 θ 를 추정
 - 가능도
 - 주어진 데이터를 설명하는 확률 분포 파라미터에 대한 함수

$$P(\mathbf{x} = x_{1:n}; \theta)$$

- 확률값 자체 (이산 확률 분포) 또는 확률 밀도 값 (연속 확률 분포)이 가능도를 표현

$$P(x_1, x_2, \dots, x_n | \theta) = P(x_1; \theta) P(x_2; \theta) \cdots P(x_n; \theta) = \prod_{i=1}^n P(x_i; \theta)$$

$$\log P(x_1, x_2, \dots, x_n | \theta) = \sum_{i=1}^n \log P(x_i; \theta)$$

- 언더플로, 연산의 빠르기 등

2.4 MLE

- 가우시안 분포에서 지수를 제거

$$J(\theta) = \log \mathcal{N}(x | \mu, \sigma^2) = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}$$

$$\text{where } \mathcal{N}(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

- 로그 가능도 (log-likelihood) 최대화 = 음의 로그 가능도 (negative log-likelihood: NLL) 최소화

2.4.1 확률 분포 함수로서의 신경망

- 확률 분포 함수로서의 신경망
 - 신경망 또한 확률 분포 함수
 - MNIST 분류 문제
 - 이산 확률 변수를 다루는 멀티눌리 분포 문제
 - 소프트맥스 연산 : 클래스별 확률값에 대한 분포를 반환
 - y 도메인 : 원핫 벡터로 표현
 - $\theta \leftarrow \theta - \lambda \nabla_{\theta} J(\theta)$
 - θ is network weight parameter, and $J(\theta)$ is negative log likelihood
 - MLE 수행
 - 신경망 가중치 파라미터 θ 가 훈련 데이터를 잘 설명하도록
 - 경사 하강법 (gradient descent) 을 통해 MLE를 수행하여 학습

2.5 정보 이론

2.5.1 정보량

- 정보 이론 (information theory)
 - 데이터를 정량화하기 위한 응용 수학의 분야 – 신경망과 밀접한 연관
- 정보량 (information content)
 - 불확실성 또는 놀람의 정도
 - 일상적인 정보, 즉 발생할 확률이 높은 정보 -> 정보의 가치 낮음
 - 아주 희귀한 정보, 즉 확률이 작은 정보 -> 정보의 가치 높음

정보량	정보
매우 낮음	내일 아침에는 해가 동쪽에서 뜹니다.
매우 높음	내일 아침에는 해가 서쪽에서 뜹니다.
매우 낮음	올 여름 평균 기온은 섭씨 28도로 예상됩니다.
매우 높음	올 여름 평균 기온은 영하 10도로 예상됩니다.

▶ 정보량이 높을수록 특정 사건의 발생 확률은 낮아짐

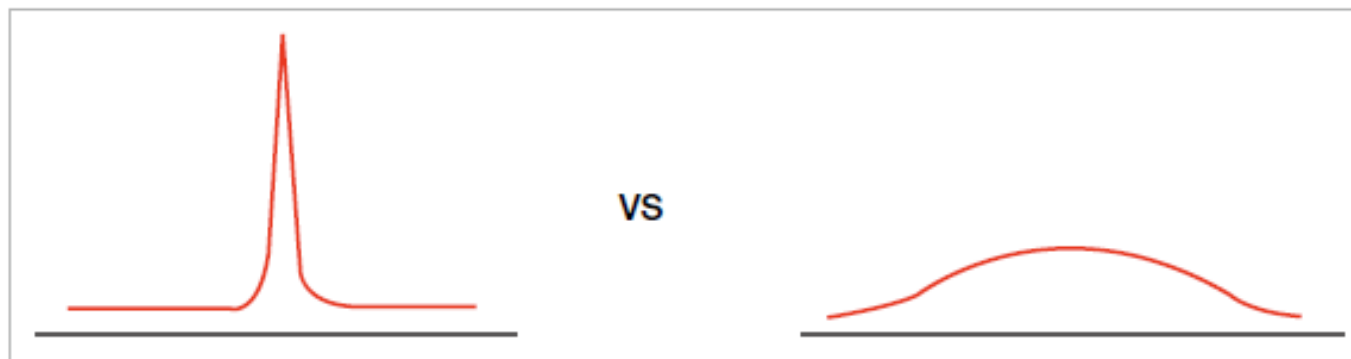
2.5.1 정보량

- 일어날 확률이 낮은 일에 대한 문장일수록 많은 정보를 갖고 있음
 - 확률이 낮은 사건에 대해 서술한 문장이 맞을수록 정보의 가치가 올라감

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

2.5.2 엔트로피



- 엔트로피 (entropy)
 - 정보량의 평균 (기댓값)
 - $H(P) = -E_{X \sim P(x)}[\log P(x)] = -\sum_{x \in X} P(x) \log P(x)$
 - 엔트로피는 분포의 대략적인 모양이 얼마나 퍼져있는지, 뾰족한지를 가늠해볼 수 있는 척도
 - 뾰족한 확률 분포일수록 특정 값에 대한 확률이 높다



▶ 뾰족한(sharp) 분포와 퍼진(flat) 분포

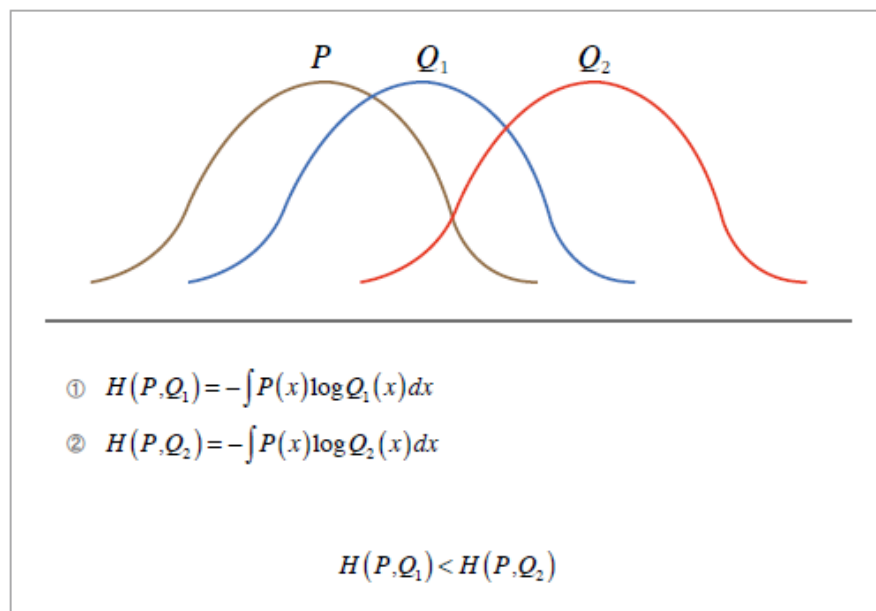
2.5.2 엔트로피

- 엔트로피

		$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$	
		Calculation	Entropy
50%	50%	$-(0.5 * \log_2 0.5 + 0.5 * \log_2 0.5) = 1$	1
100%	0%	$-(1.0 * \log_2 1 + 0.0 * \log_2 0) = 0$	0
90%	10%	$-(0.9 * \log_2 0.9 + 0.1 * \log_2 0.1) = 0.47$	0.47

2.5.2 엔트로피

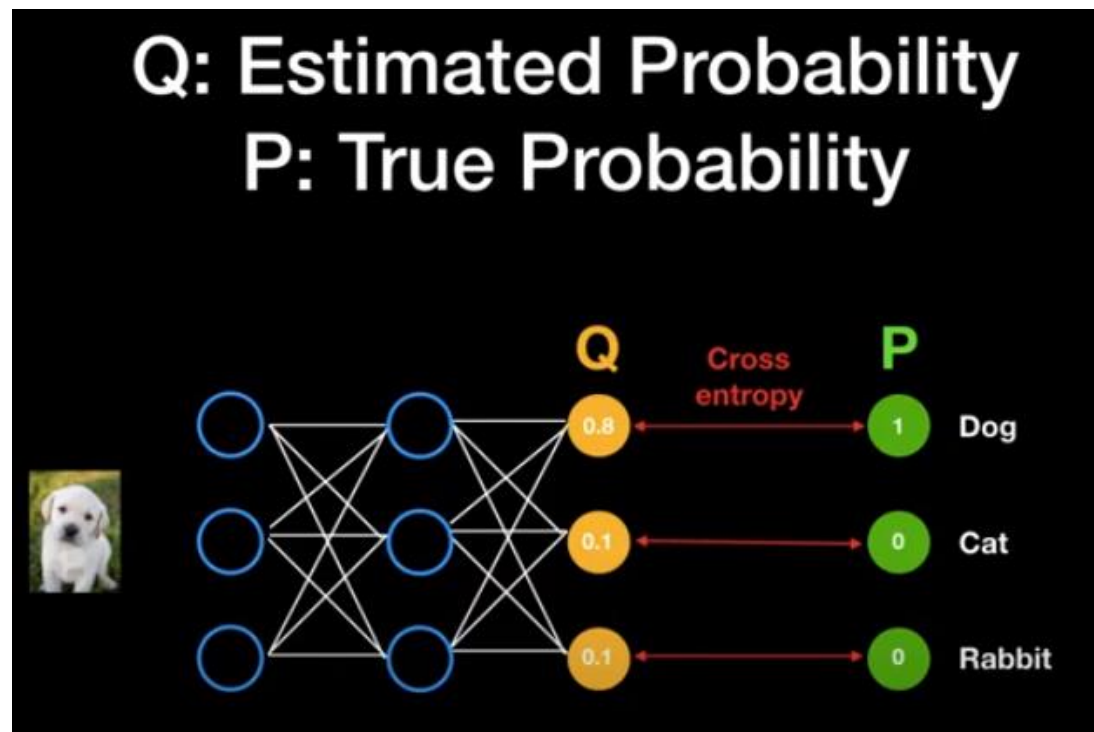
- 교차 엔트로피 (cross entropy)
 - 틀릴 수 있는 정보를 가지고 구한 최적의 엔트로피 값, 즉 불확실성 정보의 양
 - 다른 분포 P 를 사용하여 대상 분포 Q 의 엔트로피를 측정합니다.
 - $H(P, Q) = -E_{X \sim P(x)}[\log Q(x)] = -\sum_{x \in X} P(x) \log Q(x)$



▶ 교차 엔트로피의 직관적인 표현

2.5.2 엔트로피

- 교차 엔트로피
 - 이산 확률 분포
 - $\mathcal{L}(\theta) = H(P, P_\theta)$
 - $\theta \leftarrow \theta - \nabla_{\theta} \mathcal{L}(\theta)$ | 최소가 되도록 경사하강법을 수행
 - 연속 확률 분포
 - 샘플에 대한 확률값을 구할 수 없다.
 - 평균제곱오차 (Mean Square Error) 사용



2.5.2 엔트로피

- 쿨백-라이블러 발산 (Kullback-Leibler divergence: KLD)
 - 두 분포 사이의 괴리를 보여줌
 - 대칭의 개념이 아님
 - 분포 P와 분포 Q의 위치에 따라서 KLD 값이 달라짐
 - '거리' 라고 표현하지 않음
 - 교차 엔트로피를 손실 함수로 활용 = KLD를 손실 함수로 활용

$$\begin{aligned} KL(P \parallel Q) &= -\mathbb{E}_{\mathbf{x} \sim P(\mathbf{x})} \left[\log \frac{Q(\mathbf{x})}{P(\mathbf{x})} \right] \\ &= -\sum_{x \in \mathcal{X}} P(x) \log \frac{Q(x)}{P(x)} \\ &= -\sum_{x \in \mathcal{X}} P(x) \log Q(x) - \sum_{x \in \mathcal{X}} P(x) \log P(x) \\ &= H(P, Q) - H(P) \end{aligned}$$

$$\begin{aligned} \mathcal{L}(\theta) &= KL(P \parallel P_\theta) \\ &= H(P, P_\theta) - H(P) \\ \nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta KL(P \parallel P_\theta) \\ &= \nabla_\theta H(P, P_\theta) - \nabla_\theta H(P) \\ &= \nabla_\theta H(P, P_\theta) \end{aligned}$$