기계 학습과 수학

1. 선형대수
2. 확률과 통계

박달님 03.11.2021

선형대수

벡터와 행렬

- 샘플 - 특징 벡터 feature vector
  - 예) Iris 데이터 - 꽃받침 길이, 넓이, 꽃잎 길이, 넓이 4개의 특징
- 훈련집합을 담은 행렬: 설계행렬 design matrix
- 전치행렬 $\mathbf{A}^{\mathrm{T}}$

$$\mathbf{X} = \begin{pmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ \vdots & \vdots & \vdots & \vdots \\ 6.2 & 3.4 & 5.4 & 2.3 \\ 5.9 & 3.0 & 5.1 & 1.8 \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \\ x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{149,1} & x_{149,2} & x_{149,3} & x_{149,4} \\ x_{150,1} & x_{150,2} & x_{150,3} & x_{150,4} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{150}^{\mathrm{T}} \end{pmatrix}$$

벡터와 행렬

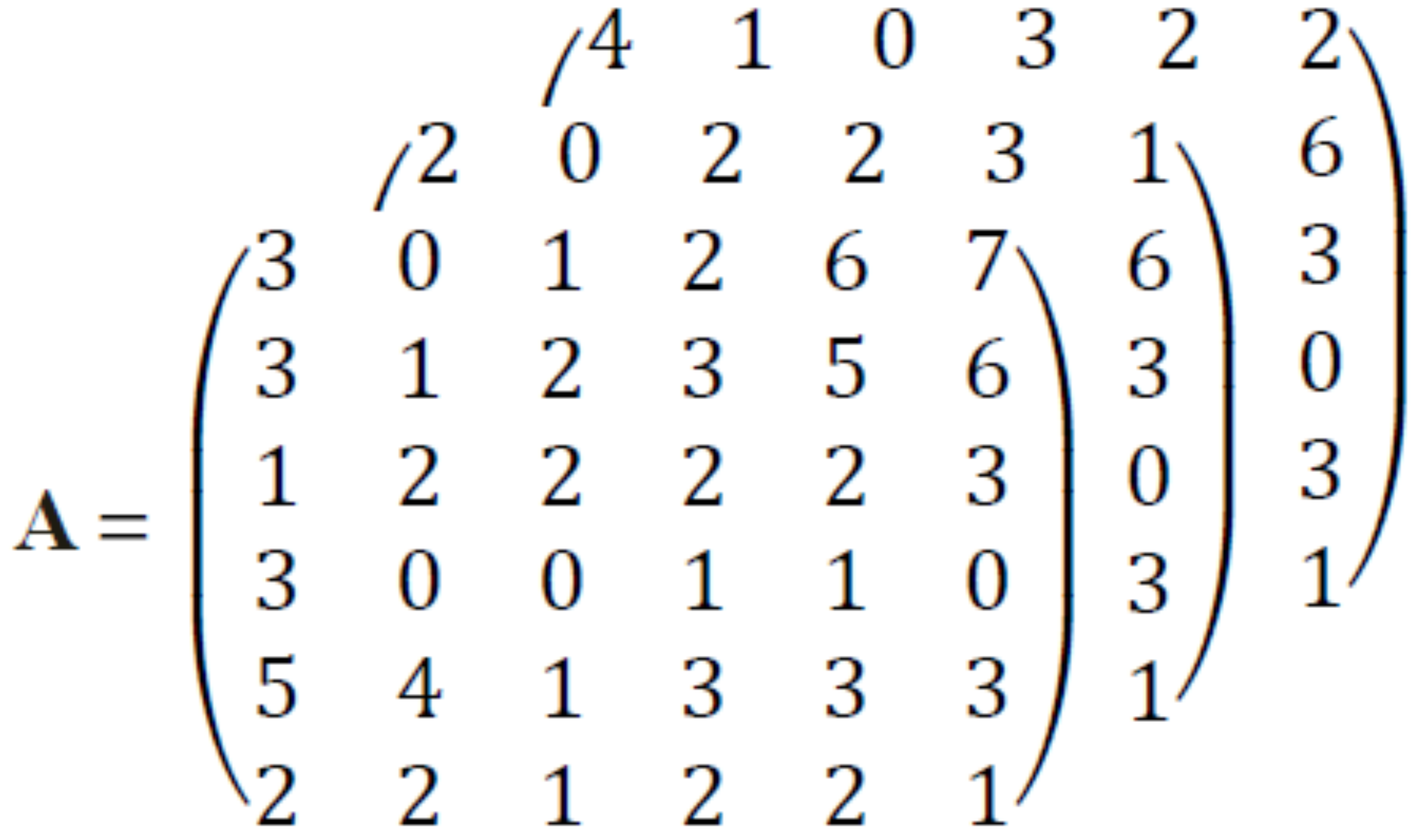
행렬 연산

- 행렬 곱셈
- 행렬 곱셈은 교환법칙 성립하지 않음. AB ≠ BA
- 분배법칙과 결합법칙은 성립함 A(B+C) = AB + AC, A(BC) = (AB)C
- C = AB, 이때 $c_{ij} = \sum_{k=1,s} a_{ik} b_{kj}$
- 차원이 같은 두 벡터 a와 b의 곱 - 내적 dot product $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathrm{T}}\mathbf{b} = \sum_{k=1,d} a_k b_k$

벡터와 행렬

텐서

- 3차원 이상의 구조를 가진 숫자 배열
- 예) RGB 컬러 영상
- 6*6*3 텐서
- 스칼라 - 0차원, 벡터 - 1차원, 행렬- 2차원 텐서



놈과 유사도

놈

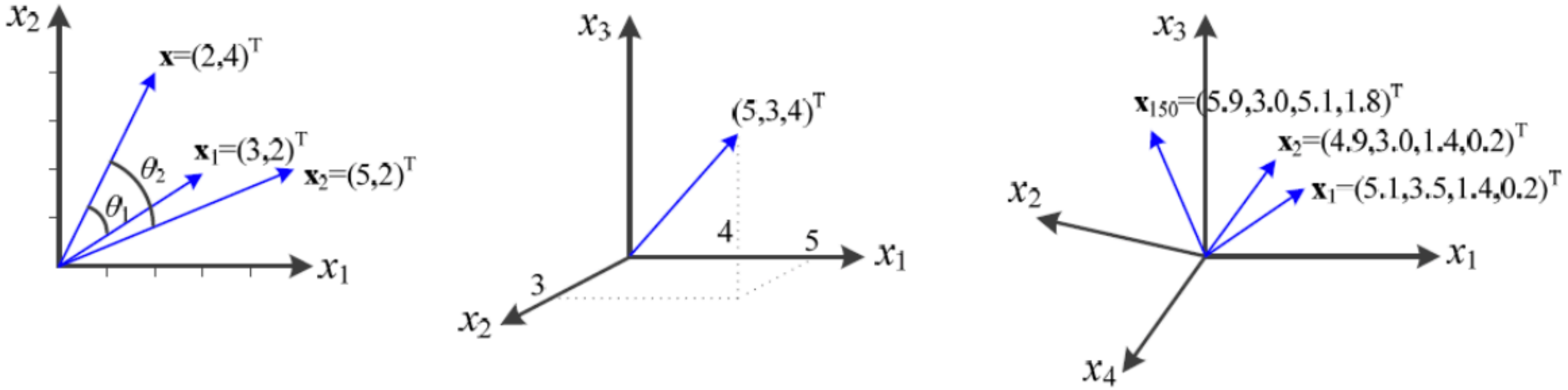
- 두 샘플의 유사도 측정
- 유사도 이용 - 비슷한 샘플을 찾아 같은 군집으로 모으거나 가장 유사한 샘플을 찾아 그 샘플이 속한 부류로 분류 가능
- 놈 norm - 벡터의 크기를 정의. P차 놈 - Lp놈 p차 놈: $\|\mathbf{x}\|_p = \left(\sum_{i=1,d} |x_i|^p\right)^{\frac{1}{p}}$
- 최대 norm $\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \cdots, |x_d|)$
- 기계 학습이 주로 사용 - 프로베니우스 놈 Frobenius noorm. 요소들의 제곱합의 제곱근

$$\|\mathbf{A}\|_F = \left(\sum_{i=1,n} \sum_{j=1,m} a_{ij}^2\right)^{\frac{1}{2}} \qquad \left\|\begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix}\right\|_F = \sqrt{2^2 + 1^2 + 6^2 + 4^2} = 7.550$$

놈과 유사도

유사도와 거리

- 벡터를 기하학적으로 해석



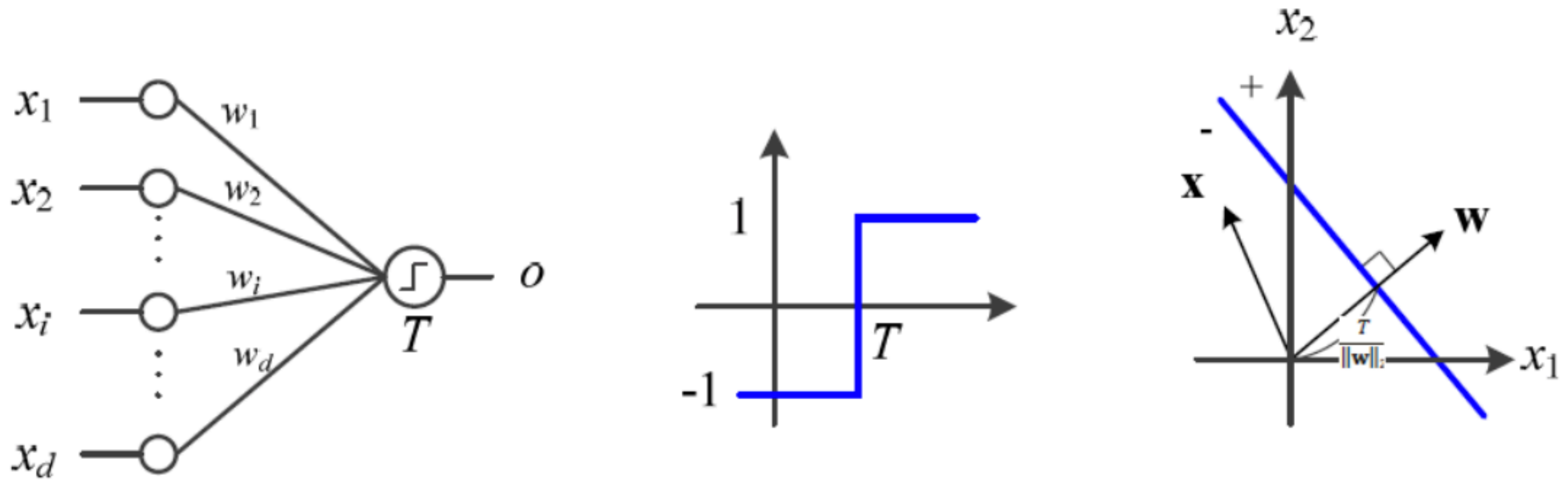
(a) 2차원 벡터 (b) 3차원 벡터 (c) 4차원 벡터(Iris 데이터)

- 코사인 유사도

$$cosine\_similarity(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|} = cos(\theta)$$

퍼셉트론의 해석

- 퍼셉트론 - 입력 샘플을 2개의 부류 중 하나로 분류하는 분류기 classifier

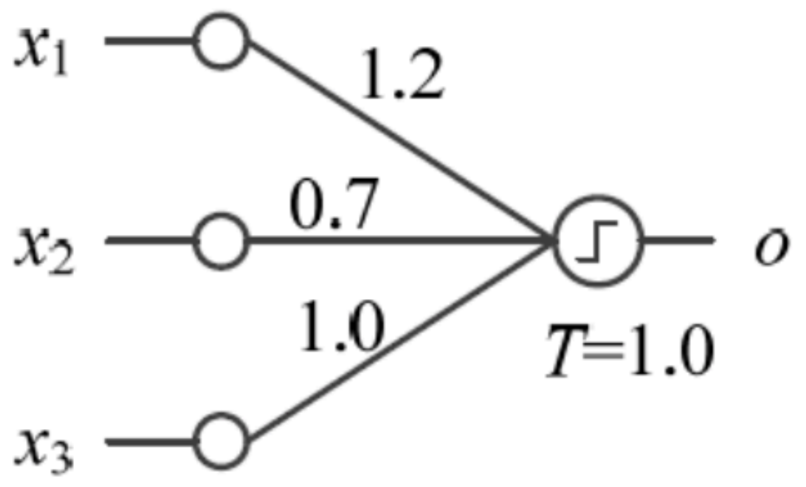


(a) 퍼셉트론 구조 (b) 계단형 활성함수(비선형) (c) 퍼셉트론의 공간 분할

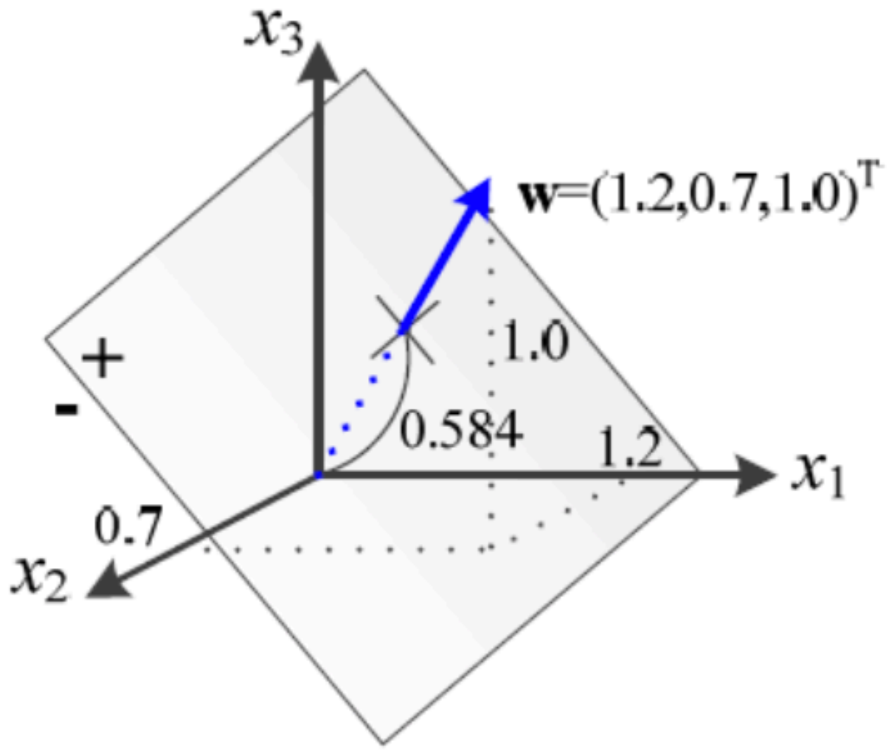
- 퍼셉트론의 동작을 수식으로 표현: $o = \tau(\mathbf{w} \cdot \mathbf{x})$, 이때 $\tau(a) = \begin{cases} 1, & a \geq T \\ -1, & a < T \end{cases}$

퍼셉트론

- 파란 직선은 두개의 부분 공간을 나누는 결정직선 decision line
- w에 수직이고 원점으로부터 $\dfrac{T}{\|\mathbf{w}\|_2}$ 만큼 떨어져 있음
- 3차원 특징공간 - 결정평면 decision plane, 4차원 이상 - 결정 초평면 decision hyperplane
- 예) 3차원 특징공간을 위한 퍼셉트론



(a) 퍼셉트론



(b) 공간 분할(2부류 분류)

퍼셉트론

학습의 정의

- 학습을 마친 프로그램을 현장에 설치했을때:
- 분류:

$$\overbrace{\mathbf{o}}^{?} = \boldsymbol{\tau}(\overbrace{\mathbf{W}}^{\text{앎}}\,\overbrace{\mathbf{x}}^{\text{앎}})$$

- 학습 과정 - 훈련집합의 샘플에 대해 식을 가장 잘 만족하는 W를 찾아내는 작업
- 학습:

$$\overbrace{\mathbf{o}}^{\text{앎}} = \boldsymbol{\tau}(\overbrace{\mathbf{W}}^{?}\,\overbrace{\mathbf{x}}^{\text{앎}})$$

- 기계 학습은 특히 선형대수를 많이 사용. 퍼셉트론을 여러 층으로 확한 것.

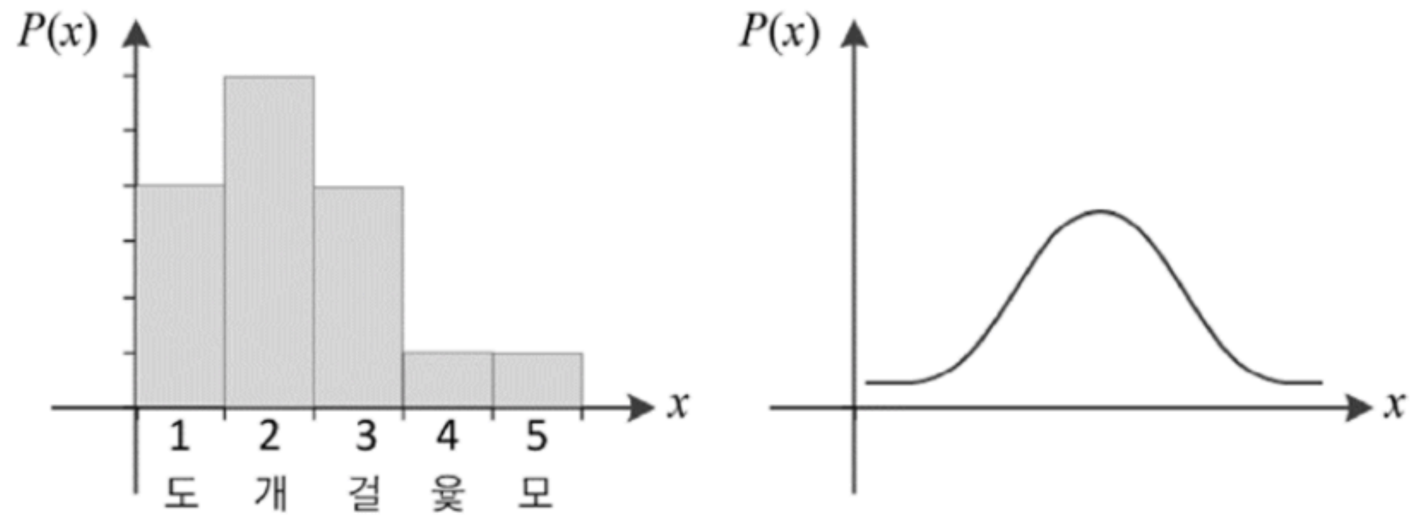
확률과 통계

5.6 Bayesian Statistics
베이즈 정리

확률 기초

확률변수와 확률분포

- 확률변수 - random variable
- 확률변수가 가질 수 있는 집합 - 정의역
    - 예) 윷놀이에서의 정의역 {도, 개, 걸, 윷, 모)
- 정의역 전체에 걸쳐 확률을 표현한 것 - 확률분포 probability distribution
- 이산 일때 확률분포 - 확률질량함수 probability mass function
- 연속인 경우 - 확률밀도함수 probability density function

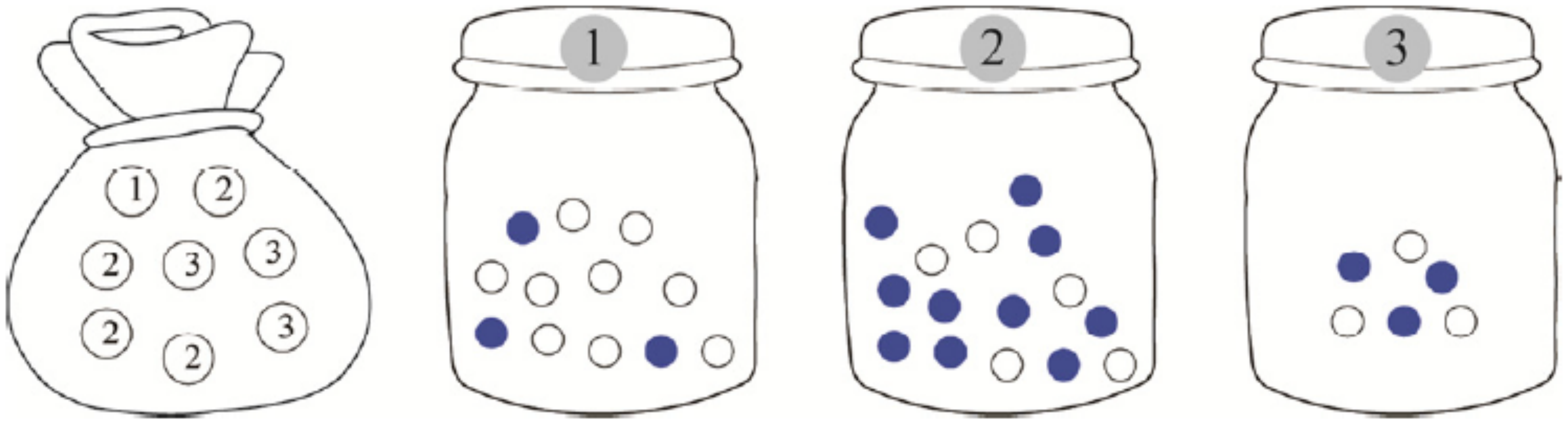


(a) 이산인 경우의 확률질량함수 (b) 연속인 경우의 확률밀도함수

확률

간단한 예시

- 주머니에서 번호를 뽑은 뒤 번호에 따라 해당 병에서 어떤 색깔의 공을 뽑음
- 번호: y, 공의 색: x라는 확률 변수로 표현. 정의역 $y \in \{①,②,③\}$, $x \in \{파랑, 하양\}$



확률

곱 규칙과 합 규칙

- ①번 카드를 뽑을 확률 $P(y{=}①){=}P(①){=}1/8$
- 카드는 ①번, 공은 하양일 확률은 $P(y{=}①,x{=}\text{하양}){=}P(①,\text{하양})$ ← 결합확률

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

- 곱 규칙 곱 규칙: $P(y,x)=P(x|y)P(y)$
- 하얀 공이 뽑힐 확률 $P(\text{하양})=P(\text{하양}|①)P(①)+P(\text{하양}|②)P(②)+P(\text{하양}|③)P(③)$
$$=\frac{9}{12}\frac{1}{8}+\frac{5}{15}\frac{4}{8}+\frac{3}{6}\frac{3}{8}=\frac{43}{96}$$
- 합 규칙

합 규칙: $P(x)=\sum P(y,x)=\sum P(x|y)P(y)$

베이즈 정리

베이즈 정리 식

- 일반적으로 x와 y가 같이 일어날 결합확률이나 y와 x가 같이 일어날 결합확률이 같음

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

- “햐얀 공이 나왔다는 사실만 알고 어느 병에서 나왔는지 모르는데, 어느 병인지 추정하라.”

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

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

베이즈 정리

베이즈 정리의 해석

- 세 가지 경우에 대해 확률을 계산하면, 3번 병일 확률이 가장 높다
- 해석:

$$\overbrace{P(y|x)}^{\text{사후확률}} = \frac{\overbrace{P(x|y)}^{\text{우도}}\ \overbrace{P(y)}^{\text{사전확률}}}{P(x)}$$

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

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

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

우도: P(x | y) = L(y|x)

알고 있음 추정해야함

베이즈 정리와 기계 학습

기계 학습에 적용

- 예) Iris 데이터 분류 문제
- 특징 벡터 x, 부류 $y \in$ {setosa, versicolor, virginica}

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

$$\xrightarrow{\text{특징추출}} \mathbf{x} = (7.0, 3.2, 4.7, 1.4)^{\mathrm{T}} \xrightarrow{\text{사후확률 추정}} \begin{array}{l} P(\text{setosa}|\mathbf{x}) = 0.18 \\ P(\text{versicolor}|\mathbf{x}) = 0.72 \\ P(\text{virginica}|\mathbf{x}) = 0.10 \end{array} \xrightarrow{\text{argmax}} \text{versicolor}$$

- 사후확률 P(y|**x**)를 직접 추정하는 일은 아주 단순한 경우를 빼고 불가능. 따라서 베이즈 정리를 이용하여 추정함.

베이즈 정리와 기계 학습

기계 학습에 적용

- 사전확률 P(y)와 우도 P(x|y)를 구할 수 있다면 베이즈 공식을 이용하여 사후확률을 간접적으로 계산 가능

- 우도 측정이 훨씬 쉽다. 부류별로 독립적으로 확률 추정 가능. **확률밀도 추정** 방법 사용.

사전확률: $P(y = c_i) = \frac{n_i}{n}$

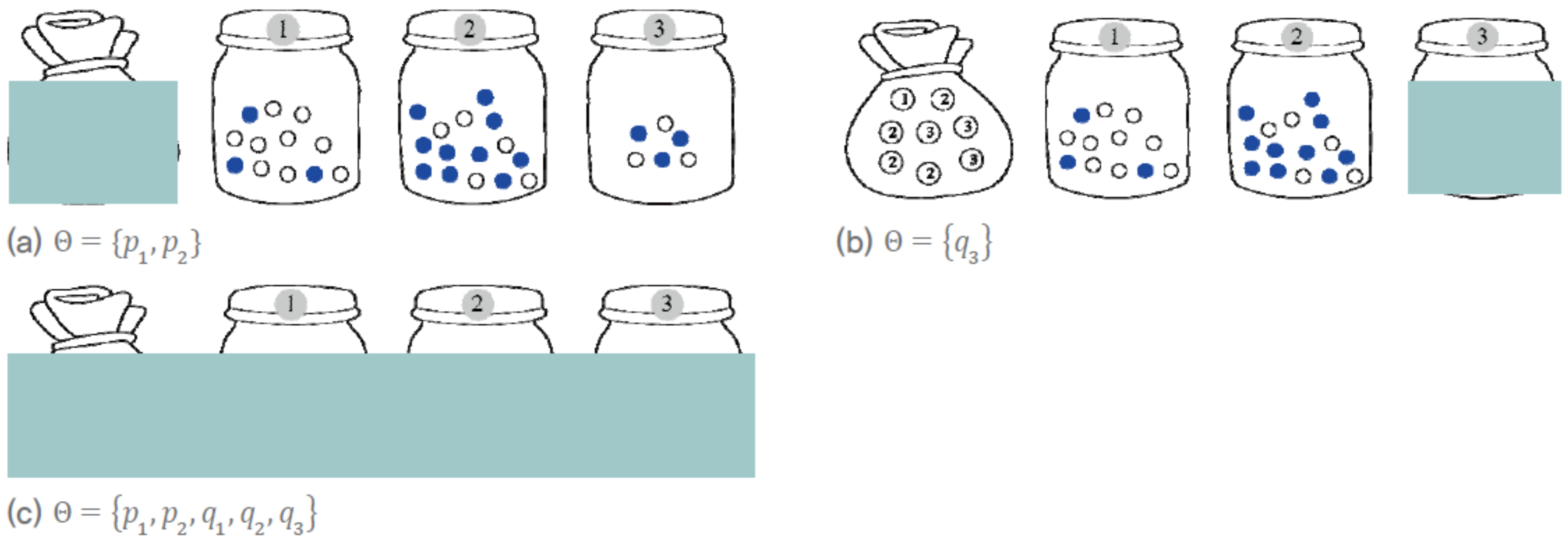
5.5 Maximum Likelihood Estimation
최대 우도

최대 우도

최대 우도법이란?

- 일부 또는 전부가 가려진 상황에서 가려진 곳에 있는 매개변수 추정

- 카드 ①, ②의 확률과 p1, p2 추정



- “데이터 X가 주어졌을 때, X를 발생시켰을 가능성을 최대로 하는 매개변수 theta = {q3}의 값을 찾아라.”

최대 우도

최대 우도법이란?

- 우도를 최대화하는 해를 구한다는 뜻 - 최대 우도 추정 MLE

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

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

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X};\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\mathrm{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\end{aligned}$$

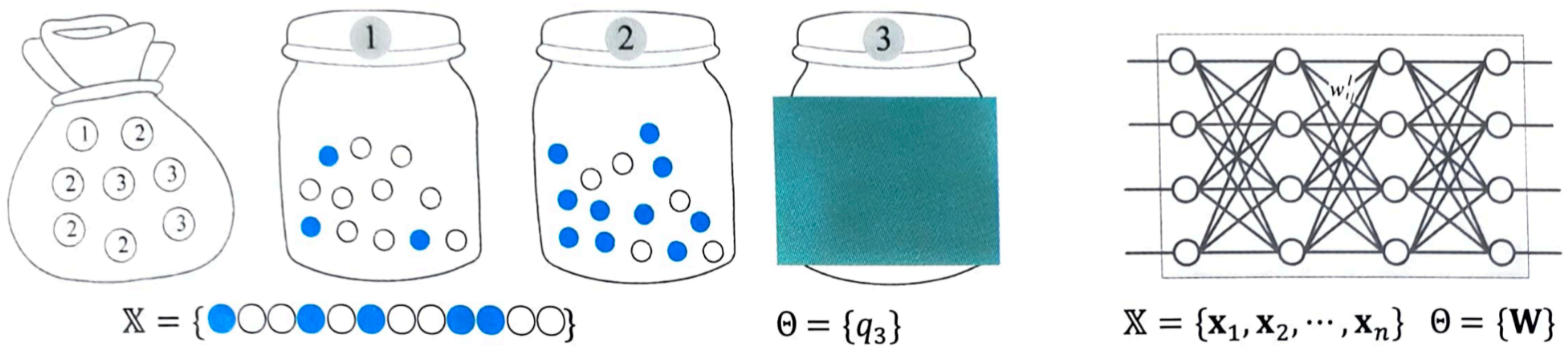
수치적인 문제 일으킬 수 있음

최대 로그우도 추정: $\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log P(\mathbb{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta)$ (2.34)

최대 우도

기계 학습에 적용

딥러닝에서 최대 우도법: $\hat{\mathbf{W}} = \underset{\mathbf{W}}{\mathrm{argmax}}\, P(\mathbb{X}|\mathbf{W})$



(a) 간단한 확률 실험

(b) 딥러닝

- 공의 색깔 집합, 추정해야 할 매개변수는 3번 병의 파란색 공의 확률 Q3 $\hat{q}_3 = \underset{q_3}{\mathrm{argmax}}\, P(\mathbb{X}|q_3)$
- 딥러닝에 적용 - 훈련집합 X = {x1, x2, …, xn}, 추정할 매개변수 - 신경망의 가중치집합 **W.** 48개의 매개변수.