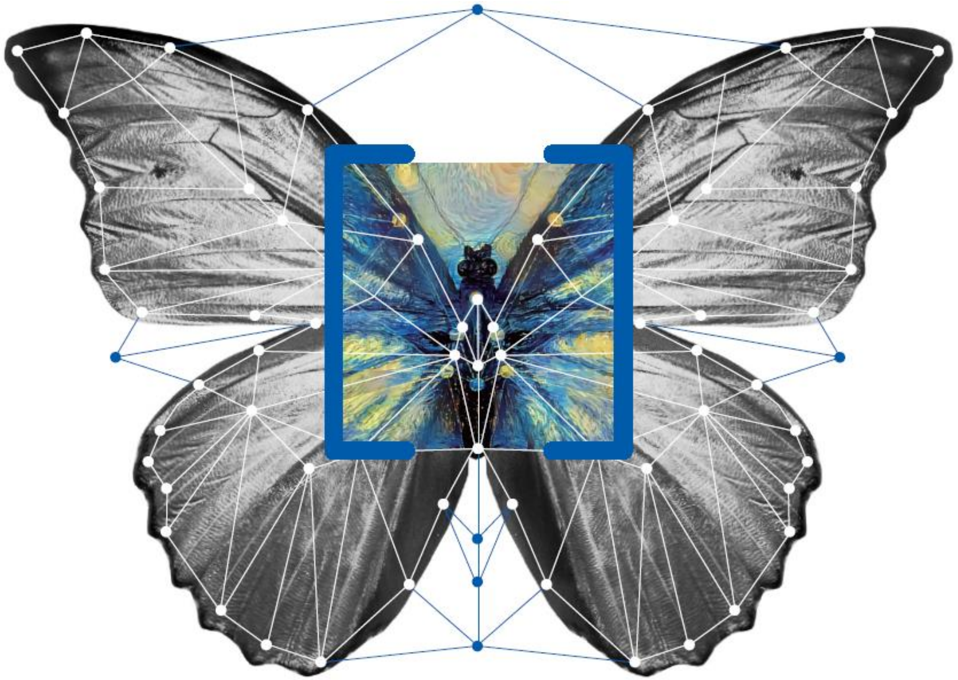


MACHINE LEARNING 기계 학습

2장. 기계 학습과 수학

PREVIEW

기계 학습에서 수학의 역할

- 수학은 목적함수를 정의하고, 목적함수가 최저가 되는 점을 찾아주는 최적화 이론 제공
- 사람은 알고리즘을 설계하고 데이터를 수집함

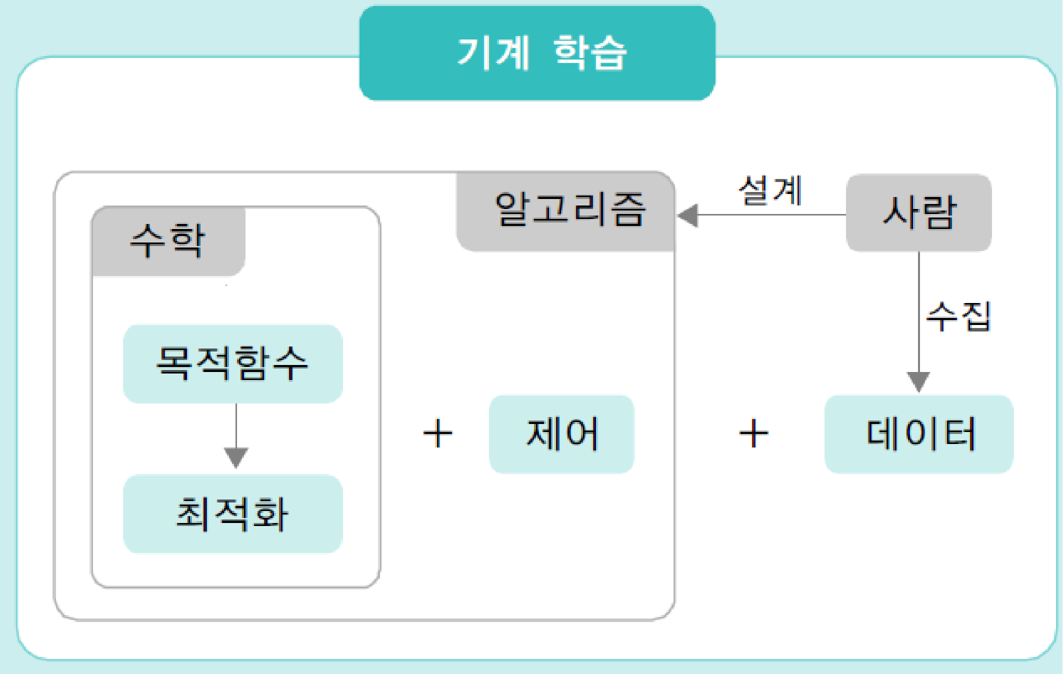


그림 2-1 기계 학습의 개념도

각 절에서 다루는 내용

- 2.1절: 선형대수를 다룬다.
- 2.2절: 확률과 통계를 다룬다.
- 2.3절: 최적화 이론을 다룬다.

- 선형대수: 이 분야의 개념을 이용하면 학습 모델의 매개변수집합, 데이터, 선형연산의 결합 등을 행렬 또는 텐서로 간결하게 표현할 수 있다. 데이터를 분석하여 유용한 정보를 알아내거나 특징 공간을 변환하는 등의 과업을 수행하는 데 핵심 역할을 한다.
- 확률과 통계: 데이터에 포함된 불확실성을 표현하고 처리하는 데 활용한다. 베이즈 이론과 최대 우도 기법을 이용하여 확률 추론을 수행한다.

2.1 선형대수

- 벡터와 행렬
- 선형결합과 벡터공간
- 역행렬

벡터와 행렬

벡터

- 샘플을 특징 벡터로feature vector 표현
- 예) Iris 데이터에서 꽃받침의 길이, 꽃받침의 너비, 꽃잎의 길이, 꽃잎의 너비라는 4개의 특징이 각각 5.1, 3.5, 1.4, 0.2인 샘플

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}$$

- 여러 개의 특징 벡터를 첨자로 구분

$$\mathbf{x}_1 = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 4.9 \\ 3.0 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 4.7 \\ 3.2 \\ 1.3 \\ 0.2 \end{pmatrix}, \cdots, \mathbf{x}_{150} = \begin{pmatrix} 5.9 \\ 3.0 \\ 5.1 \\ 1.8 \end{pmatrix}$$

벡터와 행렬

행렬

- 여러 개의 벡터를 담음
- 훈련집합을 담은 행렬을 설계행렬이라 부름
- 예) Iris 데이터에 있는 150개의 샘플을 설계 행렬 $\mathbf{X}$로 표현

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

행row

열column

벡터와 행렬

- 행렬 $\mathbf{A}$의 전치행렬 $\mathbf{A}^\mathrm{T}$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad \mathbf{A}^\mathrm{T} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}$$

예를 들어, $\mathbf{A} = \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix}$라면 $\mathbf{A}^\mathrm{T} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \\ 1 & 2 \end{pmatrix}$

  - Iris의 설계 행렬을 전치행렬 표기에 따라 표현하면,

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

벡터와 행렬

- 행렬을 이용하면 수학을 간결하게 표현할 수 있음
  - 예) 다항식의 행렬 표현

$$
\begin{aligned}
f(\mathbf{x}) &= f(x_1, x_2, x_3) \\
&= 2x_1x_1 - 4x_1x_2 + 3x_1x_3 + x_2x_1 + 2x_2x_2 + 6x_2x_3 - 2x_3x_1 + 3x_3x_2 + 2x_3x_3 + 2x_1 + 3x_2 - 4x_3 + 5 \\
&= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 2 & -4 & 3 \\ 1 & 2 & 6 \\ -2 & 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 2 & 3 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + 5 \\
&= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c
\end{aligned}
$$

- 특수한 행렬들

정사각행렬 $\begin{pmatrix} 2 & 0 & 1 \\ 1 & 21 & 5 \\ 4 & 5 & 12 \end{pmatrix}$, 대각행렬 $\begin{pmatrix} 50 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 8 \end{pmatrix}$,

단위행렬 $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, 대칭행렬 $\begin{pmatrix} 1 & 2 & 11 \\ 2 & 21 & 5 \\ 11 & 5 & 1 \end{pmatrix}$

벡터와 행렬

행렬 연산

- 행렬 곱셈 $\mathbf{C} = \mathbf{AB}$, 이때 $c_{ij} = \sum_{k=1,s} a_{ik} b_{kj}$ (2.1)

2*3 행렬 $\mathbf{A} = \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix}$와 3*3행렬 $\mathbf{B} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 0 & 5 \\ 4 & 5 & 1 \end{pmatrix}$을 곱하면 2*3 행렬 $\mathbf{C} = \mathbf{AB} = \begin{pmatrix} 14 & 5 & 24 \\ 13 & 10 & 27 \end{pmatrix}$

  - 교환법칙 성립하지 않음: $\mathbf{AB} \neq \mathbf{BA}$
  - 분배법칙과 결합법칙 성립: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}$이고 $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$

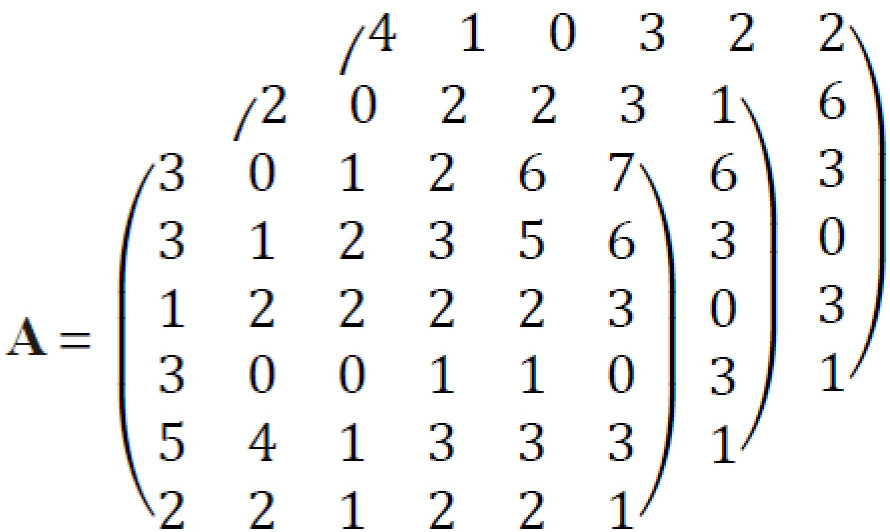
- 아다마르 곱
  - 같은 크기의 두 행렬의 각 성분을 곱하는 연산
  - 일반 행렬곱은 m x n과 n x p 의 꼴의 두 행렬을 곱하지만,
    아다마르 곱은 m x n과 m x n의 꼴의 두 행렬을 곱한다.

$$\begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 10 \end{bmatrix} \circ \begin{bmatrix} 8 & 5 \\ 1 & -3 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 8 & 15 \\ -2 & -15 \\ 0 & 70 \end{bmatrix}$$

벡터와 행렬

텐서

- 3차원 이상의 구조를 가진 숫자 배열
- 예) 3차원 구조의 RGB 컬러 영상

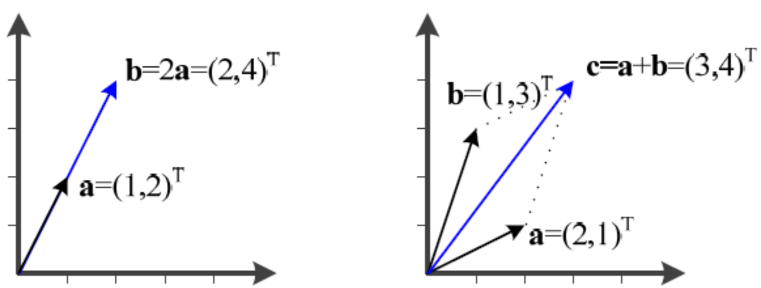


선형결합과 벡터공간

- 벡터
  - 공간상의 한 점으로 화살표 끝이 벡터의 좌표에 해당
- 선형결합이 만드는 벡터공간
  - 기저벡터 **a**와 **b**의 선형결합

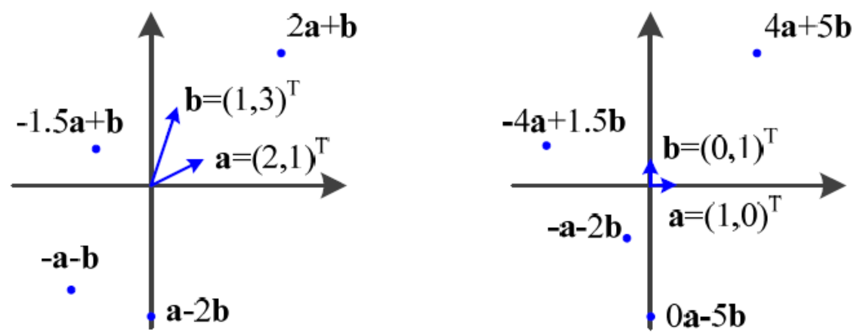
$$\mathbf{c} = \alpha_1 \mathbf{a} + \alpha_2 \mathbf{b} \tag{2.12}$$

  - 선형결합으로 만들어지는 공간을 벡터공간이라 부름



(a) 벡터에 스칼라 곱 (b) 두 벡터의 덧셈

그림 2-6 벡터의 연산

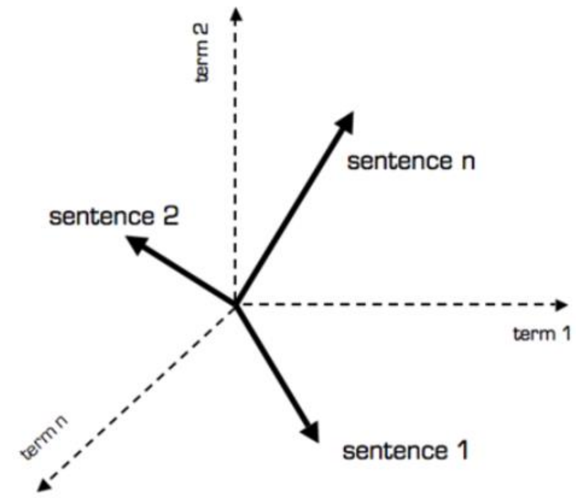


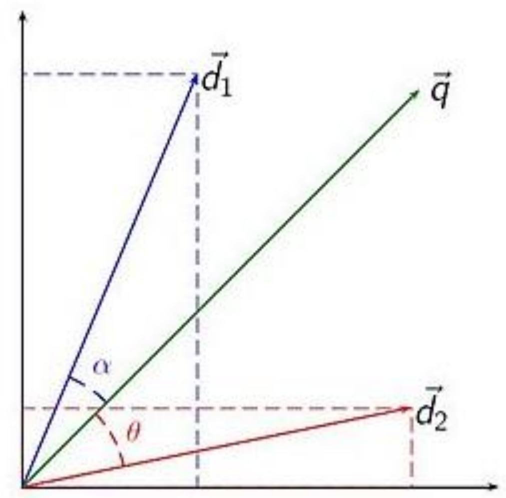
(a) 기저 벡터와 벡터공간 (b) 정규직교 기저 벡터

그림 2-7 벡터공간

유사도

- 유사도와 거리
  - 벡터를 기하학적으로 해석
  - 질의 혹은 문서를 n차원 벡터로 표현
  - 유사한 문서들은 벡터 공간 상에서 가까이 위치할 것이라고 가정
  - 벡터간의 관계로 유사도 계산 : 코사인





$$\cos(\theta) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|}$$

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

유사도 예제

w1 : Bioinformatics
w2 : Biology
w3 : Chemistry
w4 : Enzymes
w5 : Evolution
w6 : Gens
w7 : Genome(s)
w8 : Proteins

문 서

D1 Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins

D2 Proteins. Enzymes. Genes : The Interplay of Chemistry and Biology

D3 Adaptive Evolution of Genes and Genomes
D4 Advanced in Genome Biology : Genes and Genomes
D5 Bioinformatics and Genome Research
D6 Data Analysis in Molecular Biology and Evolution

$$D=\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad q=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

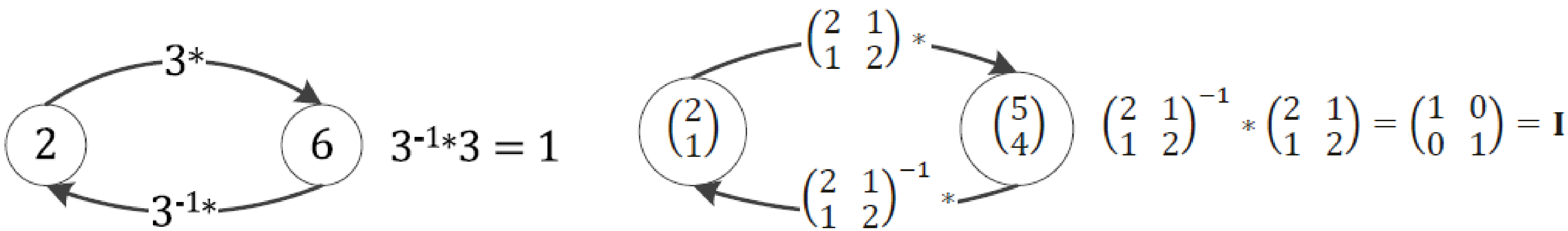
$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\|\|q\|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} d_{ij}^2}\sqrt{\sum_{i=1}^{m} q_i^2}}$$

유사도 계산

$\cos(d_1, q) = 0.408$
$\cos(d_2, q) = 0.316$
$\cos(d_3, q) = 0.816$
$\cos(d_4, q) = 0.866$
$\cos(d_5, q) = 0.050$
$\cos(d_6, q) = 0.000$

역행렬

역행렬의 원리



(a) 역수의 원리 (b) 역행렬의 원리

그림 2-9 역행렬

- 정사각행렬 **A**의 역행렬 $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- 예를 들어, $\begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix}$의 역행렬은 $\begin{pmatrix} 2 & -0.5 \\ -3 & 1 \end{pmatrix}$

역행렬

- 행렬 **A**의 행렬식 *det*(**A**)

$$\left.\begin{aligned} det\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= ad - bc \\ det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} &= aei + bfg + cdh - ceg - bdi - afh \end{aligned}\right\} \quad (2.15)$$

예를 들어 $\begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix}$의 행렬식은 $2*4-1*6=2$

- 기하학적 의미
  - 2차원에서는 2개의 행 벡터가 이루는 평행사변형의 넓이
  - 3차원에서는 3개의 행 벡터가 이루는 평행사각기둥의 부피

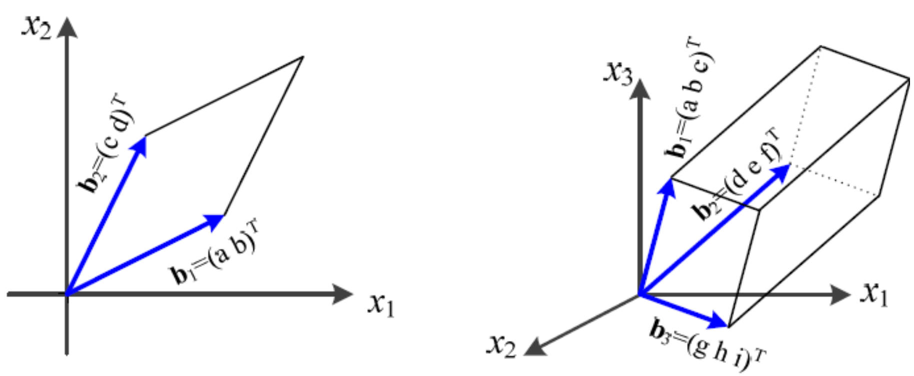


그림 2-10 행렬식의 기하학적 해석

행렬 분해 – 고윳값 분해

분해란?

- 정수 3717은 특성이 보이지 않지만, 3*3*7*59로 소인수 분해를 하면 특성이 보이듯이, 행렬도 분해하면 여러모로 유용함

고윳값과 고유 벡터

- 고유 벡터(eigen vector) $\mathbf{v}$와 고윳값(eigen value) λ

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

- 예를 들어, $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \end{pmatrix}$이고 $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1\begin{pmatrix} 1 \\ -1 \end{pmatrix}$이므로, $\lambda_1 = 3, \lambda_2 = 1$이고 $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

- 정리
  - "고유 벡터 v에 정방 행렬 A를 곱하면 스칼라 λ 를 곱하는 것과 동일한 효과"

행렬 분해 – 고윳값 분해 (의미에 집중!!)

- 고윳값 분해eigen value decomposition

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1} \qquad (2.21)$$

  - $\mathbf{Q}$는 $\mathbf{A}$의 고유 벡터를 열에 배치한 행렬이고 $\boldsymbol{\Lambda}$는 고윳값을 대각선에 배치한 대각행렬
  - 예를 들어, $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}$
  - 고윳값 분해는 정사각행렬에만 적용 가능한데, 기계 학습에서는 정사각행렬이 아닌 경우의 분해도 필요하므로 고윳값 분해는 한계를 가짐
    - N * N 행렬에 대해 N 개의 고윳값-교유벡터 쌍이 존재

행렬 분해 – 특잇값 분해 (의미에 집중!!)

- $n*m$ 행렬 **A**의 특잇값 분해SVD(singular value decomposition)

$$\mathbf{A} = \mathbf{U\Sigma V^T} \tag{2.22}$$

  - 왼쪽 특이행렬 **U**는 $\mathbf{AA^T}$의 고유 벡터를 열에 배치한 $n*n$ 행렬
  - 오른쪽 특이행렬 **V**는 $\mathbf{A^TA}$의 고유 벡터를 열에 배치한 $m*m$ 행렬
  - $\mathbf{\Sigma}$는 $\mathbf{AA^T}$의 고윳값의 제곱근을 대각선에 배치한 $n*m$ 대각행렬

예를 들어, **A**를 4*3 행렬이라고 했을 때 다음과 같이 특잇값 분해가 된다.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 2 \\ 3 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} -0.1914 & -0.2412 & 0.1195 & -0.9439 \\ -0.5144 & 0.6990 & -0.4781 & -0.1348 \\ -0.6946 & -0.6226 & -0.2390 & 0.2697 \\ -0.4651 & 0.2560 & 0.8367 & 0.1348 \end{pmatrix}$$

$$\begin{pmatrix} 3.7837 & 0 & 0 \\ 0 & 2.7719 & 0 \\ 0 & 0 & 1.4142 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.7242 & -0.4555 & -0.5177 \\ -0.6685 & 0.2797 & 0.6891 \\ 0.1690 & -0.8452 & 0.5071 \end{pmatrix}$$

2.2 확률과 통계

- 확률 기초
- 베이즈 정리와 기계 학습
- 최대 우도

- 기계 학습이 처리할 데이터는 불확실한 세상에서 발생하므로, 불확실성을 다루는 확률과 통계를 잘 활용해야 함
  - 현실 세계에서 100% 확실한 정보나 지식을 얻는다는 것은 상당히 힘들다. 따라서 우리가 사용하는 정보나 지식은 항상 어느 정도 불확실하다고 할 수 있다.

불확실성의 예

- 집에서 공항까지 가는 문제
  - 중간에 자동차가 고장 날 수도 있고,
  - 엄청난 차량 정체를 겪을 수도 있다.
  - 공항이 몹시 혼잡하여 수속에 시간이 많이 걸릴 수도 있다.

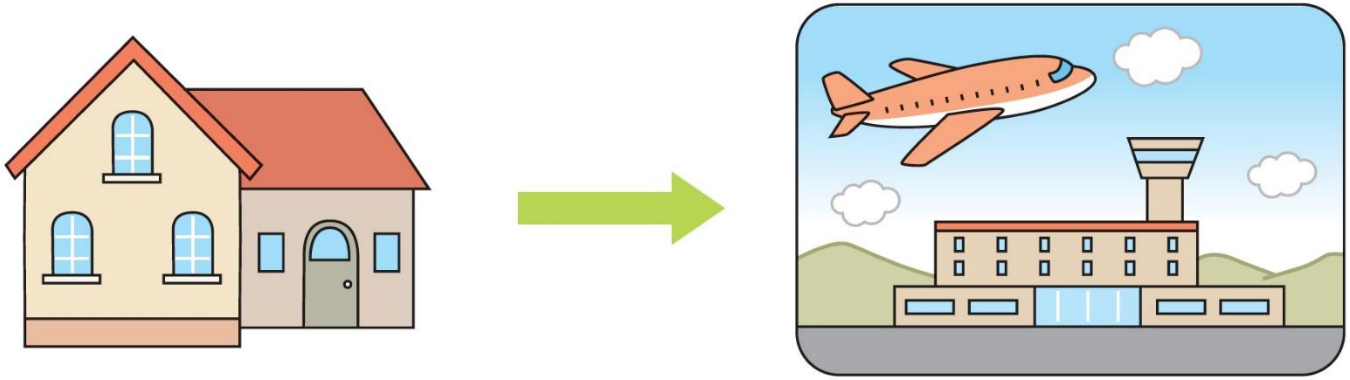
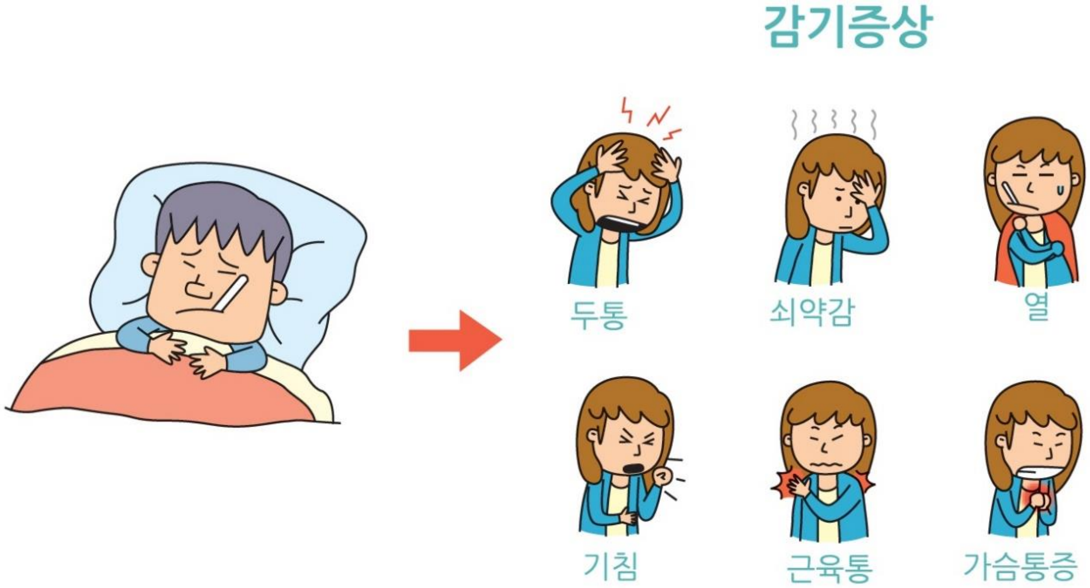


그림 7-1 불확실성이 발생하는 문제

불확실성의 예

- 내과에서 환자를 진단하는 전문가 시스템
  - 규칙: 감기이다 → 열이 있다. -> 100% 확실할 수 없다.



확률 기초

- 확률변수random variable
  - 예) 윷

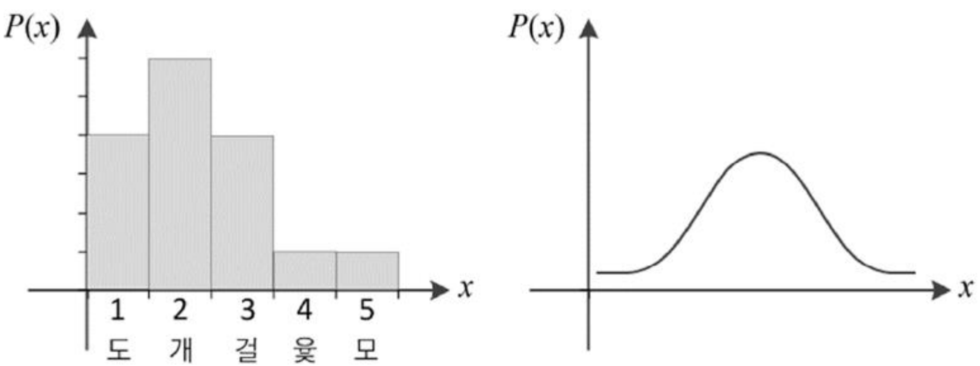
그림 2-13 윷을 던졌을 때 나올 수 있는 다섯 가지 경우(왼쪽부터 도, 개, 걸, 윷, 모)

  - 다섯 가지 경우 중 한 값을 갖는 확률변수 x
  - x의 정의역은 {도, 개, 걸, 윷, 모}

확률 기초

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

확률벡터 random vector

- 예) Iris에서 확률벡터 $\mathbf{x}$는 4차원 $\mathbf{x}=(x_1,x_2,x_3,x_4)^{\mathrm{T}}=($꽃받침 길이, 꽃받침 너비$_1$, 꽃잎

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$\}$

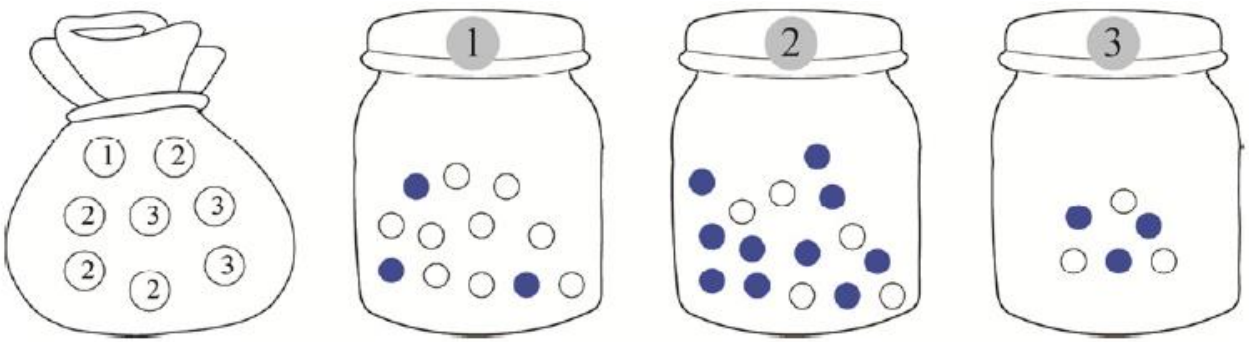


그림 2-15 확률 실험

확률 기초

곱 규칙과 합 규칙

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

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

- 곱 규칙

$$\text{곱 규칙: } P(y,x)=P(x|y)P(y) \tag{2.23}$$

- 하얀 공이 뽑힐 확률

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

- 합 규칙

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

2.2.2 베이즈 정리와 기계 학습

- 베이즈 정리 (식 (2.26))

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

$$\longrightarrow \quad P(y|x) = \frac{P(x|y)P(y)}{P(x)} \qquad (2.26)$$

  - 다음 질문을 식 (2.27)로 쓸 수 있음

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

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|x) \qquad (2.27)$$

베이즈 정리와 기계 학습

- 베이즈 정리 (식 (2.26))
  - 베이즈 정리를 적용하면, $\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x = 하양) = \underset{y}{\operatorname{argmax}} \frac{P(x = 하양|y)P(y)}{P(x = 하양)}$
  - 세 가지 경우에 대해 확률을 계산하면,

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

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

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

⟶ ③번 병일 확률이 가장 높음

- 베이즈 정리의 해석

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

베이지 정리(예: 카드 게임)

- 우리가 한 장의 카드를 가지고 있는데 카드에 얼굴이 그려져 있다는 것만 알고 있다. 이 상황에서 이 카드가 킹(king)일 확률을 추론해보자. 즉 P(King | Face) 확률을 계산해보는 것이다.

사전확률 (prior)

가능도 (likelihood)

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

사후확률 (posterior)

사전확률

카드 게임

$$P(King|Face) = \frac{P(Face|King)P(King)}{P(Face)}$$

1) 카드에 왕이 그려져 있을 확률 : P(King) = 4/52 = 1/13
2) 모든 킹 카드에는 얼굴이 그려진 카드이므로 P(Face|King) = 1
3) P(Face)는 52장의 카드 중에서 얼굴이 그려진 비율 (3*4)/52 = 3/13

$$P(King|Face) = \frac{P(Face|King)P(King)}{P(Face)} = \frac{13}{3}\frac{1}{13} = \frac{1}{3}$$

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)

$$\hat{y} = \underset{y}{\text{argmax}}\, P(y|\mathbf{x}) \qquad (2.29)$$

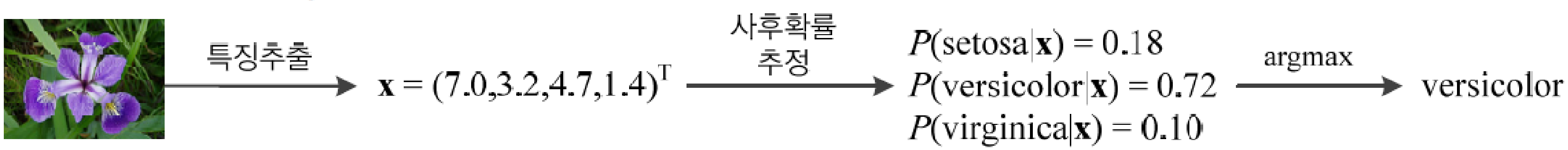


그림 2-16 붓꽃의 부류 예측 과정

- 사후확률 $P(y|\mathbf{x})$를 직접 추정하는 일은 아주 단순한 경우를 빼고 불가능
- 따라서 베이즈 정리를 이용하여 추정함
  - 사전확률은 식 (2.30)으로 추정
  - 우도는 밀도 추정 기법으로 추정

$$\text{사전확률:}\quad P(y = c_i) = \frac{n_i}{n} \qquad (2.30)$$

최대 우도

- 매개변수 Θ를 모르는 상황에서 매개변수를 추정하는 문제

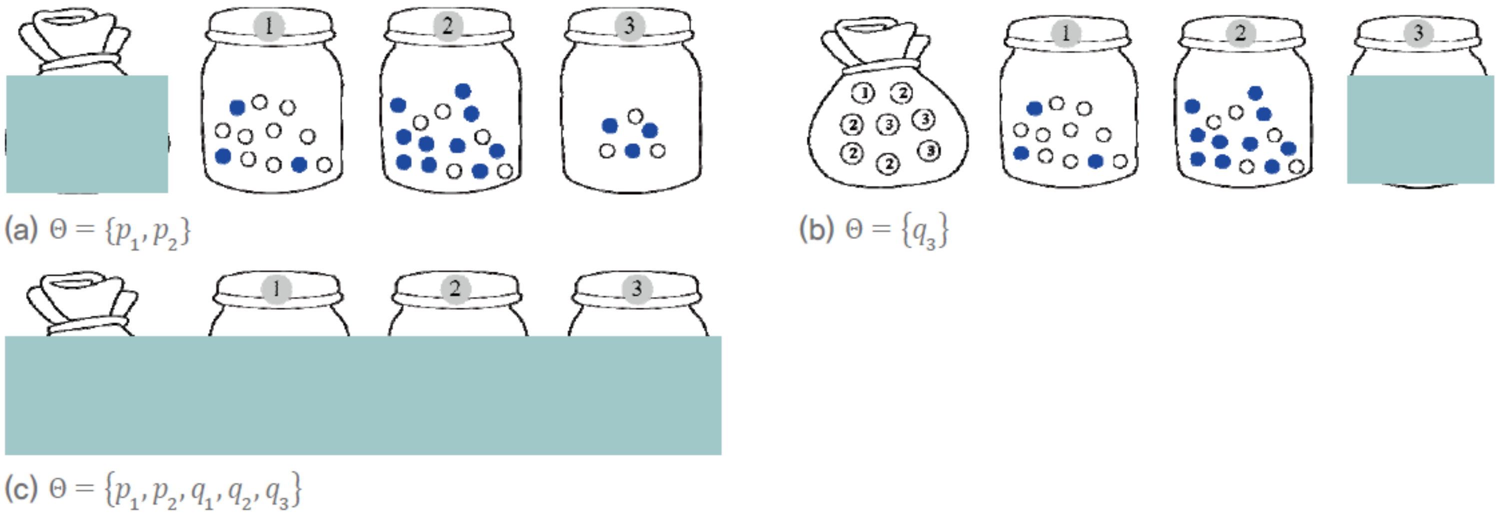


그림 2-17 매개변수가 감추어진 여러 가지 상황

  - 예) [그림 2-17(b)] 상황

    데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}

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

최대 우도

최대 우도법

- [그림 2-17(b)] 문제를 수식으로 쓰면,

$$\hat{q}_3 = \underset{q_3}{\operatorname{argmax}}\, P(\mathbb{X}|q_3) \qquad (2.31)$$

- 일반화 하면,

$$\text{최대 우도 추정:} \quad \hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, P(\mathbb{X}|\Theta) \qquad (2.32)$$

- 수치 문제를 피하기 위해 로그 표현으로 바꾸면,

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

정보이론

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

- 사건(메시지) e_i의 정보량 (단위: 비트 또는 나츠)

$$h(e_i) = -\log_2 P(e_i) \quad \text{또는} \quad h(e_i) = -\log_e P(e_i) \qquad (2.44)$$

엔트로피

- 확률변수 x의 불확실성을 나타내는 엔트로피

$$\text{이산 확률분포} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i) \quad \text{또는} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_e P(e_i) \qquad (2.45)$$

$$\text{연속 확률분포} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x) \quad \text{또는} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_e P(x) \qquad (2.46)$$

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?

2.3 최적화 이론

- 미분
- 편미분

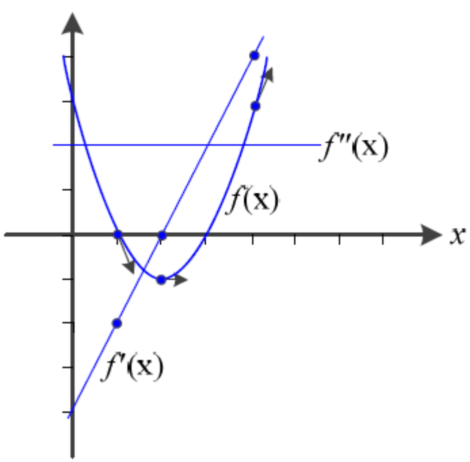
미분

미분에 의한 최적화

- 미분의 정의

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}, \qquad f''(x) = \lim_{\Delta x \to 0} \frac{f'(x+\Delta x) - f'(x)}{\Delta x} \qquad (2.51)$$

- 1차 도함수 $f'(x)$는 함수의 기울기, 즉 값이 커지는 방향을 지시함
- 따라서 $-f'(x)$ 방향에 목적함수의 최저점이 존재
- 경사 하강 알고리즘의 핵심 원리



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

$y' = f'(x) = 2x - 4$

그림 2-24 간단한 미분 예제

편미분

편미분

- 변수가 여러 개인 함수의 미분
- 여러 가지 표기: $\nabla f, \frac{\partial f}{\partial \mathbf{x}}, \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)^{\mathrm{T}}$
- 예1)

$$\left.\begin{array}{r} f(\mathbf{x}) = f(x_1, x_2) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2 \\ \nabla f = f'(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)^{\mathrm{T}} = (2x_1^5 - 8.4x_1^3 + 8x_1 + x_2, 16x_2^3 - 8x_2 + x_1)^{\mathrm{T}} \end{array}\right\} \quad (2.52)$$

예2) $z = x^2 + y^2$ 일 때 편미분의 값은?

기계 학습에서 편미분

- 매개변수 집합 Θ에 많은 변수가 있으므로 편미분을 많이 사용