인 공 지 능

[기계 학습과 수학 II]

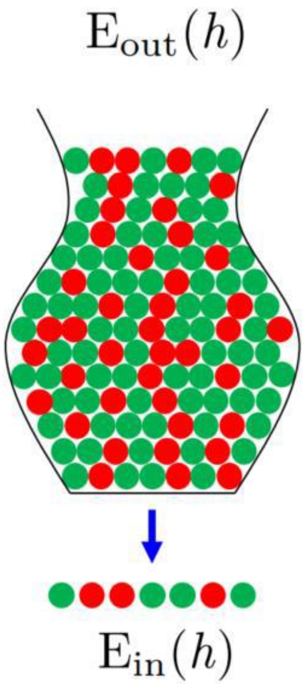
jaekoo@kookmin.ac.kr

국민대학교
KOOKMIN UNIVERSITY

2.2 확률과 통계

2.2 확률과 통계

- 기계 학습이 처리할 데이터는 불확실한 세상에서 발생하므로, 불확실성uncertainty을 다루는 확률과 통계를 잘 활용해야 함



2.2.1 확률 기초

- **확률변수**random 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 정의역domain: {도, 개, 걸, 윷, 모}

2.2.1 확률 기초

- **확률분포**probability distribution
  - 확률질량함수probability mass function: 이산discrete 확률 변수

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

  - 확률밀도함수probability density function: 연속continuous 확률 변수

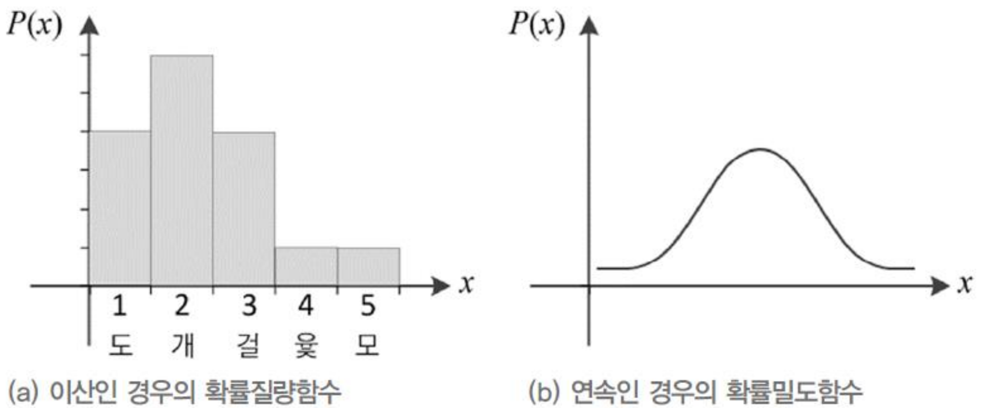


그림 2-14 확률분포

- **확률벡터**random vector
  - 확률변수를 요소로 가짐
    - 예) Iris에서 $\mathbf{x}$는 4차원 확률 벡터 $\mathbf{x}=(x_1,x_2,x_3,x_4)^{\mathrm{T}}=(\text{꽃받침 길이},\text{꽃받침 너비},\text{꽃잎 길이},\text{꽃잎 너비})^{\mathrm{T}}$

2.2.1 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$\}$

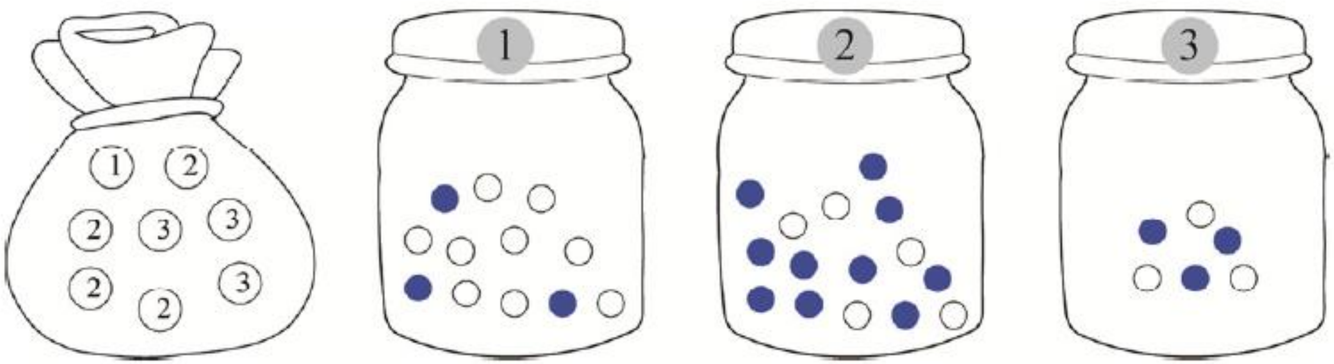


그림 2-15 확률 실험

2.2.1 확률 기초

- 곱 (AND) 규칙product rule과 합 (OR) 규칙sum rule
  - ①번 카드를 뽑을 확률은 $P(y=①) = P(①) = 1/8$
  - 예) 카드는 ①번, 공은 하얀 공일 확률: $P(y=①, x=\text{하양}) = P(①, \text{하양})$ ← 결합확률joint probability

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

← 조건부 확률conditional probability에 의한 결합확률 계산

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

  - 예) 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 규칙과 곱 규칙 의한 주변확률marginal probability 계산

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

2.2.1 확률 기초

- 조건부 확률

$$P(\mathrm{y}=y \mid \mathrm{x}=x)=\frac{P(\mathrm{y}=y, \mathrm{x}=x)}{P(\mathrm{x}=x)}$$

- 확률의 연쇄 법칙chain rule

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)})=P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

- 독립independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x}=x, \mathrm{y}=y)=p(\mathrm{x}=x) p(\mathrm{y}=y)$$

- 조건부 독립conditional independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z)=p(\mathrm{x}=x \mid \mathrm{z}=z) p(\mathrm{y}=y \mid \mathrm{z}=z)$$

- 기대값expectation

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]=\sum_{x} P(x) f(x) \quad \rightarrow \quad \text{linearity of expectations:}$$

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x)+\beta g(x)]=\alpha \mathbb{E}_{\mathrm{x}}[f(x)]+\beta \mathbb{E}_{\mathrm{x}}[g(x)]$$

2.2.2 베이즈 정리와 기계 학습

- 베이즈 정리Bayes's rule (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.”

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

2.2.2 베이즈 정리와 기계 학습

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

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

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

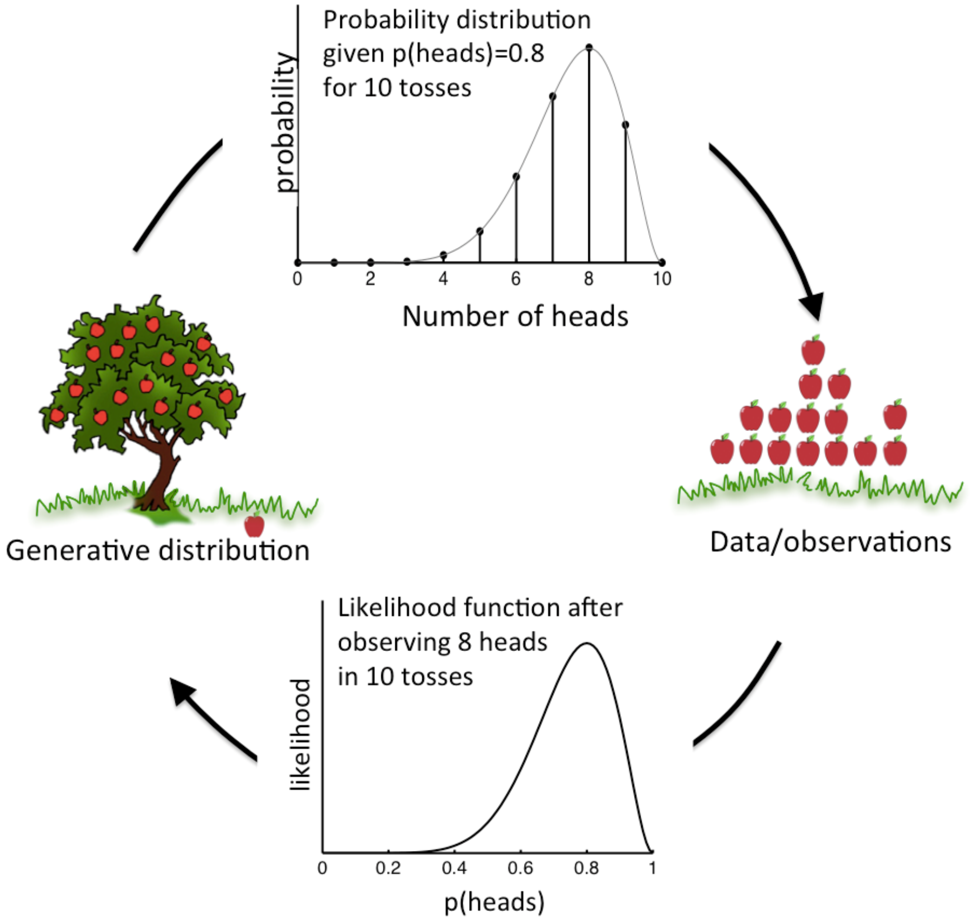
$$P(③|\text{하양}) = \frac{P(\text{하양}|③)P(③)}{P(\text{하양})} = \frac{\frac{3}{6}\frac{3}{8}}{\frac{43}{96}} = \frac{18}{43} \longrightarrow$$ ③번 병일 확률이 가장 높음

- 베이즈 정리의 해석
  - 사후posteriori확률=우도likelihood확률*사전prior확률 →

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

2.2.2 베이즈 정리와 기계 학습

- 확률과 우도

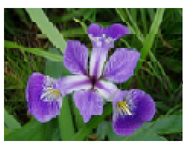


2.2.2 베이즈 정리와 기계 학습

기계 학습에 적용

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

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



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

그림 2-16 붓꽃의 부류 예측 과정

- 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 사전확률: $P(y = c_i) = \dfrac{n_i}{n}$
  - 우도확률는 6.4절의 밀도 추정density estimation 기법으로 추정

2.2.3 최대 우도

- 매개변수 (모수)parameter Θ를 모르는 상황에서 매개변수를 추정하는 문제

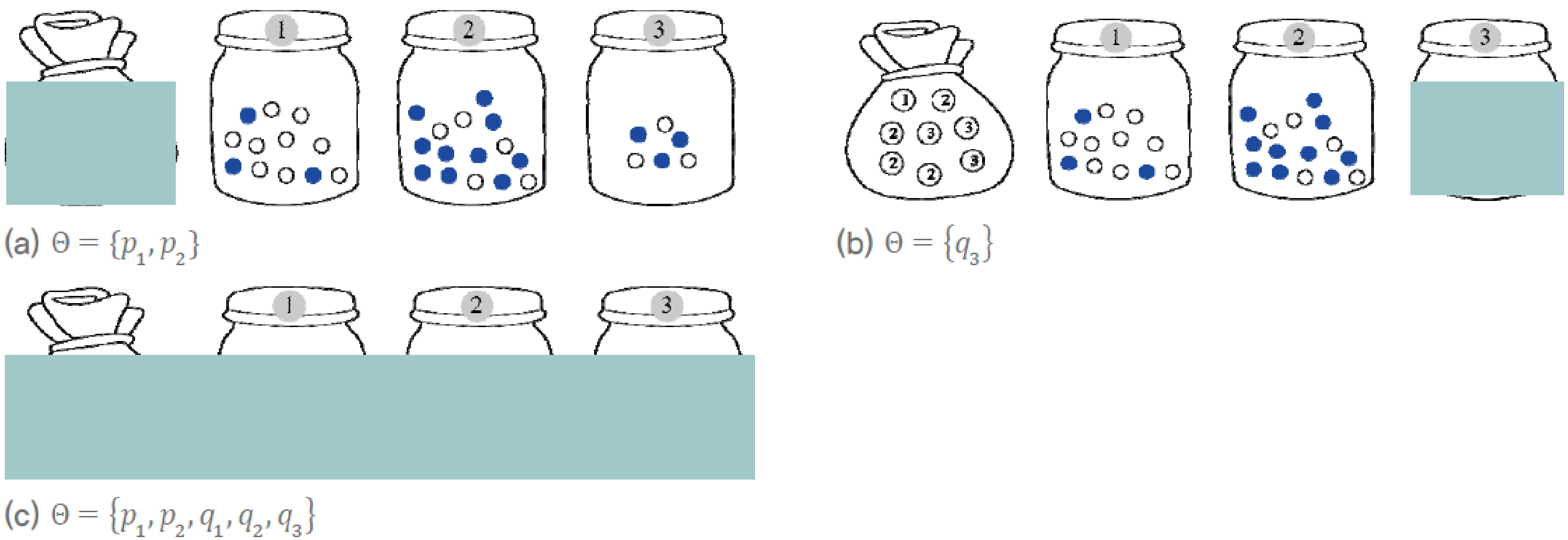


그림 2-17 매개변수가 감추어진 여러 가지 상황

- 예) [그림 2-17(b)] 상황

  관측된 데이터집합 $\mathbb{X}$={●○○●○●○○●●○○}라 할 때,

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

2.2.3 최대 우도

- 최대 우도maximum likelihood
  - 어떤 확률변수의 관찰된 값들을 토대로 그 확률변수의 매개변수를 구하는 방법
  - [그림 2-17(b)] 문제를 수식으로 쓰면,

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

  - 일반화 하면,

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

  - 수치 문제를 피하기 위해 로그 표현으로 바꾸면,

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

    - 단조 증가하는 로그 함수를 이용하여 계산 단순화

2.2.4 평균과 분산

- 데이터의 요약 정보로서 평균mean과 분산variance

평균 $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

분산 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ → $\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$

- 평균 벡터(치우침 정도)와 공분산 행렬covariance matrix (확률변수의 상관정도)

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

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

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

2.2.4 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.

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

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

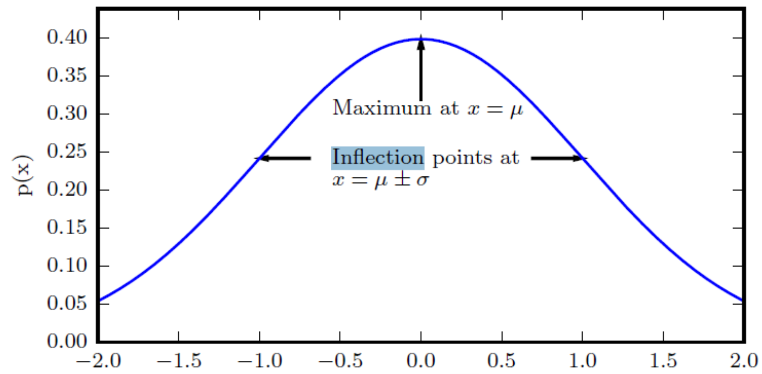
나머지 7개 샘플도 같은 계산을 한 다음, 결과를 모두 더하고 8로 나누면 다음과 같은 공분산 행렬을 얻는다.

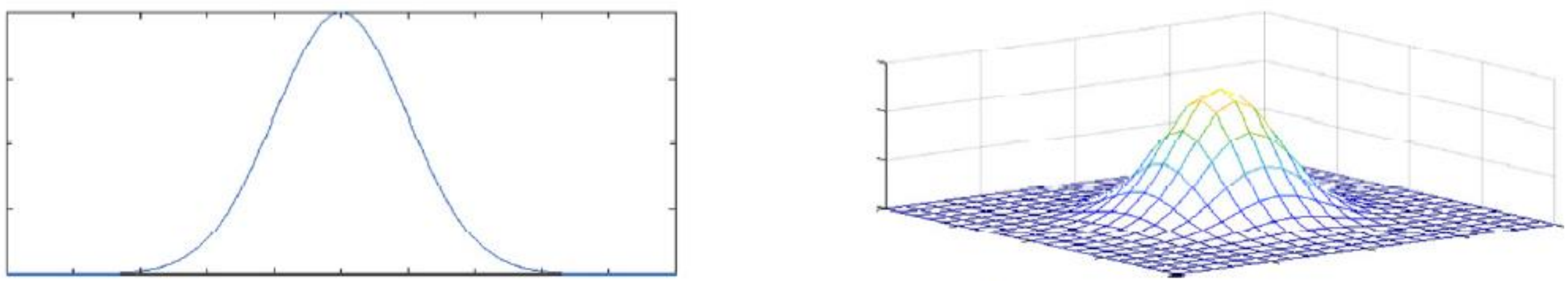
$$\boldsymbol{\Sigma}=\begin{pmatrix}0.0661 & 0.0527 & 0.0181 & 0.0083\\0.0527 & 0.0736 & 0.0181 & 0.0130\\0.0181 & 0.0181 & 0.0125 & 0.0056\\0.0083 & 0.0130 & 0.0056 & 0.0048\end{pmatrix}$$

2.2.5 유용한 확률분포

- 가우시안 분포Gaussian distribution
  - 평균 μ와 분산 σ^2으로 정의

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





(a) 1차원 (b) 2차원

그림 2-19 가우시안 분포

  - 다차원 가우시안 분포: 평균벡터 $\mathbf{\mu}$와 공분산행렬 $\mathbf{\Sigma}$로 정의

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

2.2.5 유용한 확률분포

■ 베르누이 분포Bernoulli 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 분포Binomial distribution

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

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

- 확률질량함수

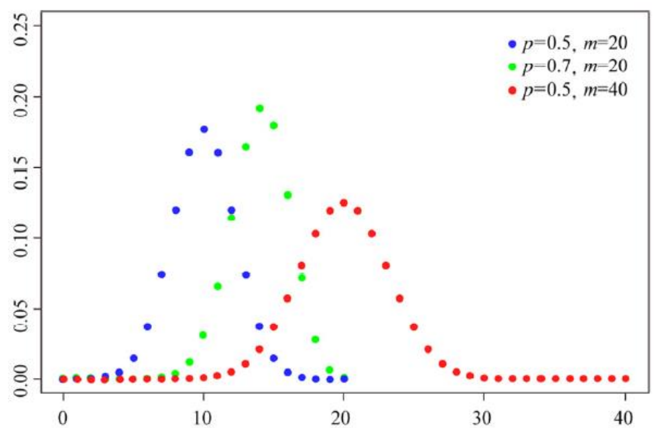
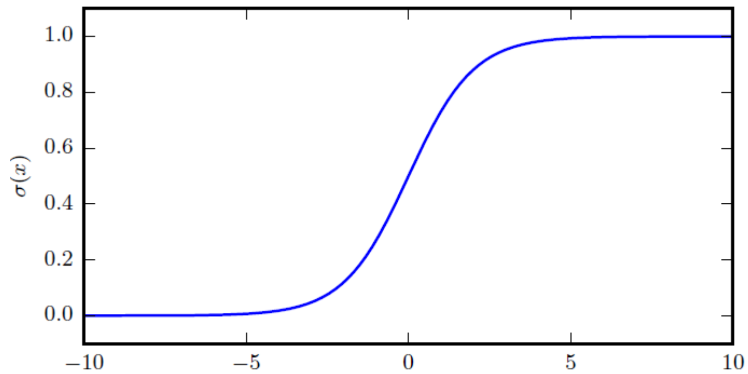


그림 2-20 이항 분포

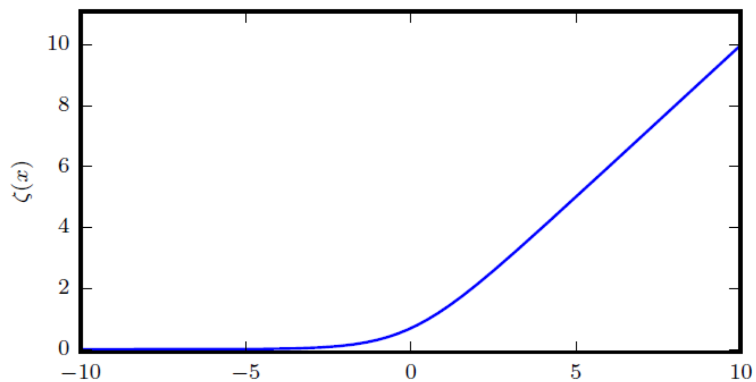
2.2.5 유용한 확률분포

확률 분포와 연관된 유용한 함수들

- 로지스틱 시그모이드 함수logistic sigmoid function
  - 일반적으로 베르누이 분포의 매개변수를 조정을 통해 얻어짐



- 소프트플러스 함수softplus function
  - 정규 분포의 매개변수의 조정을 통해 얻어짐



KMU 국민대학교 KOOKMIN UNIVERSITY

2.2.5 유용한 확률분포

- 지수 분포exponential distribution

$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp(-\lambda x)$$

- 라플라스 분포Laplace distribution

$$\text{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

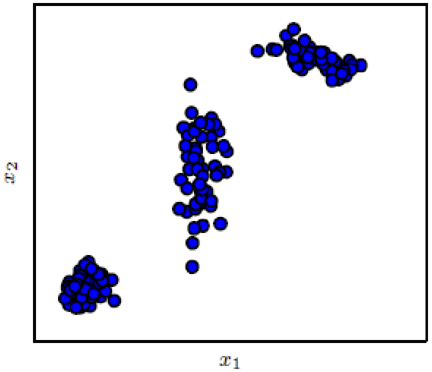
- 디랙 분포Dirac distribution

$$p(x) = \delta(x-\mu)$$

- 혼합 분포들Mixture Distributions

$$P(\mathrm{x}) = \sum_i P(\mathrm{c}=i)P(\mathrm{x} \mid \mathrm{c}=i)$$

  - 3개의 요소를 가진 가우시안 혼합 분포 예 ← 가우시안 혼합 모델 추정 가능



2.2.5 유용한 확률분포

- 변수 변환change of variables
  - 기존 확률변수를 새로운 확률 변수로 바꾸는 것
  - 변환 $y=g(x)$와 가역성을 가진 g에 의해 정의되는 x, y 두 확률변수를 가정할 때, 두 확률 변수는 다음과 같이 상호 정의될 수 있음

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left| \det \left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}} \right) \right|$$

  - 예) 확률변수 x의 확률질량함수가 다음과 같을 때,

$$\left(\frac{4}{5}\right)\left(\frac{1}{5}\right)^{x-1},\ x=1,2,\ \ldots$$

새로운 확률변수$y=x^2$의 확률질량함수는 다음과 같이 정의됨

$$y=x^2 \Rightarrow x=\sqrt{y}$$

$$f(x)=f(\sqrt{y})=\left(\frac{4}{5}\right)\left(\frac{1}{5}\right)^{\sqrt{y}-1}=g(y) \longrightarrow g(y)=\begin{cases}\left(\frac{4}{5}\right)\left(\frac{1}{5}\right)^{\sqrt{y}-1} & ,\ y=1,\ 4,\ 9,\ \ldots \\ 0 & ,\ \text{elsewhere}\end{cases}$$