기수 : 171 이름 : 윤정수

23-2 DSL 정규 세션

기초과제 1 통계적 사고

- ☑ 본 과제는「통계학입문」,「통계방법론」,「선형대수」및「수리통계학(1)」일부에 상응하는 내용의 복습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(✋)를 적극 활용하시고 학회원 간 토론, Slack 의 질의응답을 활용하시어 해결해주십시오. 단, 답안 표절은 금지합니다.
- ☑ 서술형 문제는 ✎ , 코딩 문제는 © 으로 표기가 되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 pdf 로 제출해주시고 코드 문제들은 ipynb 파일에 답안을 작성하시어 제출해주십시오.
- ☑ **1/18 (목) 23 시 59 분까지** Github 에 PDF 파일과 ipynb 파일을 모두 제출해주십시오. Github 에 제출하는 방법을 모른다면 학술부장 혹은 Slack 의 질의응답을 활용해주시오.
- ☑ 참고 도서 :
  통계학입문(3 판, 강상욱 외), Introduction to Mathematical Statistics(8 판, Hogg et.al.)

문제 1 Central Limit Theorem

중심극한정리는 확률변수의 합 형태 (Sum of Random Variables) 의 극한분포를 손쉽게 구할 수 있도록 해 주기에 통계학에서 가장 자주 사용하는 정리입니다. 이 문제에서는 중심극한정리의 정의와 그 활용에 대해 짚어보겠습니다.

1-1 ✎ : 중심극한정리(Central Limit Theorem)의 정의를 서술하시오.

- ✋ 통계학입문 (3 판) 7 장 참고
- ✋ Hogg(8 판) 4 장 2 절, 5 장 3 절 참고

양의 분산 σ^2 을 가진 임의의 모집단에서 추출한 크기 n의 확률 표본이 $X_1, X_2, \cdots X_n$ 이라면, 확률변수 $\sqrt{n}(\bar{X}-\mu/\sigma)$는 극한분포로 정규분포를 따른다.

1-2 ✎ : 중심극한정리가 통계적 추론 중 "구간추정"에서 어떻게 유용한지 서술하시오.

- ✋ Hogg(8 판) 4 장 2 절

신뢰구간은 모수가 가질 수 있는 값의 범위를 측정하는데, CLT는 신뢰 구간을 구성하는 데 사용되는 표본 평균의 표준 오차를 계산하여 준다.

1-3 © : .ipynb 파일에서 $Unif\ (0,1)$ 의 분포에 대해서 중심극한정리가 적용되는 예시가 있습니다. 코드를 참조하면서 지수분포인 $Exp(2)$ 분포에 대해서 중심극한정리가 적용되는 모습을 보이시오.

✓

문제 2 Linear Algebra

선형대수학은 머신러닝을 위한 수학 중에서 가장 중요한 요소 중 하나이며, 이 중에서 가장 중요한 것 중에서 하나는 바로 SVD (Singular Value Decomposition, 특이값 분해) 입니다. 이것을 알기 위해서 고유값과 고유벡터를 활용한 Diagonalization 에 대해서 먼저 알아본 다음에 SVD 를 사용하며 실제로 이미지를 압축하면서 적용시켜보겠습니다.

2-1 ✎ **:** Diagonalization 의 정의가 다음과 같이 주어졌습니다.

Diagonalization 이란 정방행렬(A) 를 Eigenvalue, Eigenvector 를 통해서 대각행렬 (D) 를 만드는 것이며, 즉 $D = P^{-1} A P$ 를 통해서 대각행렬 (D) 를 찾는 것입니다.

조건들 : 1.) A 는 정방행렬 (Square Matrix) 이다. 2.) A$(n \times n)$ 는 n 개의 독립인 고유벡터를 가지고 있다.

1.) A 에 대한 고유벡터들을 찾으며, 이것을 각각 $P_1, P_2, \cdots, P_n$ 으로 놓는다

2.) $P = [P_1, P_2, \cdots, P_n]$ 메트릭스를 만든다

3.) $P^{-1} A P$ 를 구하면 다음과 같은 형태가 나오게 된다

$$P^{-1} A P = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = D$$

만약에 A 가 대칭 (Symmetric) 행렬이면 다음과 같은 꼴이 나오게 됩니다.

$$A = P D P^T$$

다음과 같은 정방행렬에 Diagonalization 을 적용시켜서 나오게 되는 대각행렬을 쓰시오.

a) $\begin{bmatrix} 6 & -1 \\ 2 & 3 \end{bmatrix}$

$\det(\lambda I - A) = \begin{vmatrix} \lambda-6 & 1 \\ -2 & \lambda-3 \end{vmatrix}$. $(\lambda-6)(\lambda-3) - (-2) = \lambda^2 - 9\lambda + 20$

$= (\lambda-5)(\lambda-4)$. $\lambda_1 = 5$, $\lambda_2 = 4$

For λ_1, $\begin{bmatrix} -1 & 1 \\ -2 & 2 \end{bmatrix} \times(-1) = \begin{bmatrix} 1 & -1 \\ 2 & 2 \end{bmatrix} \times(-2) = \begin{bmatrix} 1 & -1 \\ 0 & 4 \end{bmatrix}$. $\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$x - y = 0$. $x = y$. $x = t$, $y = t$ $t\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\lambda_1 = 5$에 대응하는 eigenvector $= \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. 같은 식으로, if $\lambda = 4$, eigenvector $= \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$

Basis $= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$ (w_1, w_2) . $v_1 = w_1 = (1, 1)$

$v_2 = w_2 - \frac{w_2 \cdot v_1}{\|v_1\|^2} v_1 = w_2 - \frac{3}{4}(1, 1) = (\frac{1}{2}, 1) - (\frac{3}{4}, \frac{3}{4}) = (-\frac{1}{4}, \frac{1}{4})$

Orthonormal basis $= e_1 = \frac{v_1}{\|v_1\|} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$

$e_2 = \frac{v_2}{\|v_2\|} = v_2 \div \frac{1}{2\sqrt{2}} = (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$

$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{\sqrt{2}}{2} \\ \frac{1}{\sqrt{2}} & \frac{\sqrt{2}}{2} \end{bmatrix}$. $P^{-1}AP = D = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$

b) $\begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{bmatrix}$

c) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

문제 2 | Linear Algebra

선형대수학은 머신러닝을 위한 수학 중에서 가장 중요한 요소 중 하나이며, 이 중에서 가장 중요한 것 중에서 하나는 바로 SVD (Singular Value Decomposition, 특이값 분해) 입니다. 이것을 알기 위해서 고유값과 고유벡터를 활용한 Diagonalization 에 대해서 먼저 알아본 다음에 SVD 를 사용하며 실제로 이미지를 압축하면서 적용시켜보겠습니다.

2-1 ✎ **:** Diagonalization 의 정의가 다음과 같이 주어졌습니다.

Diagonalization 이란 정방행렬(A) 를 Eigenvalue, Eigenvector 를 통해서 대각행렬 (D) 를 만드는 것이며, 즉 $D = P^{-1} A P$ 를 통해서 대각행렬 (D) 를 찾는 것입니다.

조건들 : 1.) A 는 정방행렬 (Square Matrix) 이다. 2.) A($n \times n$) 는 n 개의 독립인 고유벡터를 가지고 있다.

1.) A 에 대한 고유벡터들을 찾으며, 이것을 각각 $P_1, P_2, \cdots, P_n$ 으로 놓는다

2.) $P = [P_1, P_2, \cdots, P_n]$ 메트릭스를 만든다

3.) $P^{-1} A P$ 를 구하면 다음과 같은 형태가 나오게 된다

$$P^{-1} A P = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = D$$

만약에 A 가 대칭 (Symmetric) 행렬이면 다음과 같은 꼴이 나오게 됩니다.

$$A = P D P^T$$

다음과 같은 정방행렬에 Diagonalization 을 적용시켜서 나오게 되는 대각행렬을 쓰시오.

a) $\begin{bmatrix} 6 & -1 \\ 2 & 3 \end{bmatrix}$

b) $\begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ → (a)와 같은 방식. A를 triangular 하게 만들면

$A = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. $(\lambda-2)^2(\lambda-1)$. $\lambda_1 = 2$, $\lambda_2 = 1$

eigenvectors : $(0,1,0), (-1,0,1), (0,-1,1)$

$P = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}$. $P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

c) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ → $(\lambda-3)(\lambda-1)(\lambda-2)$. $\lambda_1 = 3$, $\lambda_2 = 1$, $\lambda_3 = 2$

eigenvectors : $(0,1,0), (0,0,1), (1,0,0)$

$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. $P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

2-2 ✎ **:** SVD 의 정의가 다음과 같이 주어졌습니다.

SVD 란 Diagonalization 과는 달리 모든 행렬 (A) 에 대해서 사용이 가능합니다.

1.) $A^T A$, $A A^T$ 행렬들을 만듭니다. 이것은 항상 대칭 (Symmetric) 행렬이 됩니다.

2.) $A^T A = V D V^T$, $A A^T = U D' U^T$ 으로 대각화를 진행을 하고 나서 정규직교화까지 하게 된다면 U 와 V 를 얻게 됩니다.

3.) 여기에서 0 이 아닌 고유값들이 내림차순으로 나열된 것이 바로 D 가 되며, 이것은 바로 $\sum$ 행렬의 대칭 원소들이 됩니다.

4.) 결국 $A = U \sum V^T$ 의 관계를 가지기 때문에 위에서 구한 U 와 V 를 대입시키면 되며 $\sum$ 도 3.) 에서 구했던 걸로 대입을 하면 됩니다.

참고 자료 :

- https://www.youtube.com/watch?v=rziHzFk5JyU
- https://www.youtube.com/watch?v=HeGdlgB8450 (이것을 참조해서 문제를 풀으시면 됩니다)
- https://angeloyeo.github.io/2019/08/01/SVD.html#google_vignette

다음과 같은 행렬에 SVD 를 적용시켜서 나오게 되는 $\sum$ 행렬을 구하시오.

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ ✓

2-3 © **:** .ipynb 파일에서는 SVD (Singular Value Decomposition, 특이값-분해) 를 실제로 이미지 압축을 위해서 활용하는 예시를 보여주고 있습니다. 해당 코드를 본 뒤에 새로운 사진에 대해서 원본에 비해서 적은 용량을 차지하면서도 원본에 대한 정보를 유지해주는 차원 수가 무엇인지 알아냅시오.

15 정도로 차원을 조절해도 원본 인식 가능.
더 낮아지면 다소 인식이 어려워진다.

대각 (trace)의 합은 고유값의 합과 같다.

2.2) SVD of $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

(1) $A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$(\lambda-1)^3 - (\lambda-1)\cdot(-1)^2 = \lambda^3 - 3\lambda^2 + 2\lambda$

$= \lambda(\lambda^2 - 3\lambda + 2) = \lambda(\lambda-2)(\lambda-1)$.

$\lambda_1 = 2,\ \lambda_2 = 1,\ \lambda_3 = 0$

고유벡터 when λ_1 : $\lambda_1 I - A$ (or $A - \lambda_1 I$) $= \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

$x_1 - x_2 = 0,\ x_3 = 0,\ x_1 = x_2$

$V_1 = [x_1, x_2, x_3] = [1, 1, 0]$

when λ_2 : $A - \lambda_2 I = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$x_1 = 0,\ x_2 = 0,\ x_3 = x_3$

$V_2 = [0, 0, 1]$

when λ_3 : $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

$x_1 = -x_2,\ x_3 = 0,\ x_1 + x_2 = 0$

$V_3 = [-1, 1, 0]$

V_1, V_2, V_3 가 직교하기에 ($V_1 \times V_2,\ V_2 \times V_3,\ V_1 \times V_3 = 0$) Gram-Shmirf 생략가능

$u_1' = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$ $u_2' = [0, 0, 1]$ $u_3' = [-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$

$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix}$

$\sigma_1 = \sqrt{\lambda_1} = \sqrt{2},\ \sigma_2 = \sqrt{\lambda_2} = 1$

$\therefore \Sigma = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}$

문제 3 | Student's Theorem

스튜던트 정리는 통계적 추정에서 필요한 정리 중 하나로, 표본평균과 표본분산이 어떤 분포를 갖는지 알려줍니다. 이 문제에서는 스튜던트 정리의 내용을 어떻게 수식적으로 유도할 수 있는지 짚어보겠습니다.

스튜던트 정리는 다음과 같이 총 4 개의 내용으로 구성되어 있습니다.

① $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

② 표본평균 $\bar{X}$ 와 표본분산 s^2 은 서로 독립이다.

③ $(n-1)S^2/\sigma^2$ 는 $\mathcal{X}^2(n-1)$ 분포를 따른다.

④ $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

3-1 ✎ **:** ③에 있는 내용을 증명하시오.

- Hogg(8 판) 3 장 6 절 참고
- 무작위표본 $X_1, \cdots, X_n$ 이 독립적으로 동일하게 (independently and identically distributed) 평균이 μ 이고 분산이 σ^2 인 정규분포를 따를 때, 자유도가 n 인 카이 제곱 분포를 따르는 새로운 확률변수 V 를 아래와 같이 두어 증명에 활용할 수 있습니다.

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \mathcal{X}^2(n)$$

3-2 ✎ **:** ④에 있는 내용을 증명하시오.

- Hogg(8 판) 3 장 6 절 참고
- t 분포의 정의에 따르면, 표준정규분포를 따르는 확률변수와 카이제곱분포를 따르는 확률변수를 이용하 여 t 분포를 유도할 수 있습니다.

3-1) $V = \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2$

$= \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$

$= \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$. 양쪽에 mgf를 취하면,

$(1-2t)^{-n/2} = E[\exp\{t(n-1)S^2/\sigma^2\}](1-2t)^{-1/2}$

여기서 $(n-1)S^2/\sigma^2$ 의 mgf 에 대한 해를 구하면 (c)를 얻는다.

3-2) $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}}$

문제 4 통계적 방법론

t 검정은 모집단이 정규분포를 따르지만 모표준편차를 모를 때, 모평균에 대한 가설검정 방법입니다. 대개 두 집단의 모평균이 서로 차이가 있는지 파악하고자 할 때 사용하며, 표본평균의 차이와 표준편차의 비율 을 확인하여 통계적 결론을 도출합니다. ANOVA Test 의 경우 집단이 2 개보다 많을 때에 모평균이 서로 차이가 있는지 파악하고자 할 때 사용되며, 이것은 코드로만 살피겠습니다.

4-1 ✎ : 어떤 학우가 DSL 학회원(동문 포함)의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크 다는 주장을 하여, 실제로 그러한지 통계적 검정을 수행하려고 합니다. 며칠간 표본을 수집한 결과 다음의 결과를 얻었다고 합시다.

표본 개수가 충분하고, 모분산을 알고 있으므로 Z-test 사용

표본 수 : 210 명, 각 105명

X → 측정에 응한 DSL 학회원들의 평균 키 : 173.5cm / 표준편차 : 7.05cm

Y → 측정에 응한, DSL 학회원이 아닌 사람들의 평균 키 : 171.4cm / 표준편차 : 7.05cm

a) ✎ 귀무가설과 대립가설을 설정하시오.

$H_0 : \mu_X - \mu_Y = 0$, $H_1 : \mu_X - \mu_Y > 0$

b) ✎ 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.

- 통계학입문(3 판) 7 장 참고
- 어떤 검정통계량이 어떤 분포를 따르는지, 언제 귀무가설을 기각하는지 정해야 합니다.

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/n_1 + S_Y^2/n_2}} = \frac{173.5 - 171.4}{\sqrt{7.05^2/105 + 7.05^2/105}} = \frac{2.1}{0.973} = 2.158$$

$Z_{2.158} = 0.01546$.

따라서, 귀무가설을 채택한다. → DSL 학회 원로의 평균키가 더 크다

4-2 © : 또 다른 학우가 다른 학회인 ESC 의 키들도 포함이 된다고 알려주었으며, 새로운 데이터를 heights.csv 파일에 저장해놓았다고 합니다. 이 학우는 학회마다의 평균 키가 똑같다는 주장을 하고 있으며, 해당 학우가 준 데이터를 통해서 이 주장을 검정하려고 합니다.

a) ✎ 귀무가설과 대립가설을 설정하시오. $H_0 : \mu_{DSL} = \mu_{ESC} = \mu_{Else}$

$H_1 : \mu_{DSL} \neq \mu_{ESC} \neq \mu_{Else}$

b) © 파이썬의 scipy.stats 을 활용해서 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오. 결론은 .ipynb 파일에 쓰셔도 괜찮습니다.

- One-way Anova Test 를 활용해서 사용하는 문제입니다. ✓
- 활용해야 될 함수는 scipy.stats.f_oneway 입니다.

Reference

- 통계학입문(3 판, 강상욱 외)
- Introduction to Mathematical Statistics(8 판, Hogg et.al)
- 23-2 기초과제 1 (9 기 이성균)

Data Science Lab

담당자 : 학술부(신재우)

jaewoo356@gmail.com