DAS 511. Introduction to Statistical Learning

Preliminary 1. Some Basics

Jun Song

Department of Statistics
Korea University

표기법: 관측 단위

- 관측 단위 i에 대한 m개 변수들의 관측값들의 벡터는 열벡터로 다음과 같이 표기할 수 있습니다:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix}$$

  즉, x_{ij}는 관측 단위 i에 대한 변수 j의 값을 나타냅니다.
- 관측값들은 일반적으로 소문자로 표기됩니다.
- 단변량 데이터의 경우 $m = 1$입니다.

표기법: 다변량 데이터

- 다변량 데이터 집합은 일반적으로 n개의 행과 m개의 열을 가진 데이터 행렬 $\mathbf{X}$로 표기됩니다 (n = 총 관측 단위 수, m = 변수의 수).

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

- $\mathbf{X}$의 i번째 행은 따라서 $\mathbf{x}_i^T$입니다.

확률변수

- 확률변수(r.v.)는 변수의 각 결과에 실수를 할당하는 매핑입니다.
- 관심 변수가 '성별'이라고 하면, 이 변수의 결과는 '남성'과 '여성'입니다. 확률변수 X는 이러한 결과에 숫자를 할당합니다. 즉,

$$X = \begin{cases} 1 & \text{만약 여성이면} \\ 0 & \text{만약 남성이면} \end{cases}$$

- 값이 알려지지 않은 확률변수는 일반적으로 대문자로 표기됩니다.

확률변수의 기댓값

- X가 확률질량함수 $P(X)$를 가진 이산 확률변수라면, X의 기댓값(평균이라고도 함)은 다음과 같이 정의됩니다:

$$\mu = E[X] = \sum_x xP(X = x)$$

- X가 공간 $\mathbb{R}$에서 정의된 확률밀도함수 $f(x)$를 가진 연속 확률변수라면:

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

- 함수 $u(X)$의 기댓값은 다음과 같이 주어집니다:
  1. $\mu = E[u(X)] = \sum_x u(x)P(X = x)$ (이산 r.v.의 경우)
  2. $\mu = E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)dx$ (연속 r.v.의 경우)

분산, 공분산 및 상관계수

- 확률변수 $X_1, X_2, \ldots, X_m$을 고려해 봅시다.
- X_i의 분산은 다음과 같이 정의됩니다:

$$\mathsf{Var}[X_i] = E[(X_i - E[X_i])^2] = E[X_i^2] - (E[X_i])^2$$

- X_i와 X_j의 공분산은 다음과 같이 정의됩니다:

$$\mathsf{Cov}[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]$$

- X_i와 X_j의 상관계수는 다음과 같이 정의됩니다:

$$\mathsf{Cor}[X_i, X_j] = \frac{\mathsf{Cov}[X_i, X_j]}{\sqrt{\mathsf{Var}[X_i]\mathsf{Var}[X_j]}}$$

- 상관계수는 공분산의 '정규화된' 형태이며, 확률변수들이 단위분산을 가질 때 둘 사이에 같음이 성립합니다.

공분산 행렬

- 확률변수 집합 $\mathbf{X} = (X_1, X_2, \ldots, X_m)$의 분산과 공분산을 행렬을 사용하여 기록하는 것이 편리합니다:

$$\boldsymbol{\Sigma} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}$$

  여기서 $s_{ij} = \mathrm{Cov}[X_i, X_j]$이고 $s_{ii} = \mathrm{Var}[X_i]$입니다.
- 이 행렬을 공분산 행렬이라고 부릅니다.
- $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{X}] = \mathrm{Var}[\mathbf{X}]$

독립성

- 두 확률변수 X_1과 X_2가 독립이라는 것은 다음이 성립할 때입니다:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$$

- 두 독립 확률변수 X_1과 X_2에 대해:

$$E[X_1 X_2] = E[X_1]E[X_2]$$

선형결합

- a와 b가 상수이고 확률변수 X가 기댓값 μ와 분산 σ^2를 가진다고 하면:
  1. $E[aX + b] = aE[X] + b = a\mu + b$
  2. $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X] = a^2\sigma^2$
- X_1과 X_2를 각각 평균 μ_1과 μ_2, 분산 σ_1^2과 σ_2^2를 가진 두 독립 확률변수라 하고, a_1과 a_2를 상수라 하면:
  3. $E[a_1X_1 + a_2X_2] = a_1E[X_1] + a_2E[X_2] = a_1\mu_1 + a_2\mu_2$
  4. $\mathrm{Var}[a_1X_1 + a_2X_2] = a_1^2\mathrm{Var}[X_1] + a_2^2\mathrm{Var}[X_2] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$
- X_1과 X_2가 독립이 아니라면 위의 (c)는 여전히 성립하지만 (d)는 다음으로 대체됩니다:
  5. $\mathrm{Var}[a_1X_1 + a_2X_2] = a_1^2\mathrm{Var}[X_1] + a_2^2\mathrm{Var}[X_2] + 2a_1a_2\mathrm{Cov}[X_1, X_2]$

공분산에 대한 선형결합

a, b, c, d가 상수이고 X, Y, W, Z가 0이 아닌 분산을 가진 확률변수라고 하면:

1. $\text{Cov}[aX+b, cY+d] = ac\text{Cov}[X,Y]$
2. $\text{Cov}[aX+bY, cW+dZ] = ac\text{Cov}[X,W] + ad\text{Cov}[X,Z] + bc\text{Cov}[Y,W] + bd\text{Cov}[Y,Z]$

- 증명: 연습문제 (나머지와 마찬가지로, 정의에서 대수적 조작).

행렬 표현: 기댓값

- 일반적인 $a_1X_1 + \cdots + a_mX_m$ 경우에서도 비슷한 아이디어가 적용됩니다.
- $E[a_1X_1 + a_2X_2 + \cdots + a_mX_m] = a_1\mu_1 + a_2\mu_2 + \cdots a_m\mu_m$임을 기억하세요.
- $\mathbf{a} = (a_1, a_2, \ldots, a_m)^T$를 상수 벡터라 하고
  $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$를 확률변수 벡터라 하면:
- 다음과 같이 쓸 수 있습니다:

$$a_1X_1 + a_2X_2 + \ldots a_mX_m = (a_1, a_2, \cdots, a_m)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} = \mathbf{a}^T\mathbf{X}$$

- 따라서 $E[\mathbf{a}^T\mathbf{X}] = \mathbf{a}^T\boldsymbol{\mu}$로 쓸 수 있습니다. 여기서 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)^T$입니다.

행렬 표현: 분산

- 마찬가지로 $\mathsf{Var}[a_1X_1 + a_2X_2 + \cdots + a_mX_m] = \mathsf{Var}[\mathbf{a}^T\mathbf{X}]$이고,

$$
\begin{aligned}
\mathsf{Var}[\mathbf{a}^T\mathbf{X}] &= a_1^2\mathsf{Var}[X_1] + a_2^2\mathsf{Var}[X_2] + \cdots + a_m^2\mathsf{Var}[X_m] && (1)\\
&\quad + a_1a_2\mathsf{Cov}[X_1, X_2] + \cdots + a_1a_m\mathsf{Cov}[X_1, X_m] && (2)\\
&\quad + \cdots + a_{m-1}a_m\mathsf{Cov}[X_{m-1}, X_m] && (3)\\
&= \sum_{i=1}^{m} a_i^2\mathsf{Var}[X_i] + \sum_{i=1}^{m}\sum_{j\neq i}^{m} a_ia_j\mathsf{Cov}[X_i, X_j] && (4)\\
&= \sum_{i=1}^{m} a_i^2 s_{ii} + \sum_{i=1}^{m}\sum_{j\neq i}^{m} a_ia_js_{ij} && (5)\\
&= \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} && (6)
\end{aligned}
$$

행렬 표현: 공분산

- $U = \mathbf{a}^T\mathbf{X}$이고 $V = \mathbf{b}^T\mathbf{X}$라고 하면:

$$\mathsf{Cov}[U,V] = \sum_{i=1}^{m} a_i b_i s_{ii} + \sum_{i=1}^{m}\sum_{j\neq i}^{m} a_i b_j s_{ij}$$

- 행렬 표기법으로:

$$\mathsf{Cov}[U,V] = \mathbf{a}^T\mathbf{\Sigma}\mathbf{b} = \mathbf{b}^T\mathbf{\Sigma}\mathbf{a}$$

- 증명: 연습문제

예제

- $E[X_1] = 2$, $\text{Var}[X_1] = 4$, $E[X_2] = 0$, $\text{Var}[X_2] = 1$이고 $\text{Cor}[X_1, X_2] = 1/3$라고 하자.
- 연습문제:
  - $X_1 + X_2$의 기댓값과 분산은?
  - $X_1 - X_2$의 기댓값과 분산은?