기계학습 (Machine Learning)

L08
- Probability and Statistics in ML
- Objective Functions

한밭대학교

정보통신공학과

최 해 철

국립 한밭대학교

VISUAL MEDIA LAB.
HANBAT NATIONAL UNIVERSITY

ToC

Probability and Statistics in ML

오일석, 기계학습, 2.2 확률과 통계

확률 기초

◆ ______________________ 확률변수, 랜덤변수

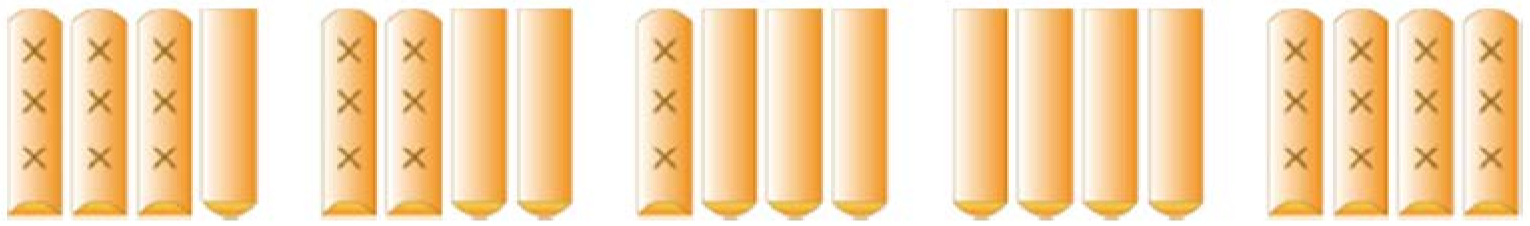


그림 2-13 윷을 던졌을 때 나올 수 있는 다섯 가지 경우(왼쪽부터 도, 개, 걸, 윷, 모)

- 다섯 가지 경우 중 한 값을 갖는 random variable x
- x 의 __________은 {도, 개, 걸, 윷, 모}

확률 기초

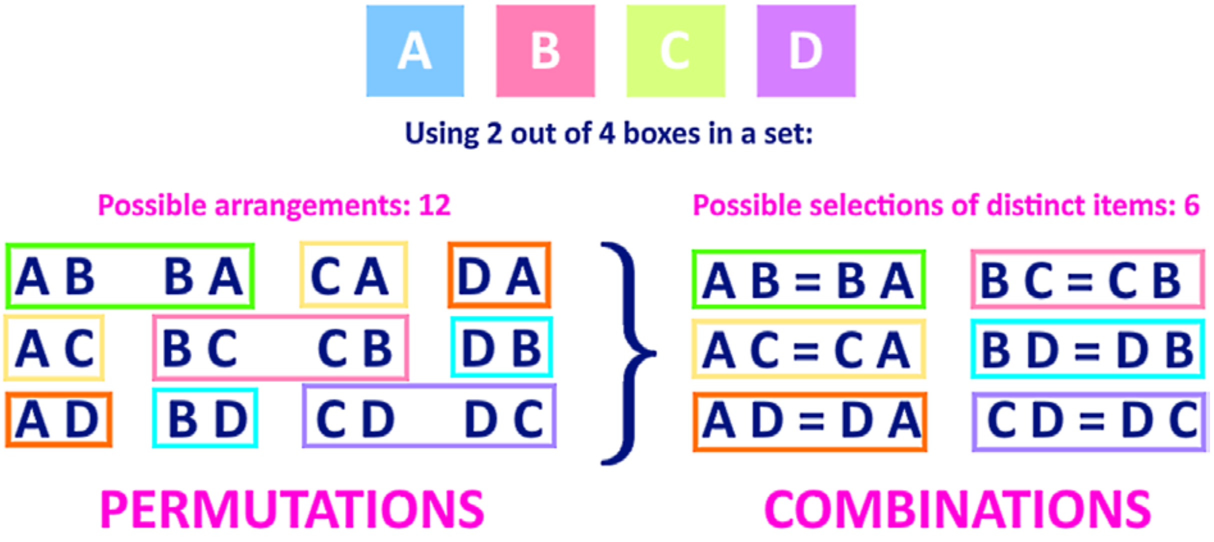
◆ Permutation순열과 combination조합

- Permutation: 순서가 있는 경우의 수

$$_{n}P_{r} = \frac{n!}{(n-r)!}$$

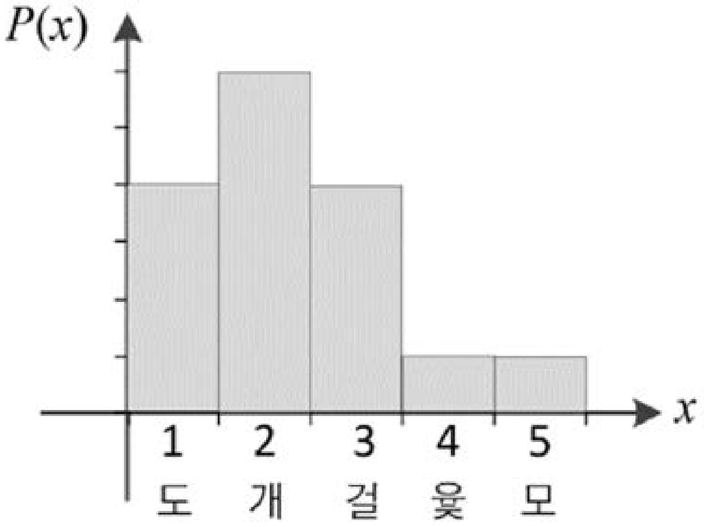
- Combination: 순서가 없는 경우의 수

$$_{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

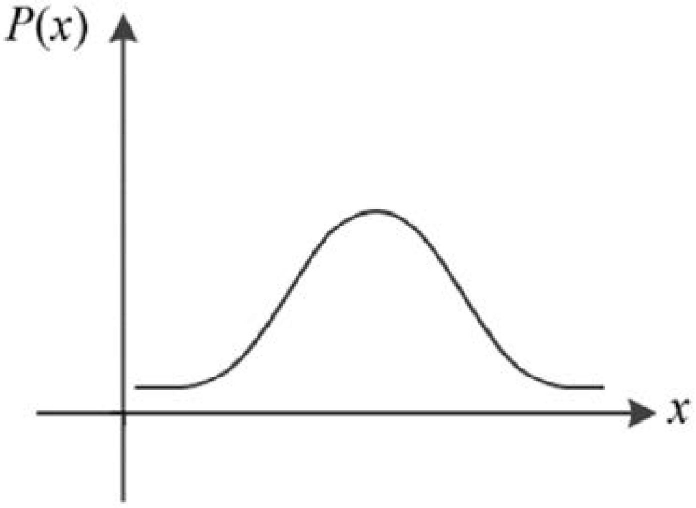


◆ Probability distribution 확률분포

$$P(x = \text{도}) = \frac{4}{16}, P(x = \text{개}) = \frac{6}{16}, P(x = \text{걸}) = \frac{4}{16}, P(x = \text{윷}) = \frac{1}{16}, P(x = \text{모}) = \frac{1}{16}$$



(a) 이산인 경우의 확률질량함수



(b) 연속인 경우의 확률밀도함수

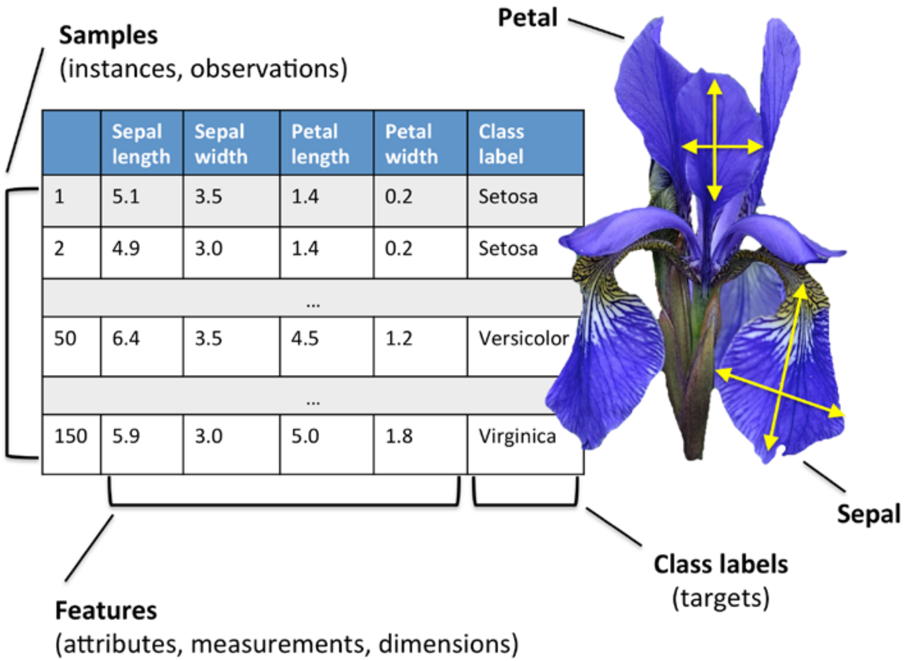
그림 2-14 확률분포

확률 기초

◆ ______________ 확률벡터

- 예) Iris에서 확률벡터 $\mathbf{x}$는

$$\mathbf{x} = (x_1, x_2, x_3, x_4)^{\mathrm{T}} = (\text{꽃받침 길이}, \text{꽃받침 너비}_1, \text{꽃잎 길이}, \text{꽃잎 너비})^{\mathrm{T}}$$



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

확률 기초

◆ 간단한 확률실험 장치

- 주머니에서 번호를 뽑은 다음, 번호에 따라 해당 병에서 공을 뽑고 색을 관찰함
- 번호를 y, 공의 색을 x라는 확률변수로 표현하면 정의역은 $y \in \{①,②,③\}$, $x \in \{$파랑, 하양$\}$

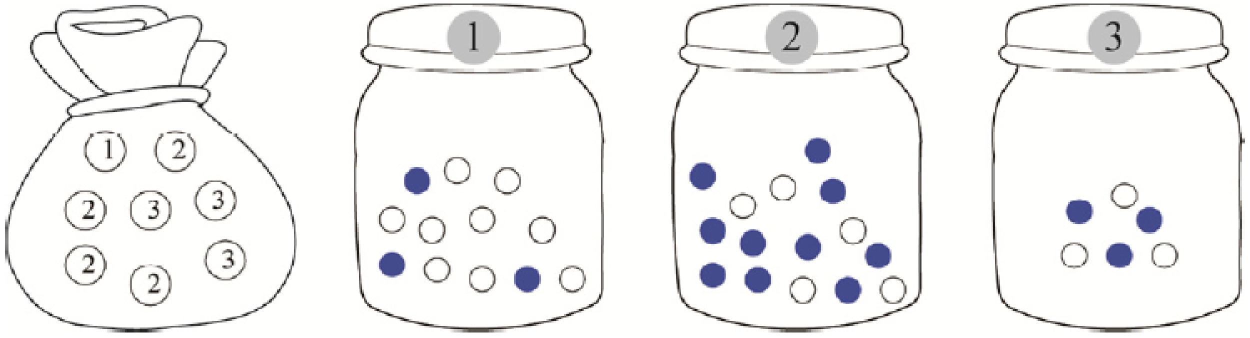


그림 2-15 확률 실험

확률 기초

◆ **Prior probability** 사전 확률

- ①번 카드를 뽑을 확률: $P(y = ①) = P(①) = 1/8$

◆ **Joint probability** 결합 확률

- 카드는 ①번, 공은 하양일 확률: $P(y = ①, x =$ 하양 $) = P(①,$ 하양$)$

◆ **Conditional probability** 조건부 확률

- 카드는 ①번을 뽑은 경우, 공이 하양일 확률 : $P(x =$ 하양 $| y = ①) = P($하양 $| ①) =9/12$

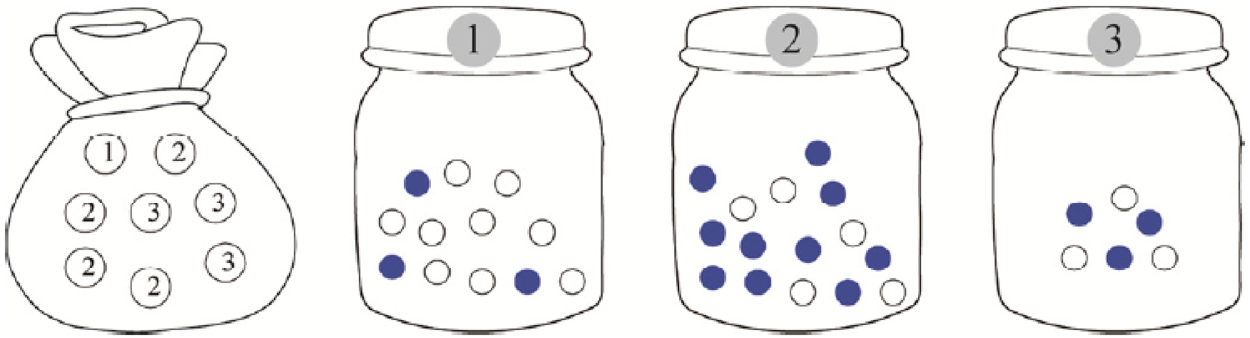


그림 2-15 확률 실험

확률 기초

◆ 곱규칙

$$\text{곱규칙:}\quad P(x,y) = P(x|y)P(y) = P(y|x)P(x) \qquad (2.23)$$

$$P(y=①, x=\text{하양}) = P(x=\text{하양}|y=①)P(y=①) = \frac{9}{12}\frac{1}{8} = \frac{3}{32}$$

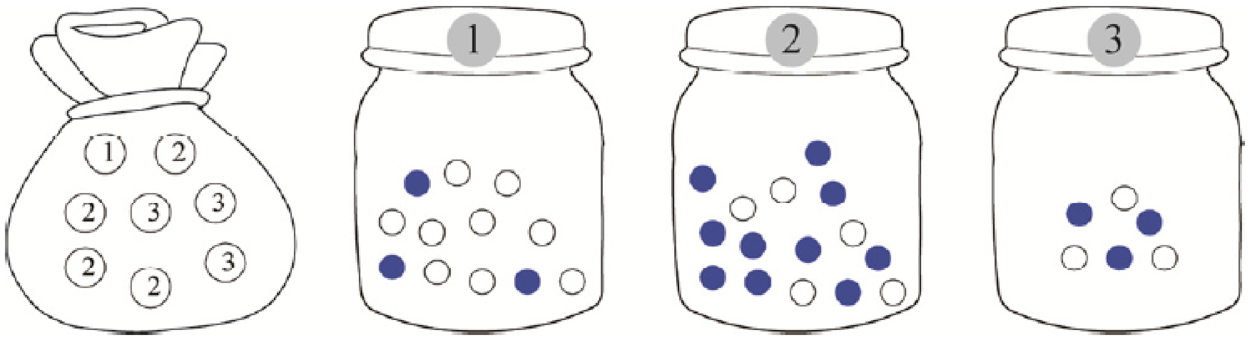


그림 2-15 확률 실험

확률 기초

◆ 합규칙

$$\text{합 규칙: } P(x) = \sum_{y} P(y,x) = \sum_{y} P(x|y)P(y) \tag{2.24}$$

● 하얀 공이 뽑힐 확률

$$P(\text{하양}) = P(\text{하양}|①)P(①) + P(\text{하양}|②)P(②) + P(\text{하양}|③)P(③)$$
$$= \frac{9}{12}\frac{1}{8} + \frac{5}{15}\frac{4}{8} + \frac{3}{6}\frac{3}{8} = \frac{43}{96}$$

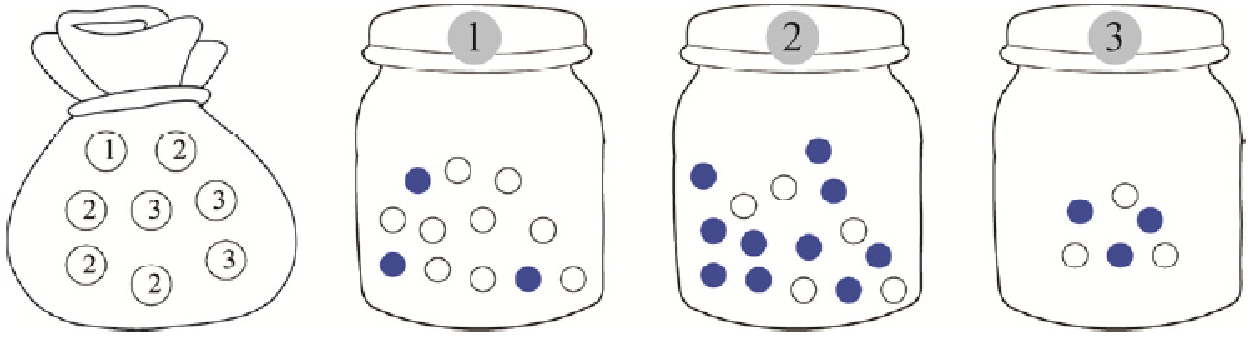


그림 2-15 확률 실험

베이즈 정리와 기계 학습

◆ 베이즈 정리Bayes formula

$$P(y,x) = P(x|y)P(y) = P(x,y) = P(y|x)P(x)$$

$$\longrightarrow \boxed{}$$

베이즈 정리와 기계 학습

◆ 질문: "하얀 공이 나왔다는 사실만 알고 어느 병에서 나왔는지 모르는데, 어느 병인지 추정하라."

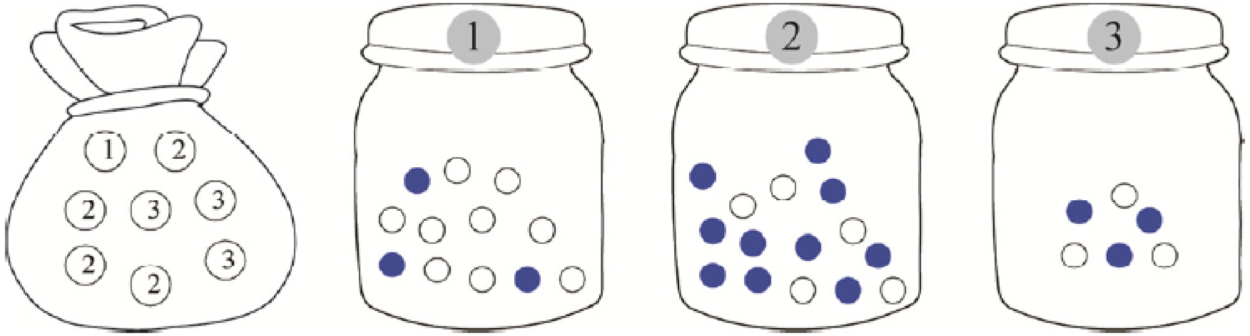


그림 2-15 확률 실험

베이즈 정리와 기계 학습

- 베이즈 정리를 적용하면,

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x = \text{하양}) = \underset{y}{\operatorname{argmax}} \frac{P(x = \text{하양}|y)P(y)}{P(x = \text{하양})}$$

$$P(①|\text{하양}) = \frac{P(\text{하양}|①)\mathrm{P}(①)}{P(\text{하양})} = \frac{\frac{9}{12}\frac{1}{8}}{\frac{43}{96}} = \frac{9}{43}$$

$$P(②|\text{하양}) = \frac{P(\text{하양}|②)\mathrm{P}(②)}{P(\text{하양})} = \frac{\frac{5}{15}\frac{4}{8}}{\frac{43}{96}} = \frac{16}{43}$$

$$P(③|\text{하양}) = \frac{P(\text{하양}|③)\mathrm{P}(③)}{P(\text{하양})} = \frac{\frac{3}{6}\frac{3}{8}}{\frac{43}{96}} = \frac{18}{43}$$

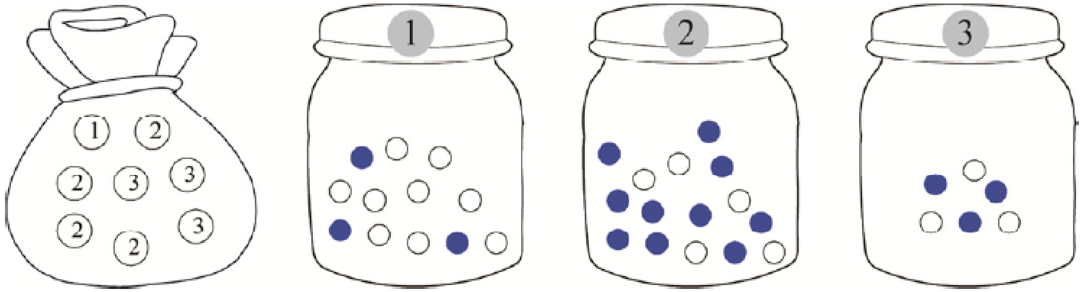


그림 2-15 확률 실험

베이즈 정리와 기계 학습

◆ 베이즈 정리의 해석

$$\overbrace{P(y|x)} = \frac{\overbrace{P(x|y)}\,\overbrace{P(y)}}{P(x)}$$

사후확률(a posterior probability)
우도/가능도(Likelihood, a conditional probability)
사전확률(a prior probability)

정의역: $y \in \{①,②,③\}$, $x \in \{$파랑, 하양$\}$

y에 의해서 x가 발생
($y \rightarrow x$)

최종사건

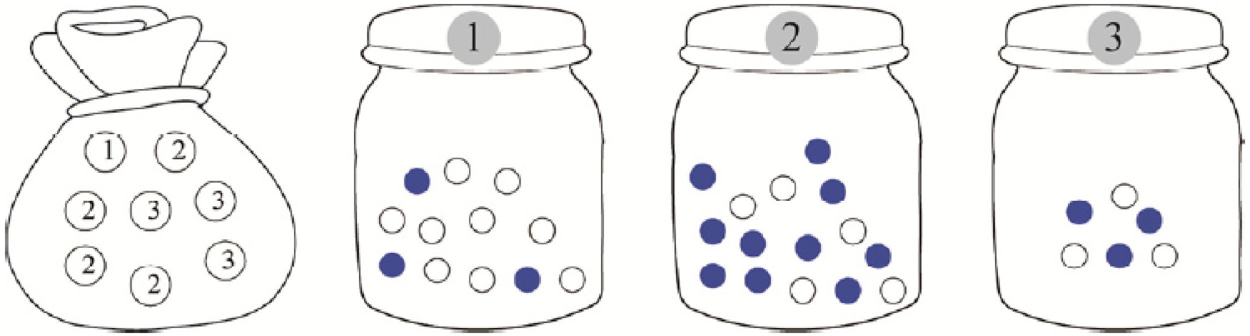


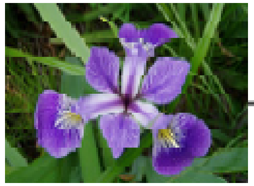
그림 2-15 확률 실험

베이즈 정리와 기계 학습

◆ 기계 학습에 적용

- 예) Iris 데이터 분류 문제
  - 특징 벡터 $\mathbf{x}$, 부류 $y \in$ {setosa, versicolor, virginica}
  - 분류 문제를 argmax로 표현하면 식 (2.29)

Maximum A Posteriori (MAP) :



$\xrightarrow{\text{특징추출}}$ $\mathbf{x} = (7.0, 3.2, 4.7, 1.4)^{\mathrm{T}}$ $\xrightarrow{\text{사후확률 추정}}$

$P(\text{setosa}|\mathbf{x}) = 0.18$

$P(\text{versicolor}|\mathbf{x}) = 0.72$

$P(\text{virginica}|\mathbf{x}) = 0.10$

$\xrightarrow{\text{argmax}}$ versicolor

그림 2-16 붓꽃의 부류 예측 과정

베이즈 정리와 기계 학습

◆ 기계 학습에 적용 (cont'd)

- 사후확률 $P(y|\mathbf{x})$를 직접 추정하는 일은 아주 단순한 경우를 빼고 불가능

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|\mathbf{x})$$

- 따라서 베이즈 정리를 이용하여 추정함

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|\mathbf{x}) = \underset{y}{\operatorname{argmax}} \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

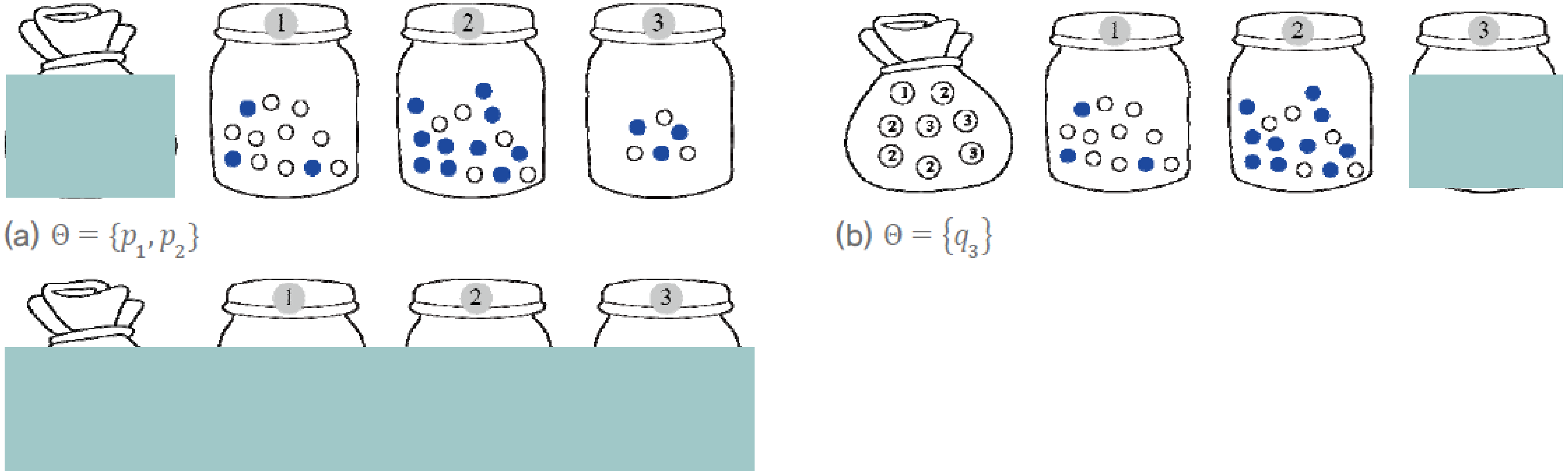
  - 사전확률은 다음 식으로 추정 사전확률: $P(y = c_i) = \frac{n_i}{n}$
  - $P(\mathbf{x})$는 고려하지 않아도 된다.
  - Likelihood는 밀도 추정density estimation 기법으로 추정 (추후 다룸, 오일석, 기계학습 6.4절)

최대 우도(Maximum Likelihood)

최대우도

- 딥러닝 목적함수에 likelihood를 이용 (오일석, 기계학습, 5.1절)
- 매개변수 Θ를 모르는 상황에서 Θ를 추정하는 문제

데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}



(a) $\Theta = \{p_1, p_2\}$

(b) $\Theta = \{q_3\}$

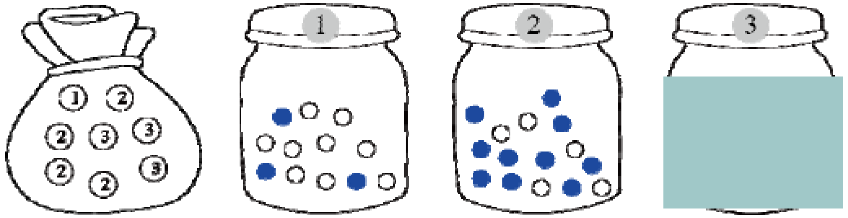
(c) $\Theta = \{p_1, p_2, q_1, q_2, q_3\}$

그림 2-17 매개변수가 감추어진 여러 가지 상황

최대 우도(Maximum Likelihood)

"데이터 $\mathbb{X}$가 주어졌을 때, $\mathbb{X}$를 발생시켰을 가능성을 최대로 하는 매개변수 $\Theta = \{q_3\}$의 값을 찾아라."

데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}



(b) $\Theta = \{q_3\}$

◆ 수식으로 표현

$$\hat{q}_3 = \underset{q_3}{\operatorname{argmax}} P(\mathbb{X}|q_3)$$

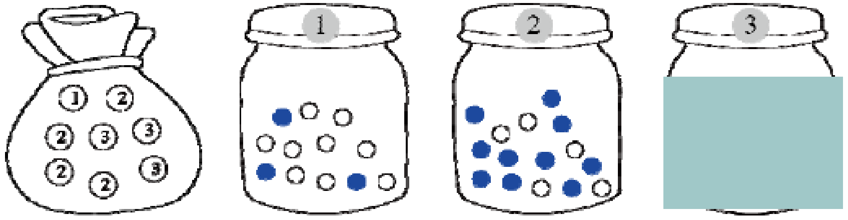
◆ 일반화

최대 우도 추정: $$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} P(\mathbb{X}|\Theta)$$

최대 우도(Maximum Likelihood)

"데이터 𝕏가 주어졌을 때, 𝕏를 발생시켰을 가능성을 최대로 하는 매개변수 $\Theta = \{q_3\}$의 값을 찾아라."

데이터집합 𝕏={●○○●○●○○●●○○}



(b) $\Theta = \{q_3\}$

◆ $\boldsymbol{X} = \{x_1, x_2, x_3, \ldots, x_n\}$이 동일분포iid(independent and identically distributed)인 경우

$$P(\boldsymbol{X}|\theta) = P(x_1, x_2, x_3, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

◆ 수치 문제를 피하기 위해 로그 표현으로 바꾸면

최대 로그우도 추정: $\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log P(\mathbb{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}}$

평균과 분산

◆ 데이터의 요약 정보로서 평균과 분산

$$\left.\begin{aligned} \text{평균} \quad \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \text{분산} \quad \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}\right\}$$

◆ 평균 벡터와 공분산 행렬

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

평균과 분산

◆ 평균 벡터와 공분산 행렬 예제

예제 2-7

Iris 데이터베이스의 샘플 중 8개만 가지고 공분산 행렬을 계산하자.

$$\mathbb{X} = \left\{\mathbf{x}_1 = \begin{pmatrix}5.1\\3.5\\1.4\\0.2\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}4.9\\3.0\\1.4\\0.2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}4.7\\3.2\\1.3\\0.2\end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix}4.6\\3.1\\1.5\\0.2\end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix}5.0\\3.6\\1.4\\0.2\end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix}5.4\\3.9\\1.7\\0.4\end{pmatrix}, \mathbf{x}_7 = \begin{pmatrix}4.6\\3.4\\1.4\\0.3\end{pmatrix}, \mathbf{x}_8 = \begin{pmatrix}5.0\\3.4\\1.5\\0.2\end{pmatrix}\right\}$$

먼저 평균벡터를 구하면 $\boldsymbol{\mu} = (4.9125,\ 3.3875,\ 1.45,\ 0.2375)^{\mathrm{T}}$이다. 첫 번째 샘플 $\mathbf{x}_1$을 식 (2.39)에 적용하면 다음과 같다.

$$(\mathbf{x}_1 - \boldsymbol{\mu})(\mathbf{x}_1 - \boldsymbol{\mu})^{\mathrm{T}} = \begin{pmatrix}0.1875\\0.1125\\-0.05\\-0.0375\end{pmatrix}\begin{pmatrix}0.1875 & 0.1125 & -0.05 & -0.0375\end{pmatrix}$$

$$= \begin{pmatrix}0.0325 & 0.0211 & -0.0094 & -0.0070\\0.0211 & 0.0127 & -0.0056 & -0.0042\\-0.0094 & -0.0056 & 0.0025 & 0.0019\\-0.0070 & -0.0042 & 0.0019 & 0.0014\end{pmatrix}$$

나머지 7개 샘플도 같은 계산을 한 다음, 결과를 모두 더하고 8로 나누면 다음과 같은 공분산 행렬을 얻는다.

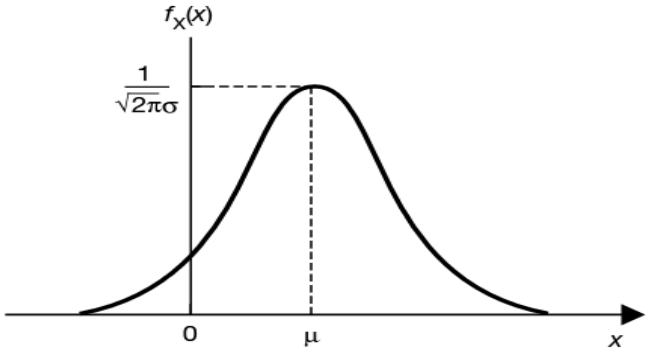
$$\boldsymbol{\Sigma} = \begin{pmatrix}0.0661 & 0.0527 & 0.0181 & 0.0083\\0.0527 & 0.0736 & 0.0181 & 0.0130\\0.0181 & 0.0181 & 0.0125 & 0.0056\\0.0083 & 0.0130 & 0.0056 & 0.0048\end{pmatrix}$$

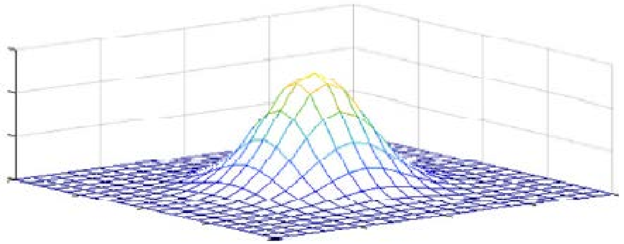
유용한 확률분포

◆ 가우시안 분포 Gaussian distribution

● 평균 μ와 분산 σ^2으로 정의

$$N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$





(b) 2차원

● 다차원 가우시안 분포: 평균벡터 $\boldsymbol{\mu}$와 공분산행렬 $\boldsymbol{\Sigma}$로 정의

$$N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

유용한 확률분포

◆ 베르누이 분포 Bernoulli Distribution

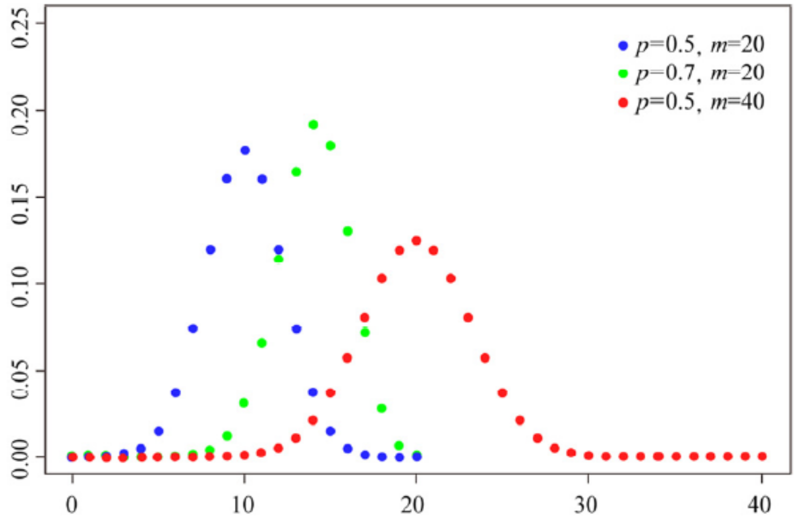
- 성공(x=1) 확률 p이고 실패(x=0) 확률이 1-p인 분포

$$Ber(x;p) = p^x(1-p)^{1-x} = \begin{cases} p, & x = 1\text{일 때} \\ 1-p, & x = 0\text{일 때} \end{cases}$$

◆ 이항 분포 Binomial distribution

- 성공 확률이 p인 베르누이 실험을 m번 수행할 때 성공할 횟수의 확률분포

$$B(x;m,p) = C_m^x p^x(1-p)^{m-x} = \frac{m!}{x!\,(m-x)!}p^x(1-p)^{m-x}$$



Information Theory

Self Information

- 메시지가 지닌 정보를 수량화할 수 있나?
  - "고비 사막에 눈이 왔다"와 "대관령에 눈이 왔다"라는 두 메시지 중 어느 것이 더 많은 정보를 가지나?
  - 정보이론의 기본 원리 → 확률이 작을수록 많은 정보

- 자기 정보self information
  - 특정 사건(메시지) e_i의 정보량 (단위: 비트 또는 나츠)

$$h(e_i) = \boxed{} \text{ 또는 } h(e_i) = -\log_e P(e_i) \qquad (2.44)$$

Entropy

◆ 엔트로피Entropy

- 확률변수 x의 불확실성uncertainty을 나타냄 (확률 분포의 무질서도 or 불확실성)
- ______________________

$$\text{이산 확률분포} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i) \quad \text{또는} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_e P(e_i)$$

$$\text{연속 확률분포} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x) \quad \text{또는} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_e P(x)$$

예제

◆ 자기 정보와 엔트로피 예제

예제 2-8

윷을 나타내는 확률변수를 x라 할 때 x의 엔트로피는 다음과 같다.

$$H(x) = -\left(\frac{4}{16}\log_2\frac{4}{16} + \frac{6}{16}\log_2\frac{6}{16} + \frac{4}{16}\log_2\frac{4}{16} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16}\right) = 2.0306\text{비트}$$

주사위는 눈이 6개인데 모두 1/6이라는 균일한 확률을 가진다. 이 경우 엔트로피를 계산하면 다음과 같다.

$$H(x) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 2.585\text{ 비트}$$

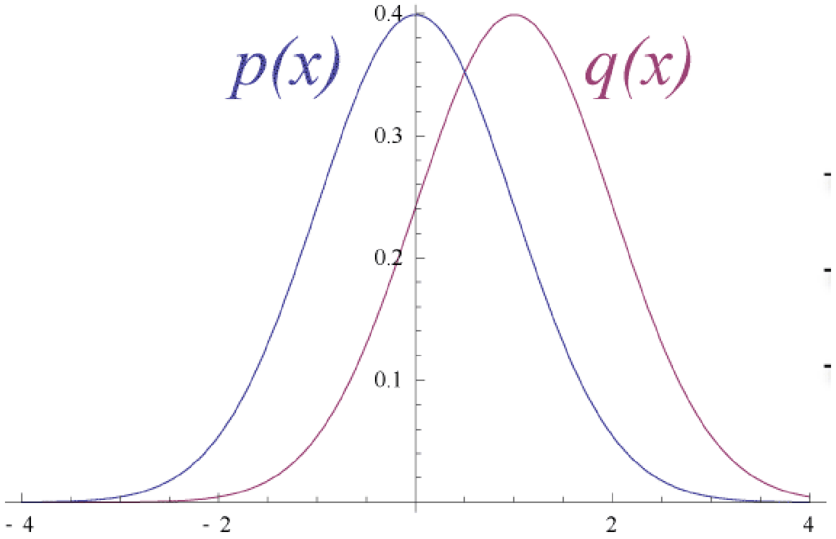
● 주사위가 윷보다 엔트로피가 높은 이유는?

Cross Entorpy

◆ Cross Entropy (CE) 교차 엔트로피

- 두 확률분포에서 P에 대한 Q의 CE

$$H(P,Q) = -\sum_{x} \boxed{\qquad\qquad} = -\sum_{i=1,k} P(e_i)\log_2 Q(e_i)$$



Original Gaussian PDF's

  - 하나의 변수(x)가 서로 다른 분포(P, Q)를 가질 경우, 해당 분포들의 차이를 의미
    - ✓ 두 확률 분포(P, Q)의 차이에 대한 정량적 지표
  - 실제 분포($P(x)$)에 대해서 알지 못하는 상태에서 모델링을 통해 구한 분포($Q(x)$)를 통해 실제 분포를 예측
    - ✓ $P(x)$: 학습 데이터의 분포(GT), Q(x) : 모델로 추정한 분포

Cross Entorpy

◆ Cross Entropy (CE)교차 엔트로피(cont'd)

- CE 식을 전개하면,

$$H(P,Q) = -\sum_x P(x)\log_2 Q(x)$$

$$= -\sum_x P(x)\log_2 P(x) + \sum_x P(x)\log_2 P(x) - \sum_x P(x)\log_2 Q(x)$$

$$= \underbrace{H(P)}_{\text{P's entropy}} + \underbrace{\sum_x P(x)\log_2 \frac{P(x)}{Q(x)}}_{\text{KL divergence}}$$

$$\boxed{\begin{aligned} P\text{와 } Q\text{의 교차 엔트로피 } H(P,Q) &= H(P) + \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \\ &= P\text{의 엔트로피} + P\text{와 } Q \text{ 간의 } KL \text{ 다이버전스} \end{aligned}}$$

KL Divergence

◆ KL divergence^KL 다이버전스

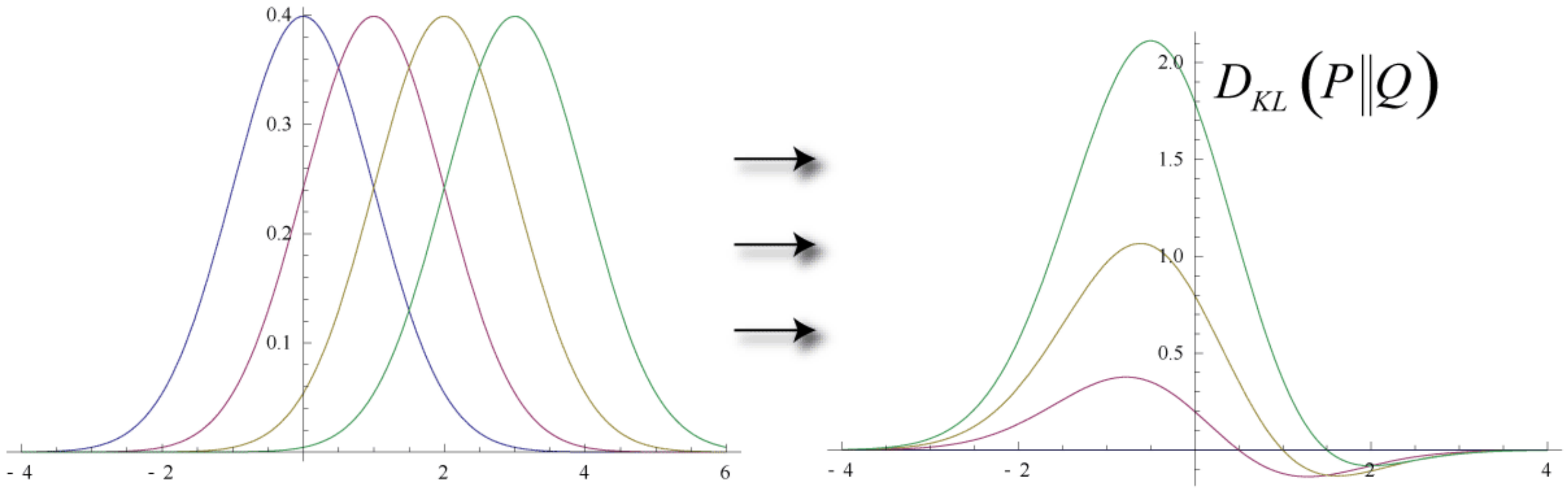
- ______________________________(얼마나 다른지)를 계산할 때 주로 사용, 하지만 $KL(P \| Q) \neq KL(Q \| P)$

$$KL(P \parallel Q) = \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (2.48)$$

예)

$P(x)$: 실제(GT)

$Q(x)$: 예측치

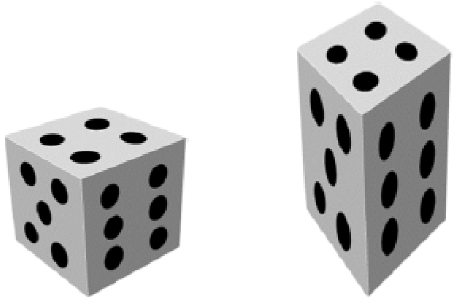


https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#/media/File:KL-Gauss-Example.png

예제 2-9

[그림 2-21]과 같이 정상적인 주사위와 찌그러진 주사위가 있는데, 정상적인 주사위의 확률분포는 P, 찌그러진 주사위의 확률분포는 Q를 따르며, P와 Q가 다음과 같이 분포한다고 가정하자.

$$P(1)=\frac{1}{6},\ P(2)=\frac{1}{6},\ P(3)=\frac{1}{6},\ P(4)=\frac{1}{6},\ P(5)=\frac{1}{6},\ P(6)=\frac{1}{6}$$
$$Q(1)=\frac{3}{12},\ Q(2)=\frac{1}{12},\ Q(3)=\frac{1}{12},\ Q(4)=\frac{1}{12},\ Q(5)=\frac{3}{12},\ Q(6)=\frac{3}{12}$$



(a) 정상 주사위 (b) 찌그러진 주사위

그림 2-21 확률분포가 다른 두 주사위

확률분포 P와 Q 사이의 교차 엔트로피와 KL 다이버전스는 다음과 같다.

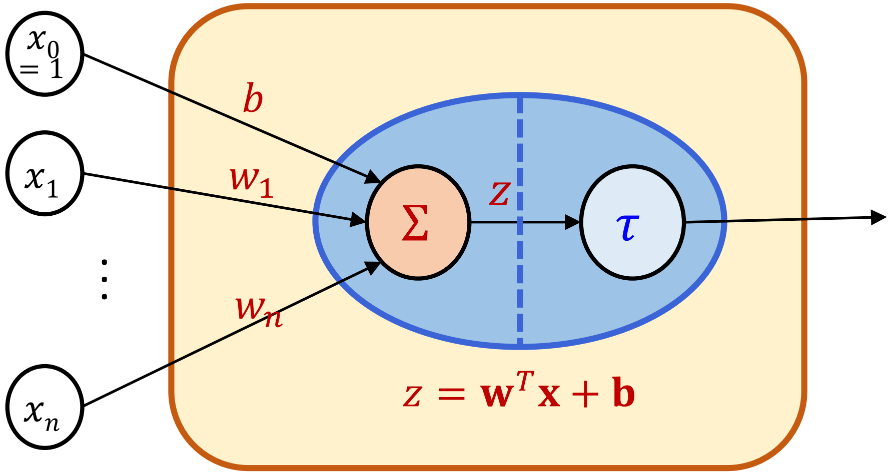
$$H(P,Q)=-\left(\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{3}{12}\right)=2.7925$$
$$KL(P\parallel Q)=\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2\frac{2}{3}=0.2075$$

[예제 2-8]에서 P의 엔트로피 $H(P)$는 2.585이었다. 따라서 식 (2.49)가 성립함을 알 수 있다.

Objective Functions

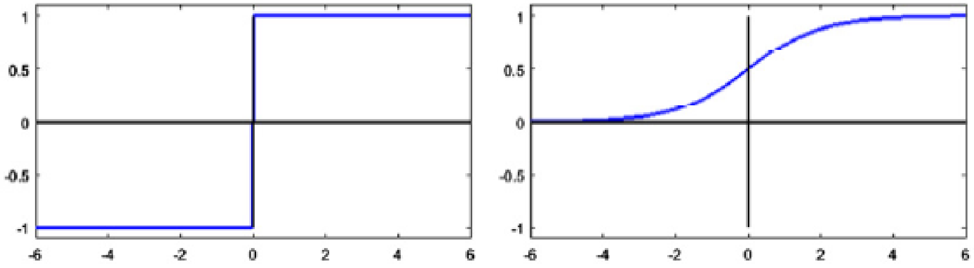
오일석, 기계학습, 5.1 목적함수: 교차 엔트로피와 로그우도

(Review) 퍼셉트론 + 활성함수



$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \tau(\mathbf{w}^T\mathbf{x} + \mathbf{b}) = \hat{y} = o$$

$$\tau(s) = \begin{cases} 1 & s \geq 0 \\ -1 & s < 0 \end{cases}$$

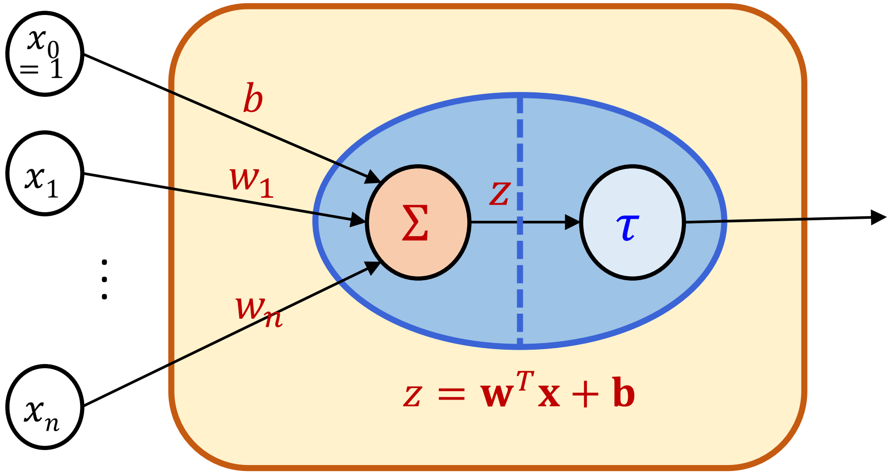


(a) 계단 함수 (b) 로지스틱 시그모이드

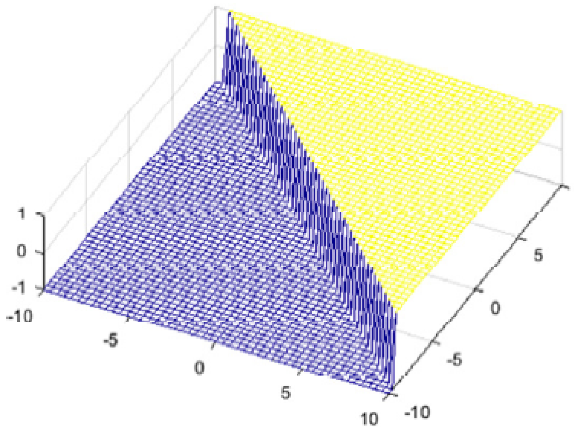
$$\tau(s) = \frac{1}{1 + e^{-s}}$$

Perceptron Logistic regression

(Review) 퍼셉트론 + 활성함수

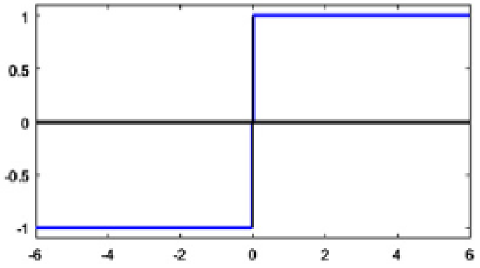


$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \tau(\mathbf{w}^T\mathbf{x} + \mathbf{b}) = \hat{y} = o$$



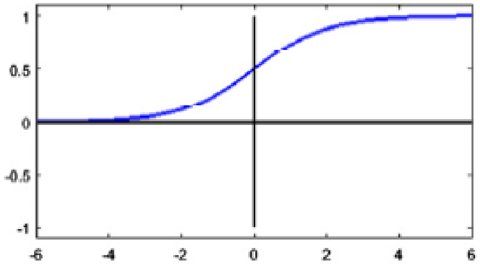
(a) 계단함수의 딱딱한 공간 분할

그림 3-13 퍼셉트론의 공간 분할 유형



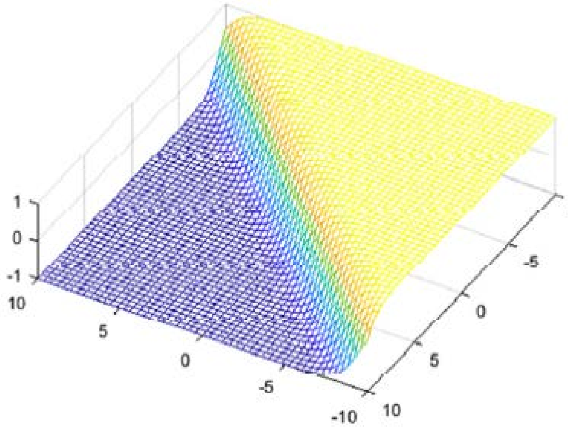
(a) 계단 함수

Perceptron



(b) 로지스틱 시그모이드

Logistic regression

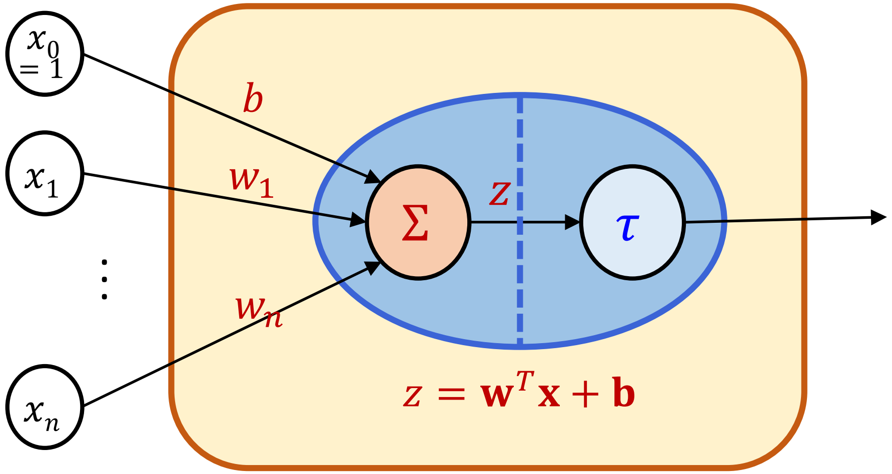


(b) 로지스틱 시그모이드의 부드러운 공간 분할

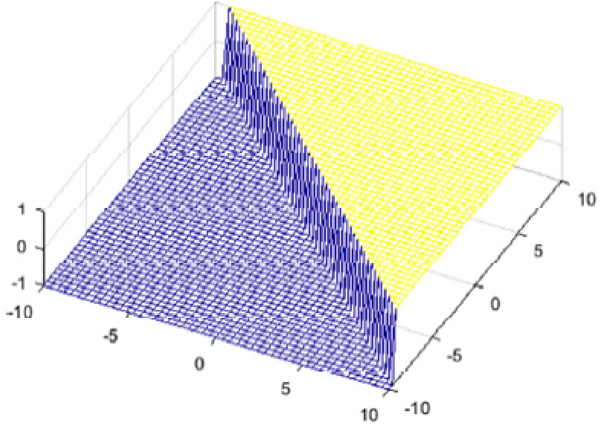
(Review) 퍼셉트론 + 활성함수

학습목적: $\Theta = (\boldsymbol{w}, b)$ 구하기

- ____________________
- ____________________

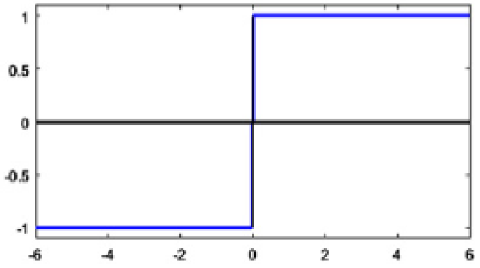


$h_{\boldsymbol{\theta}}(\mathbf{x}) = \tau(\mathbf{w}^T\mathbf{x} + \mathbf{b}) = \hat{y} = o$

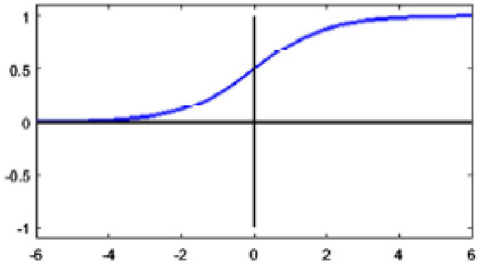


(a) 계단함수의 딱딱한 공간 분할

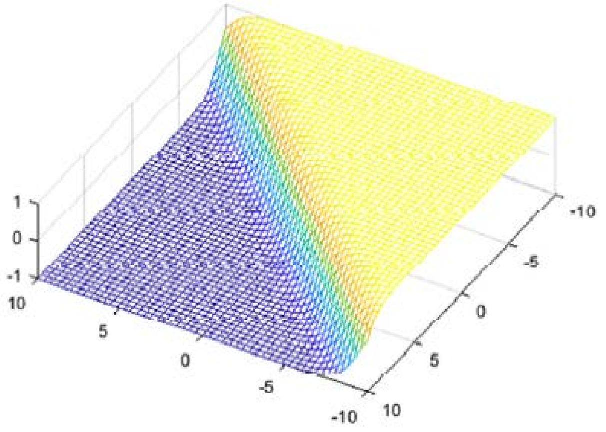
그림 3-13 퍼셉트론의 공간 분할 유형



(a) 계단 함수



(b) 로지스틱 시그모이드



(b) 로지스틱 시그모이드의 부드러운 공간 분할

Perceptron

Logistic regression

MSE 목적 함수

◆ Mean Square Error (MSE)평균제곱오차 목적 함수

- 오차가 클수록 e값이 크므로 목적 함수로 훌륭함.

- Sigmoid 활성 함수와 결합 시, 에러가 클 때 gradient가 오히려 작을 수 있다. 이 때문에, MSE의 느린 학습 문제 발생 (뒤에서 다룸)

Cross Entropy 목적 함수

◆ Cross Entropy교차 엔트로피

- 레이블에 해당하는 y가 확률변수 (부류가 2개라고 가정하면 $y \in \{0,1\}$)
- 확률 분포: P는 정답 레이블, Q는 신경망(퍼셉트론) 출력

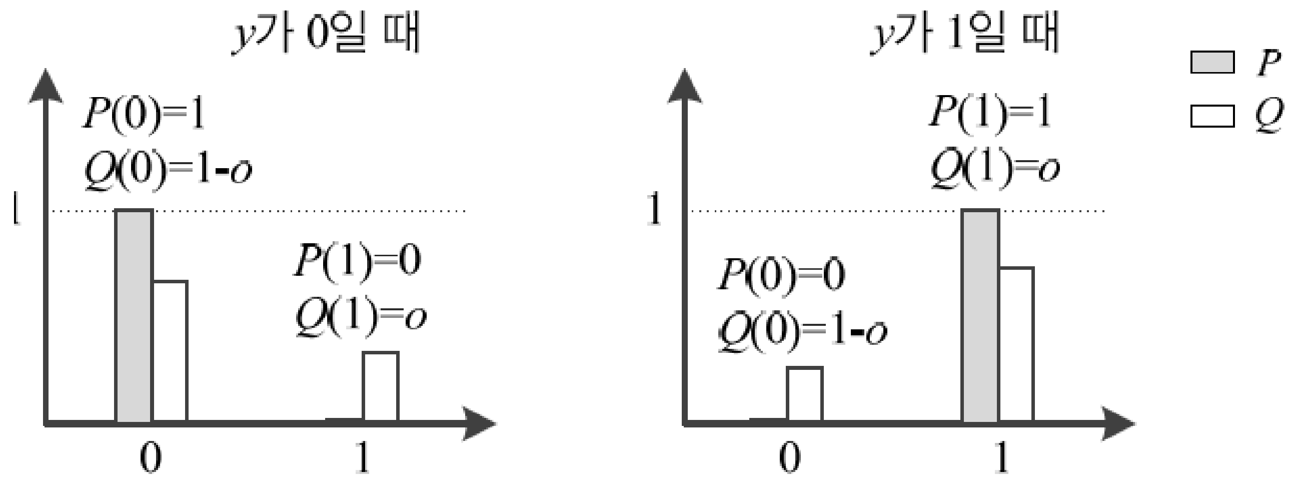
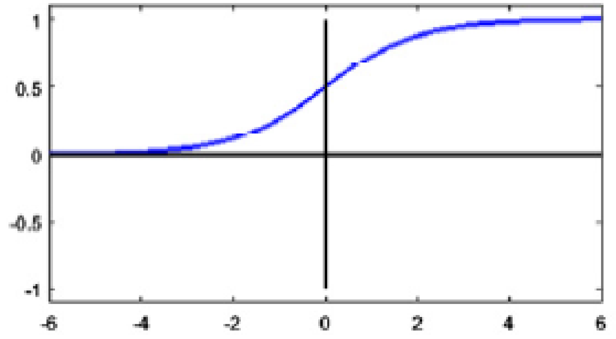


그림 5-3 레이블 y가 0일 때와 1일 때의 P와 Q의 확률분포

$$P(0) = 1 - y \quad Q(0) = 1 - o$$
$$P(1) = y \quad Q(1) = o$$



◆ Cross Entropy 목적 함수

- P 대한 Q의 교차 엔트로피

$$H(P,Q) = -\sum_{y\in\{0,1\}} P(y)\log_2 Q(y)$$
$$= -(P(0)\log_2 Q(0) + P(1)\log_2 Q(1))$$
$$= -\big((1-y)\log_2(1-o) + y\log_2 o\big)$$

$$P(0) = 1-y \qquad Q(0) = 1-o$$
$$P(1) = y \qquad Q(1) = o$$

- CE 목적함수 (for binary classification)

$$J(\boldsymbol{\Theta}) = e = -(y\log_2 o + (1-y)\log_2(1-o))$$

Cross Entropy 목적 함수

◆ Cross Entropy 목적 함수 (cont'd)

$$J(\boldsymbol{\Theta}) = e = -(y \log_2 o + (1-y)\log_2(1-o))$$ (for binary classification)

- 제구실 하는지 확인
  - y가 1, o가 0.98일 때 (예측이 잘된 경우)
    - ✓ 오류 $e = -(1\ \log_2 0.98 + (1-1)\log_2(1-0.98)) = 0.0291$로서 낮은 값
  - y가 1, o가 0.0001일 때 (예측이 엉터리인 경우)
    - ✓ 오류 $e = -(1\ \log_2 0.0001 + (1-1)\log_2(1-0.0001)) = 13.2877$로서 높은 값

- Sigmoid 활성 함수와 결합 시, 에러가 크면 gradient도 크게 만든다. 따라서, MSE의 느린 학습 문제 해결 (뒤에서 다룸)

Cross Entropy 목적 함수

◆ Cross Entropy 목적 함수 (cont'd)

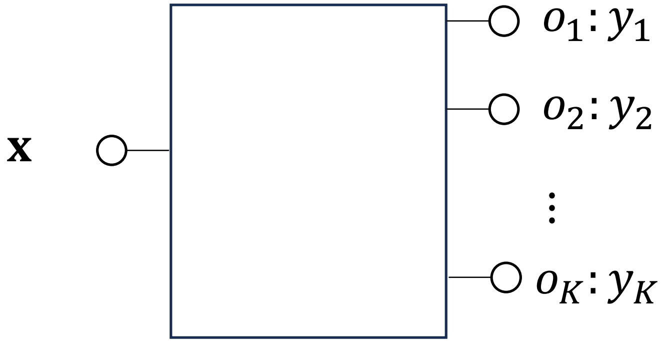
● 1개의 샘플에 대한 에러 (추론에서의 활용)

(for binary classification)

$$e = -(y\log(o) + (1-y)\log(1-o))$$

(for multiclass classification)

$$e =$$



(참고) CE $H(P,Q) = -\sum_{x} P(x)\log_2 Q(x)$

Cross Entropy 목적 함수

◆ Cross Entropy 목적 함수 (cont'd)

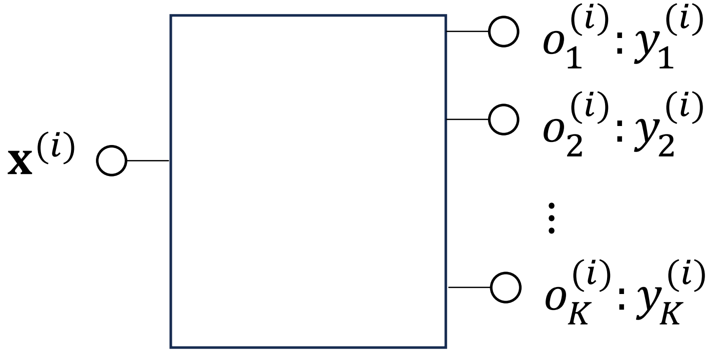
- n 개의 샘플에 대한 에러 (학습에서의 활용)

(for binary classification)

$$e = -\sum_{i=1}^{n}\left(y^{(i)}\log(o^{(i)}) + \left(1 - y^{(i)}\right)\log(1 - o^{(i)})\right)$$

(for multiclass classification)

$$e = \sum_{i=1}^{n}\sum_{k}^{K} -y_k^{(i)}\log(o_k^{(i)})$$



(참고) CE $H(P,Q) = -\sum_{x} P(x)\log_2 Q(x)$

MSE vs CE 학습 속도

시험에서는 틀린 만큼 합당한 벌점을 받는 것이 중요하다. 그래야 다음 시험에서 심기일전으로 공부하여 틀리는 개수를 줄일 가능성이 크기 때문이다. 틀린 개수에 상관없이 비슷한 벌점을 받는다면 나태해져 성적을 올리는 데 지연이 발생할 것이다. 이러한 원리가 기계 학습에도 적용될까?

◆ (Review L03) 경사 하강법 가중치 갱신 규칙 $\mathbf{\Theta} = \mathbf{\Theta} - \rho\mathbf{g}$

MSE vs CE 학습 속도

◆ MSE의 허점

- 왼쪽 상황은 $e = 0.2815$, 오른쪽 상황은 $e = 0.4971$이므로 오른쪽이 더 큰 벌점을 받아야 마땅함

$$J(\boldsymbol{\Theta}) = e = \frac{1}{2}\|y - o\|_2^2$$

$$\boldsymbol{\Theta} = \boldsymbol{\Theta} - \rho\mathbf{g}$$

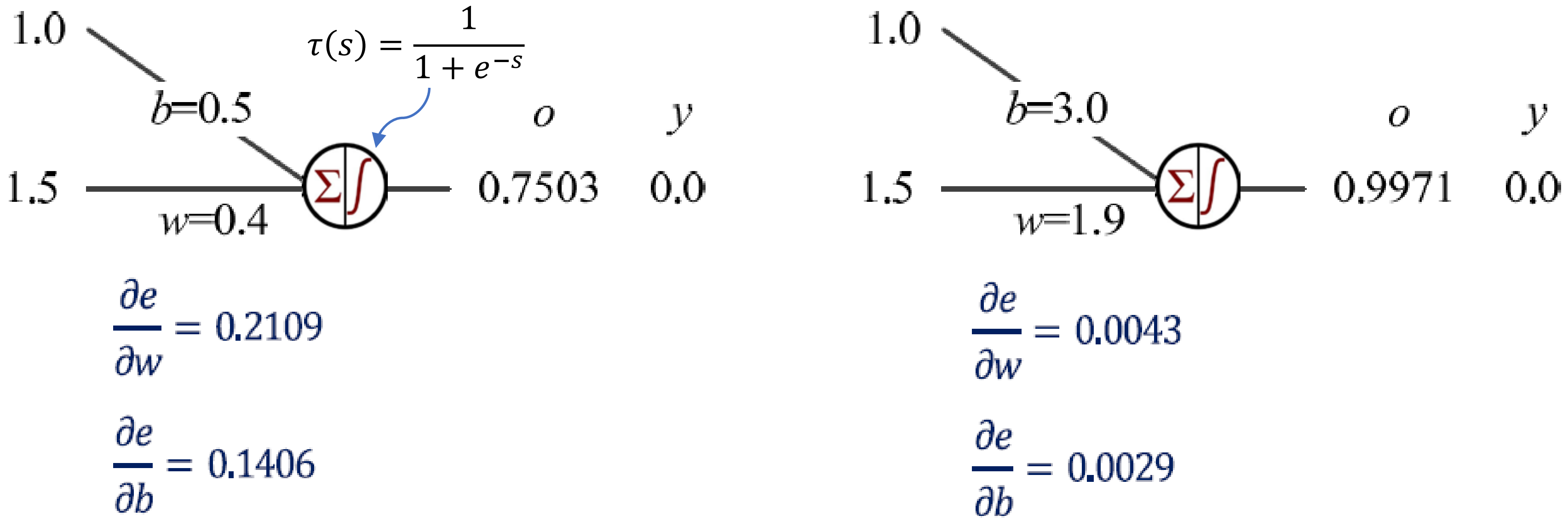


그림 5-1 MSE가 목적함수로서 부적절한 상황

더 많은 오류를 범한 상황이 더 낮은 벌점을 받는 꼴 → 학습이 더딘 부정적 효과

MSE vs CE 학습 속도

$$\frac{\partial J(\Theta)}{\partial w} = \frac{\partial}{\partial w}\left(\frac{1}{2}(y-o)^2\right)$$

$$= \frac{1}{2} \times 2 \times (y-o) \times (-1) \times \frac{\partial o}{\partial w}$$

$$= -(y-o)o(1-o)\frac{\partial z}{\partial w}$$

$$= -(y-o)o(1-o)x$$

$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial o}\frac{\partial o}{\partial w}$$

$$\frac{\partial o}{\partial w} = \frac{\partial \tau(z)}{\partial w} = \frac{\partial \tau(z)}{\partial z}\frac{\partial z}{\partial w} = \tau(z)(1-\tau(z))\frac{\partial z}{\partial w}$$

$$\frac{\partial z}{\partial w} = \frac{\partial(wx+b)}{\partial w} = x$$

표 3-1 활성함수로 사용되는 여러 함수

함수 이름	함수	1차 도함수	범위
계단	$\tau(s) = \begin{cases} 1 & s \geq 0 \\ -1 & s < 0 \end{cases}$	$\tau'(s) = \begin{cases} 0 & s \neq 0 \\ \text{불가} & s = 0 \end{cases}$	-1과 1
로지스틱 시그모이드	$\tau(s) = \frac{1}{1+e^{-as}}$	$\tau'(s) = a\tau(s)(1-\tau(s))$	(0,1)

L07, p.9 (오일석 기계학습)

MSE vs CE 학습 속도

$$\frac{\partial J(\Theta)}{\partial b} = \frac{\partial}{\partial b}\left(\frac{1}{2}(y-o)^2\right)$$

$$= \frac{1}{2} \times 2 \times (y-o) \times (-1) \times \frac{\partial o}{\partial b}$$

$$= -(y-o)o(1-o)\frac{\partial z}{\partial b}$$

$$= -(y-o)o(1-o) \times 1$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial o}\frac{\partial o}{\partial b}$$

$$\frac{\partial o}{\partial b} = \frac{\partial \tau(z)}{\partial b} = \frac{\partial \tau(z)}{\partial z}\frac{\partial z}{\partial b} = \tau(z)(1-\tau(z))\frac{\partial z}{\partial b}$$

$$\frac{\partial z}{\partial b} = \frac{\partial(wx+b)}{\partial b} = 1$$

표 3-1 활성함수로 사용되는 여러 함수

함수 이름	함수	1차 도함수	범위
계단	$\tau(s) = \begin{cases} 1 & s \geq 0 \\ -1 & s < 0 \end{cases}$	$\tau'(s) = \begin{cases} 0 & s \neq 0 \\ \text{불가} & s = 0 \end{cases}$	-1과 1
로지스틱 시그모이드	$\tau(s) = \frac{1}{1+e^{-as}}$	$\tau'(s) = a\tau(s)(1-\tau(s))$	(0,1)

L07, p.9 (오일석 기계학습)

MSE vs CE 학습 속도

◆ MSE의 허점

● 이유

▪ $z = wx + b$(아래 그래프의 가로축에 해당)가 커지면 ______________________

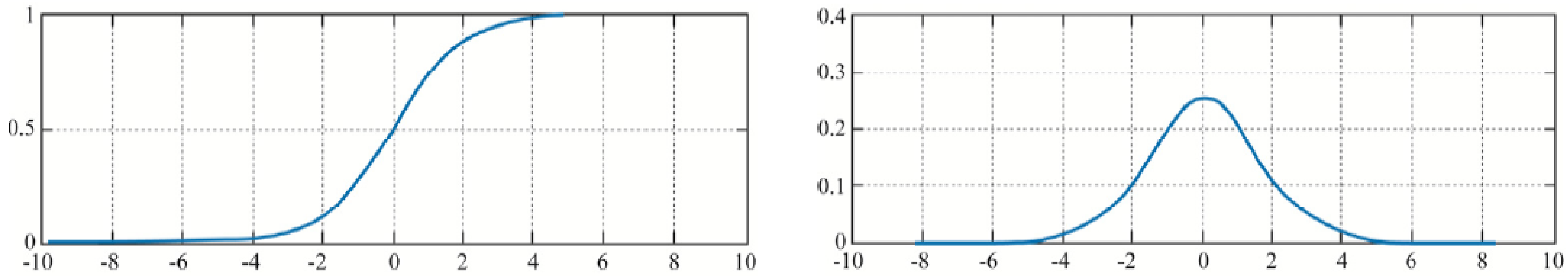


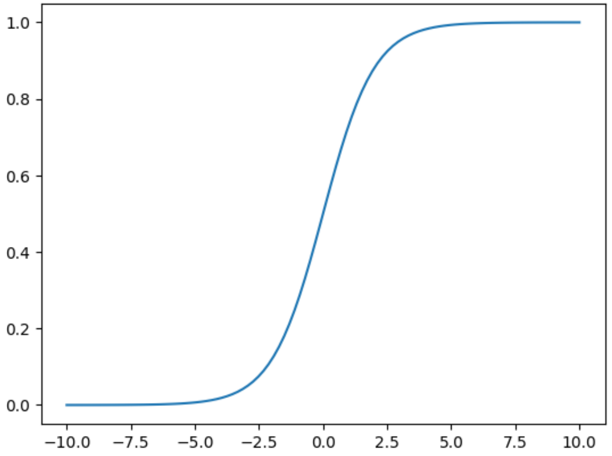
그림 5-2 로지스틱 시그모이드함수와 도함수

$$\tau(s) = \frac{1}{1+e^{-s}}$$

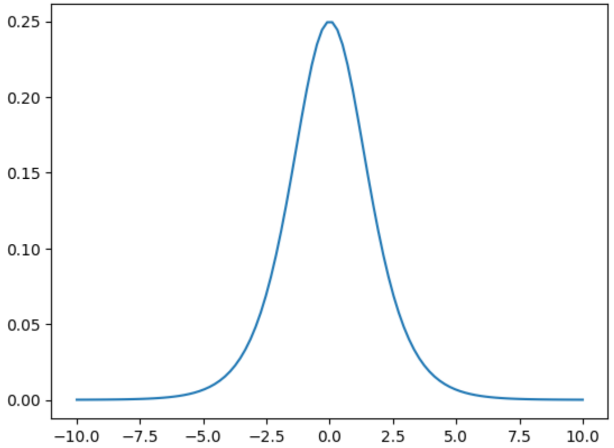
$$\tau(s)' = \tau(s)\big(1-\tau(s)\big) = \frac{1}{1+e^{-s}}\left(1-\frac{1}{1+e^{-s}}\right)$$

예제) 코드

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
x = np.linspace(-10,10, 100)
y = 1/(1+np.exp(-x))
plt.plot(x,y)
plt.show()
```




```
y2 = y*(1-y)
plt.plot(x,y2)
plt.show()
```



MSE vs CE 학습 속도

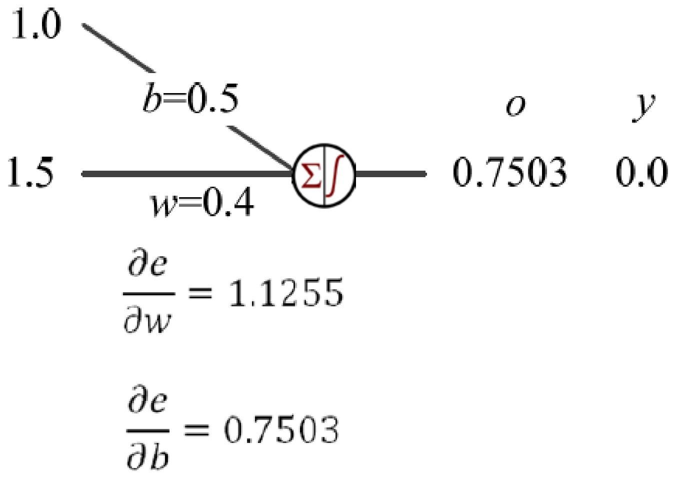
◆ CE의 경우

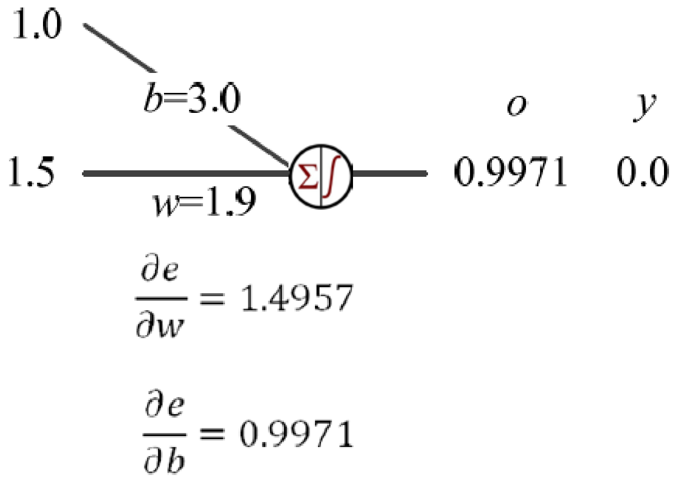
$$J(\theta) = e = -(y \log(o) + (1-y)\log(1-o))$$

$$\frac{\partial J(\theta)}{\partial w} = x(o-y) \qquad \frac{\partial J(\theta)}{\partial b} = (o-y)$$

- 그레이디언트를 계산해 보면, 오류가 더 큰 오른쪽에 더 큰 벌점 부과 (MSE 단점 해결)
- ________________

$$\boldsymbol{\Theta} = \boldsymbol{\Theta} - \rho\mathbf{g}$$





MSE vs CE 학습 속도

$$\frac{\partial J(\Theta)}{\partial w} = \frac{\partial}{\partial w}\left(-(y\log(o) + (1-y)\log(1-o))\right)$$

$$= -\left(\frac{y}{o} + \frac{1-y}{1-o} \times (-1)\right) \times \frac{\partial o}{\partial w}$$

$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial o}\frac{\partial o}{\partial w}$$

$$= -\left(\frac{y}{o} - \frac{1-y}{1-o}\right) o(1-o)\frac{\partial z}{\partial w}$$

$$\frac{\partial o}{\partial w} = \frac{\partial \tau(z)}{\partial w} = \frac{\partial \tau(z)}{\partial z}\frac{\partial z}{\partial w} = \tau(z)(1-\tau(z))\frac{\partial z}{\partial w}$$

$$= -\left(\frac{y}{o} - \frac{1-y}{1-o}\right) o(1-o)x$$

$$\frac{\partial z}{\partial w} = \frac{\partial(wx+b)}{\partial w} = x$$

$$= (o-y)x$$

표 3-1 활성함수로 사용되는 여러 함수

함수 이름	함수	1차 도함수	범위
계단	$\tau(s) = \begin{cases} 1 & s \geq 0 \\ -1 & s < 0 \end{cases}$	$\tau'(s) = \begin{cases} 0 & s \neq 0 \\ \text{불가} & s = 0 \end{cases}$	-1과 1
로지스틱 시그모이드	$\tau(s) = \frac{1}{1+e^{-as}}$	$\tau'(s) = a\tau(s)\big(1-\tau(s)\big)$	(0,1)

L07, p.9 (오일석 기계학습)

MSE vs CE 학습 속도

$$\frac{\partial J(\Theta)}{\partial b} = \frac{\partial}{\partial b}\big(-(y\log(o) + (1-y)\log(1-o))\big)$$

$$= -\left(\frac{y}{o} + \frac{1-y}{1-o} \times (-1)\right) \times \frac{\partial o}{\partial b}$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial o}\frac{\partial o}{\partial b}$$

$$= -\left(\frac{y}{o} - \frac{1-y}{1-o}\right) o(1-o)\frac{\partial z}{\partial b}$$

$$\frac{\partial o}{\partial b} = \frac{\partial \tau(z)}{\partial b} = \frac{\partial \tau(z)}{\partial z}\frac{\partial z}{\partial b} = \tau(z)(1-\tau(z))\frac{\partial z}{\partial b}$$

$$= -\left(\frac{y}{o} - \frac{1-y}{1-o}\right) o(1-o) \times 1$$

$$\frac{\partial z}{\partial b} = \frac{\partial(wx+b)}{\partial b} = 1$$

$$= (o-y)$$

표 3-1 활성함수로 사용되는 여러 함수

함수 이름	함수	1차 도함수	범위
계단	$\tau(s) = \begin{cases} 1 & s \ge 0 \\ -1 & s < 0 \end{cases}$	$\tau'(s) = \begin{cases} 0 & s \ne 0 \\ \text{불가} & s = 0 \end{cases}$	-1과 1
로지스틱 시그모이드	$\tau(s) = \frac{1}{1+e^{-as}}$	$\tau'(s) = a\tau(s)(1-\tau(s))$	(0,1)

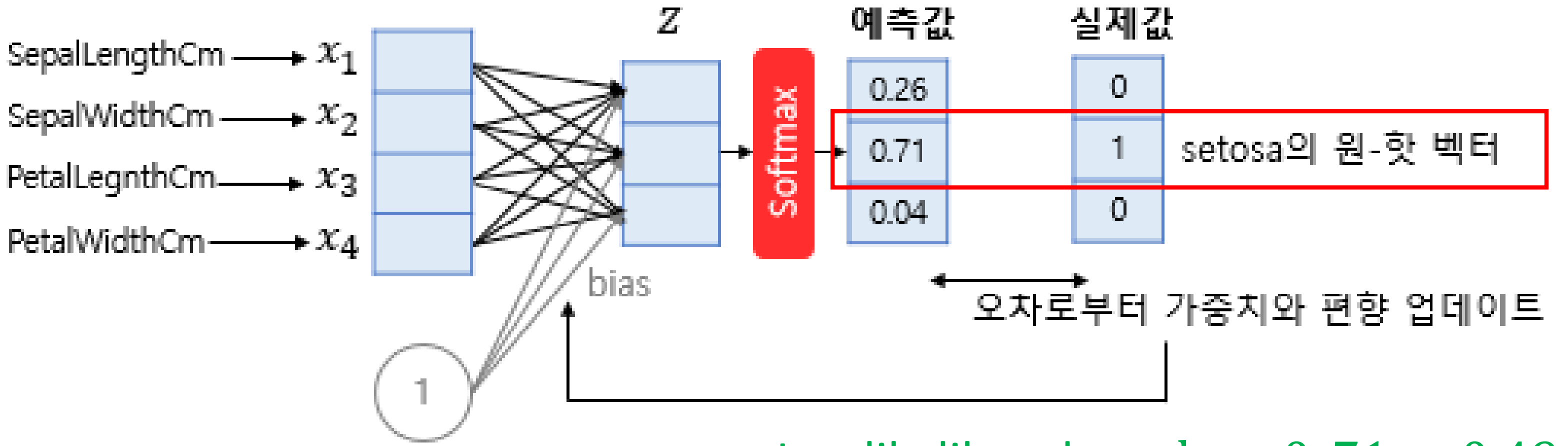
L07, p.9 (오일석 기계학습)

Maximum Likelihood 목적 함수

◆ (Negative) ____________________ (음의) 로그우도 목적 함수

$$J(\boldsymbol{\Theta}) = e = -\log_2 o_y$$

● 모든 출력 노드값을 사용하는 MSE나 교차 엔트로피와 달리 o_y라는 ______________만 사용



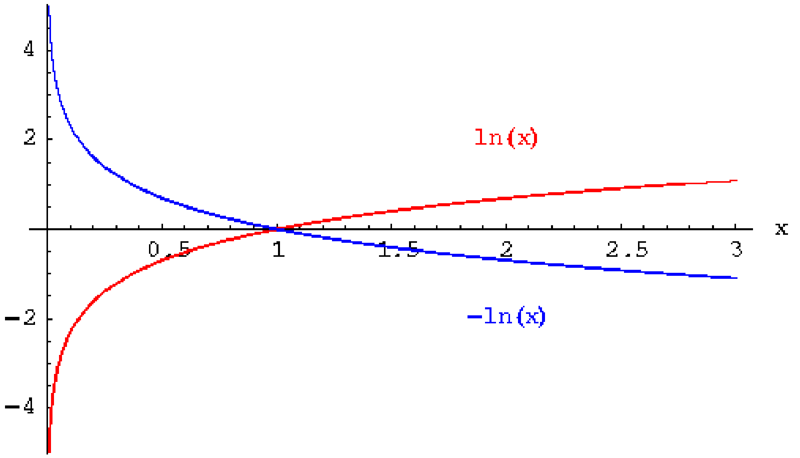
Log likelihood = $-\log_2 0.71 = 0.49$

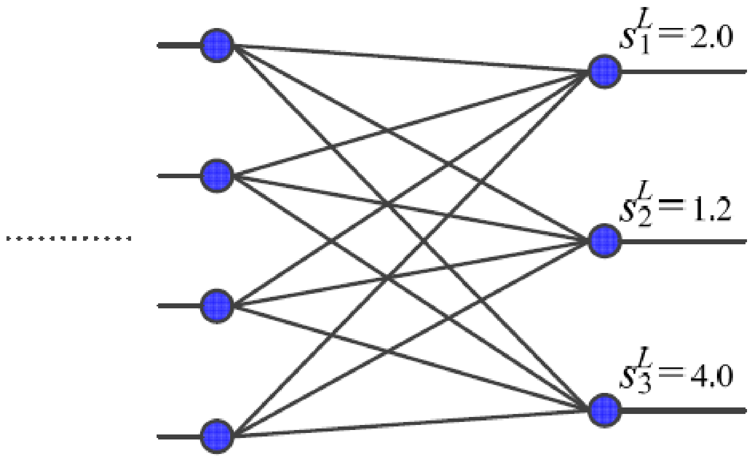
https://wikidocs.net/images/page/35476/softmax6_final.PNG

Maximum Likelihood 목적 함수

◆ Softmax와 로그우도

- Softmax는 최댓값이 아닌 값을 억제하여 0에 가깝게 만든다는 의도 내포
- 학습 샘플이 알려주는 부류에 해당하는 노드만 보겠다는 로그우도와 잘 어울림
- 따라서 ______________________________ 결합하여 사용하는 경우가 많음





로지스틱 시그모이드	max	softmax	Log likelihood
0.8808	0.0	0.1131	3.14
0.7685	0.0	0.0508	4.30
0.9820	1.0	0.8360	0.2584

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$o = f(x_i; W)$$

$$P(Y = k | X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$ Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

$$L_i = -\log P(Y = y_i | X = x_i)$$

	Unnormalized log-probabilities / logits	$\xrightarrow{\text{exp}}$	unnormalized probabilities	$\xrightarrow{\text{normalize}}$	probabilities
cat	**3.2**		**24.5**		**0.13**
car	5.1		164.0		0.87
frog	-1.7		0.18		0.00
	s_k		e^{s_k}		$\frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$

→ $L_i = -\log(0.13)$ **= 2.04**

Maximum Likelihood Estimation
Choose weights to maximize the likelihood of the observed data

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$\boxed{o = f(x_i; W)}$$

$$\boxed{P(Y = k | X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}}$$ Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

$$L_i = -\log P(Y = y_i | X = x_i)$$

	Unnormalized log-probabilities / logits s_k	$\xrightarrow{\text{exp}}$	unnormalized probabilities e^{s_k}	$\xrightarrow{\text{normalize}}$	probabilities $\frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$		Correct probs
cat	**3.2**		**24.5**		**0.13**	→ compare ←	**1.00**
car	5.1		164.0		0.87		0.00
frog	-1.7		0.18		0.00		0.00

Softmax Classifier (Multinomial Logistic Regression)

Want to interpret raw classifier scores as **probabilities**

$$o = f(x_i; W)$$

$$P(Y = k | X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$ Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

$$L_i = -\log P(Y = y_i | X = x_i)$$

	Unnormalized log-probabilities / logits		unnormalized probabilities		probabilities		Correct probs
cat	3.2	exp →	24.5	normalize →	0.13	→ compare ←	1.00
car	5.1		164.0		0.87		0.00
frog	-1.7		0.18		0.00		0.00

Kullback–Leibler divergence

$$D_{KL}(P \| Q) = \sum_y P(y) \log \frac{P(y)}{Q(y)}$$

감사합니다.