예측모델 과제 #1

고려대학교 산업경영공학과 석사과정 2020021326 윤훈상

Notation

- *X: Event X / $P(X)$: Event X의 확률*
- *Y: Event Y / $P(Y)$: Event Y의 확률*

조건부 확률

정의
조건부 확률이란 두 사건을 포함하는 확률이며, 한 사건이 given일때, 다른 사건이 일어날 확률을 말한다.

$P(X|Y) = \frac{P(X, Y)}{P(Y)} (P(Y)>0)$ 또는 $\frac{P(X\cap Y)}{P(Y)} (P(Y)>0)$
=Y 사건이 주어졌을 때, X의 확률을 뜻한다.

특징

- $P(X|Y) \neq P(Y|X)$
  = 교환법칙이 성립하지 않아, (Y가 given일때, X의 확률)과 (X가 given일때, Y의 확률)은 같지 않다.

- 하지만 위의 교환법칙이 성립하지 않는 부분은 Bayes 정리로 이어진다.

  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

  - $P(A|B) = \frac{P(A, B)}{P(B)} \Rightarrow P(A, B) = P(A|B)P(B)$

  - $P(B|A) = \frac{P(A, B)}{P(A)} \Rightarrow P(A, B) = P(B|A)P(A)$

  - $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$

  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)} $

  - 베이즈 정리를 통해 $P(X|Y) \neq P(Y|X)$임을 파악할 수 있다.

문제

코오로나라는 질병이 나타나게 되어, 고려대 학생들에게 코오로나 검사를 하고자 한다.
코오로나는 전체 학생들의 5%만에게만 영향을 주며, 코오로나 검사는 70% 정확도를 갖고 있다고 한다.
코오로나 검사 결과 양성이 나타나면, 해당 학생이 코오로나에 걸려있을 확률은 어떻게 될까?

- P(C) = 코오로나 질병에 걸릴 확률 = 5%
- P(~C) = 코오로나 질병에 안 걸릴 확률 = 95%

코오로나 검사 정확도가 70%라고 해서 P(T)라고 바로 정의하면 안된다. 정확도란 본디 옳은 것을 옳게
그른 것을 그르다고 평가해야 되기 때문이다.

- P(T|C) = 코오로나에 걸림 $\Rightarrow$ 양성이 나타날 확률 = 70%
- P(~ T|~C) = 코오로나에 안 걸림 $\Rightarrow$ 음성이 나타날 확률 = 70%
- P(~T|C) = 코오로나에 걸림 $\Rightarrow$ 음성이 나타날 확률 = 30%
- P(T|~C) = 코오로나에 안 걸림 $\Rightarrow$ 양성이 나타날 확률 = 30%

양성이 나타났을 때, 코오로나에 걸릴 확률을 표현하면

$P(C|T) = \frac{P(T|C) P(C)} {P(T)}$로 나타낼 수 있다.

이는 여기서 분모를 분해해야 한다. P(T)는 양성이 나타난 경우인데, 검사가 양성이 나타나는 경우의 수는 두가지

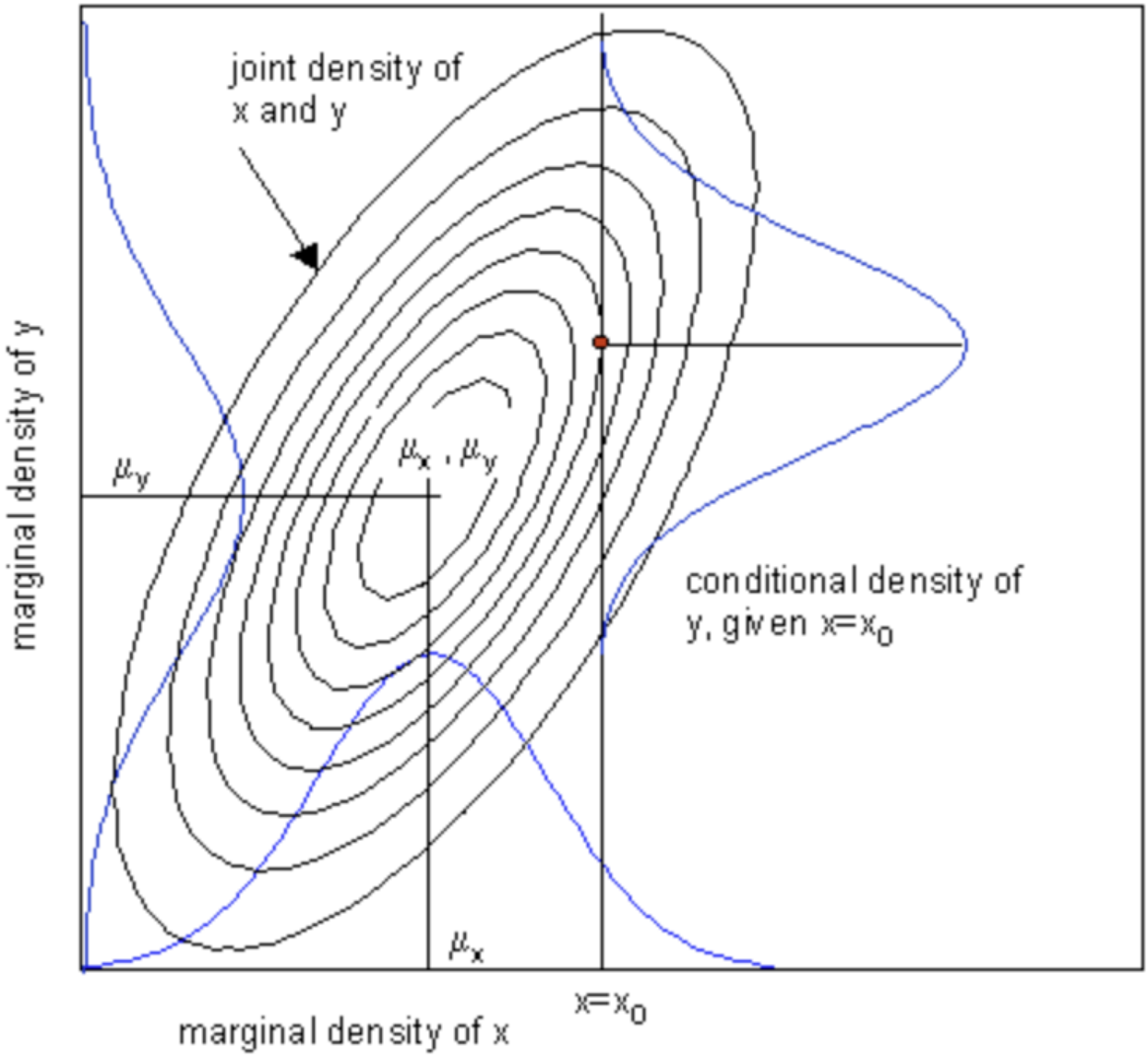
- 코오로나 질병이 있는데 양성인 경우: P(T|C)P(C)
- 코오로나 질병이 없는데 양성인 경우: P(T|~ C)P(~C)

따라서

$P(C|T) = \frac{P(T|C) P(C)} {P(T)} = \frac{P(T|C) P(C)} {P(T|C)P(C)+P(T|C^c)P(C^c))}$ 이므로

$\frac {0.7 X 0.05} {0.7 X 0.05 + 0.3 X 0.95} $ = 7/64 = 0.1 정도의 확률을 갖는다!

결합확률분포 / 조건부확률분포 / 주변확률분포



정의

- **결합확률분포 $P(X, Y)$**

'결합'이라는 말에서 알 수 있듯이, 두 사건이 함께 일어나는 확률의 분포이다. 결합분포에 대한 그래프는 보통 2차원보다 3차원으로 그려내야 하는데, 분포를 이루는 구성요소가 2개 이상이기 때문이다.

- **조건부확률분포** $P(X|Y)$:

특정한 조건이 주어졌을 때의 확률의 분포를 칭한다. 보통 $P(X|Y)$와 같이 표현하며, 이는 Y가 given일 때, X의 확률을 뜻한다.

조건부 확률분포는 두 사건이 하나의 확률의 요소로써 표현되기에 결합확률분포와 헷갈리는 경우가 많은데 조건부 확률분포 $P(X|Y)$는 $Y=y$가 정해진 후에 구할 수 있는 것이지만, 결합확률분포 $P(X, Y)$는 X,Y 사건이 동시에 발생해야 한다.

위의 그림을 보면 $x=x_0$인 선 위에 y의 분포가 그려져 있는 것을 볼 수 있다. 이는 x가 x_0으로 결정된 후의 y의 분포를 나타낸 것이다.

결합확률과 조건부확률은 자주 헷갈리는 개념이기도 하다. 흔히 이에 대한 설명으로 결합확률은 동시에, 조건부 확률은 Sequential, 즉 연속적으로 나타나야 한다 라는 설명을 봤지만, 이는 엄밀히 말하면 틀렸다고 할 수 있다.

예를 들어 확률 계산에서 다루는 사건이 하루에 걸쳐서 나타나는 사건이라고 해보자. 즉, 월요일과 화요일에 비가 올 확률을 생각해보자.

- P(월요일 비, 화요일 비): 월요일과 화요일, 모두 비가 올 확률
- P(화요일 비|월요일 비): 월요일에 비가 왔을 때, 화요일에 비가 올 확률

해당 사건은 연속적인 사건이다. 이 역시 결합확률로 표현하는데에 전혀 문제가 없다.

- 주변확률분포:

주변확률분포란 결합확률분포의 '주변'이라고 생각하면 된다. 즉, 전체의 뭉텅이에서 각 개별 분포만 보겠다는 뜻이다.

따라서 결합분포 $P(X, Y)$에서

- X에 대해 Marginalize하면 $P(Y)$
- Y에 대해 Marginalize하면 $P(X)$ 를 구할 수 있다.

특징
$P(A|B)\Rightarrow$ 조건부확률분포 = $\frac{P(A, B) = 결합확률분포}{P(B)=주변확률분포}$
$P(A|B) \neq P(B|A)$

문제

현 게임 시장에서 가장 유명하고 인기 있는 게임을 꼽자면, League Of Legends(LOL)이다. 모든 과제를 끝내고 LOL을 세 판을 하려고 하는 상훈이는 게임 도중 욕설 횟수, 자신감에 따라 승률이 다르다. 이에 대한 승률표가 주어졌을 때 결합확률분포, 조건부확률분포, 주변확률분포를 구하라

	욕설 1회	**욕설 2회**	**욕설 3회**
자신감 낮음	3/30	1/30	1/30
자신감 중간	5/30	2/30	2/30
자신감 높음	8/30	5/30	3/30

1) **욕설과 자신감에 대한 결합확률분포를 구하라**
위의 표 자체가 결합확률분포이다.

2) **욕설과 자신감에 대한 각기의 주변확률분포를 구하라**

아래의 표는 이어지는 문제들에 도움이 될 것이기 때문에 각 Cell들의 행과 열별의 합을 추가한 것이다.

	욕설 1회	**욕설 2회**	**욕설 3회**	**$P_{자신감}(자신감)$**
자신감 낮음	3/30	1/30	1/30	5/30
자신감 중간	5/30	2/30	2/30	9/30
자신감 높음	8/30	5/30	3/30	16/30
$P_{욕설}(욕설)$	16/30	8/30	6/30	1

- 욕설에 대한 주변확률 분포는 = $P_{욕설}(욕설)$ 행
- 자신감에 대한 주변확률분포 = $P_{자신감}(자신감)$ 열로 나타낼 수 있다.

3) **욕설을 2회 했을 때, 자신감의 조건부 확률 분포를 구하라**
즉, $P(자신감|욕설=2회)$의 분포를 구하면 된다.
이는 1번에서 만들어낸 표를 통해 구할 수 있다. 조건이 '욕설=2회'이므로 해당 열만으로 분포를 구성하면 된다.

즉, $P(A|B) = \frac{P(A, B)}{P(B)}$처럼, 결합 분포내에서 조건을 통해 정규화를 해줘야 한다.
따라서, 아래의 표와 같이 정규화를 거친 열이 조건부확률분포가 된다.

	욕설 2회	욕설 2회(정규화)
자신감 낮음	1/30	1/8
자신감 중간	2/30	2/8
자신감 높음	5/30	5/8
	8/30	1

독립 확률 변수

결합 확률밀도함수가 주변 확률밀도함수들의 곱으로 나타낼 수 있다면 두 확률변수 X와 Y는 독립이다.
$P(X, Y) = P(X) \cdot P(Y)$

함수 식에서 볼 수 있듯, 독립인 경우 각각의 주변 확률밀도함수 만으로도 결합 확률밀도함수를 표현할 수 있는데, 이는 X와 Y가 서로 영향을 받지 않음을 나타낸다.

문제
A 교수님의 연구실의 학생들은 매일 점심을 먹은 뒤에 가위바위보를 통해 커피를 계산할 사람을 고른다. 5일 동안의 가위바위보 참여자 명수가 다음과 같을 때, 1번 이하로만 질 확률을 구하라

월	화	수	목	금
7	6	5	4	3

- 5일 중 한번만 질 확률:
    - **(월)** $\frac {1}{7}$ X **(화)** $\frac {5}{6}$ X **(수)** $\frac {4}{5}$ X **(목)** $\frac {3}{4}$ X **(금)** $\frac {2}{3}$ --월요일에 걸림 +
    - **(월)** $\frac {6}{7}$ X **(화)** $\frac {1}{6}$ X **(수)** $\frac {4}{5}$ X **(목)** $\frac {3}{4}$ X **(금)** $\frac {2}{3}$ --화요일에 걸림 +
    - **(월)** $\frac {6}{7}$ X **(화)** $\frac {5}{6}$ X **(수)** $\frac {1}{5}$ X **(목)** $\frac {3}{4}$ X **(금)** $\frac {2}{3}$ --수요일에 걸림 +
    - **(월)** $\frac {6}{7}$ X **(화)** $\frac {5}{6}$ X **(수)** $\frac {4}{5}$ X **(목)** $\frac {1}{4}$ X **(금)** $\frac {2}{3}$ --목요일에 걸림 +
    - **(월)** $\frac {6}{7}$ X **(화)** $\frac {5}{6}$ X **(수)** $\frac {4}{5}$ X **(목)** $\frac {3}{4}$ X **(금)** $\frac {1}{3}$ --금요일에 걸림 +
      = $\frac {87}{210}$

- 5일 중 한번도 안 질 확률:
    - **(월)** $\frac {6}{7}$ X **(화)** $\frac {5}{6}$ X **(수)** $\frac {4}{5}$ X **(목)** $\frac {3}{4}$ X **(금)** $\frac {2}{3}$ --한 번도 걸리지 않음
      = $\frac {2}{7}$

두 결과를 더하면 $\frac {7}{10}$의 결과가 나타나게 된다.

상관계수 / 자기 상관계수

상관계수란 두 변수가 같이 변하는 정도를 정규화를 거친 수치이다. 두 변수가 연관이 있을 경우 상관이 있다고 하고, 방향과 강도를 나타낸다.

$r = \frac {X와 Y가 함께 변하는 정도} {X와 Y가 각각 변하는 정도} = \frac {cov(X, Y)} {\sigma_X \sigma_Y}$

$r = \frac{\sum(x_i-\overline x)(y_i-\overline y)} {\sqrt{\sum(x_i-\overline x)^2}\sqrt{\sum(y_i-\overline y)^2}}$

특징
1) 상관계수는 단위에 영향을 받지 않는다.
공분산은 단위에 의해 영향을 많이 받게 된다. 예를 들어, 똑같은 돈이라도 한 쪽은 만원 단위로, 그리고 한 쪽은 백원의 단위로 계산하여, 공분산을 계산하게 되면, 상관계수는 비슷해도, 공분산의 양의 차이가 매우 클 것이다.

2) 상관계수는 -1~1의 값을 갖는다.

3) 두 변수를 바꿔도 상관계수의 값은 똑같다.

문제
단순히 자리에 앉아 있는 시간으로 학생의 학업량을 판단할 수는 없지만, 대개 비례한다. 다음과 같이 자리에 앉아 있는 시간과 학업량에 대한 표를 보고 상관계수를 구하라 (학업량은 1~100의 수치로 표현)

학생	A	B	C	D	E
착석시간	2	4	6	10	8
공부량	3	5	7	8	8

$\overline x$ = (2 + 4 + 6 + 10 + 8) / 5 = 6
$\overline y$ = (3 + 4 + 7 + 8 + 8) / 5 = 6

$r = \frac{\sum(x_i-\overline x)(y_i-\overline y)} {\sqrt{\sum(x_i-\overline x)^2}\sqrt{\sum(y_i-\overline y)^2}}$

$\sum(x_i-\overline x)(y_i-\overline y)$
=(2-6)(3-6) + (4-6)(5-6) + (6-6)(7-6) + (10-6)(8-6) + (8-6)(8-6) = 12 + 2 + 0 + 8 + 4 = 26

$\sqrt{\sum(x_i-\overline x)^2}$ = $\sqrt{(2-6)^2 + (4-6)^2 +(6-6)^2 +(10-6)^2 + (8-6)^2}$ = $2\sqrt{10}$

$\sqrt{\sum(y_i-\overline y)^2}$ = $\sqrt{(3-6)^2 + (5-6)^2 +(7-6)^2 +(8-6)^2 + (8-6)^2}$ = $\sqrt{19}$

따라서, 계산과정은 복잡하고 숫자도 깔끔하게 떨어지지는 않지만
r = $\frac {26}{2\sqrt{10}\sqrt{19}}$ = $\frac {13}{\sqrt{10}\sqrt{19}}$이다.

자기상관계수

상관계수는 '두' 변수 사이의 선형 관계를 크기를 측정하지만, 자기상관은 자기 자신, 즉 단일한 변수의 시간 순의 선형관계를 측정한다.

$R_k = \frac {Autocovariance} {Variance}$ = $\frac{\sum_{t=k+1}^T(y_t-\overline y)(y_{t-k}-\overline y)} {\sum_{t=1}^T(y_t-\overline y)^2}$

여기서 k는 시간 사이의 간격이다. 따라서, k가 2면, 1과 3/ 2와 4 같이 2칸씩 건너뛰어서 자기상관계수를 계산하게된다.

문제
담배를 피우는 것을 좋아하는 상현이는 담배를 핀 후 다시 자리에 앉았을 때, 점점 감소하는 체내 니코틴 농도의 변화 정도를 알고 싶다. 다음과 같은 표가 주어졌을 때, 니코틴의 자기상관계수를 구하라.

시간	1	2	3	4	5
니코틴 농도	10	9	6	4	1

$\overline x$ = 6 / k = 1

$\sum_{t=k+1}^T(y_t-\overline y)(y_{t-k}-\overline y)$ =

$(9-6)(10-6) + (6-6)(9-6) + (4-6)(6-6) + (1-6)(4-6)$ = 22

$\sum_{t=1}^T(y_t-\overline y)^2 = (10-6)^2 + (9-6)^2 +(6-6)^2 +(4-6)^2 + (1-6)^2$=54

r = 22/54 = 11/27

EigenVector / EigenValue

EigenVector와 EigenValue를 식으로 표현하면 다음과 같다.

- $Av = \lambda v$
  A = Matrix (Linear Transformation)
  v = EigenVector
  λ = EigenValue

정의: 특정 선형변환 A로 인한 v의 변환 결과가 상수배와 같을 때, v를 고유벡터, 상수배 값 λ를 고유값이라고 한다.

행렬 A에 따라 고유벡터, 고유값이 존재할 수도 하지 않을 수도 있다.

문제
$$A = \begin{bmatrix}
2 & 3\\
1 & 4
\end{bmatrix}$$ 의 EigenValue λ / EigenVector v를 구하라

1. $Ax = \lambda x$로 나타냈을 때, λx를 좌변으로 이항해준다.

2. $(A-\lambda I)x = 0$

3. $(A-\lambda I)$가 역행렬을 지니면 양변에 역행렬을 곱해서 $x=0$만을 남겨버릴 수 있으므로 역행렬을 지니면 안된다. 따라서 Determinant = 0를 만족해야 한다.

$$det \begin{vmatrix}
2-\lambda & 3\\
1 & 4-\lambda
\end{vmatrix}=0$$

4. $(2-\lambda)(4-\lambda)-3=0$
5. $5-6\lambda+\lambda^2 = 0$ 이므로
6. $ \lambda_1= 1$ or $\lambda_2 = 5$을 만족하게 된다.

자 이제 λ를 찾았으니, 그에 상응하는 v만을 찾으면 된다.

- λ_1 = 1일 경우)

$AX = \lambda X \Rightarrow AX = X$

$\begin{bmatrix}
{2} & {3}\\
{1} & {4}
\end{bmatrix}
\begin{bmatrix}
{x_1} \\
{x_2}
\end{bmatrix} =
\begin{bmatrix}
{x_1} \\
{x_2}
\end{bmatrix}
\Rightarrow x_1=-3x_2$

- λ_1 = 5일 경우)

$\begin{bmatrix}
{2} & {3}\
{1} & {4}
\end{bmatrix}
\begin{bmatrix}
{x_1} \
{x_2}
\end{bmatrix} =
5\begin{bmatrix}
{x_1} \
{x_2}
\end{bmatrix}
\Rightarrow x_1=x_2$

Positive Definite Matrix

Positive Definite Matrix에 대해서 알기 위해선 행렬의 Quadradic Form을 먼저 알아야 한다.

Quadradic Form

기본적으로 행렬은 선형 방정식을 표현할 때 사용 가능하다.

$$\begin{bmatrix}
a_{11} & a_{12}\
a_{21} & a_{22}
\end{bmatrix}
\begin{bmatrix}
x_1 \
x_2
\end{bmatrix} =
\begin{bmatrix}
5 & 10\
\end{bmatrix}$$

위의 행렬 연산은 다음과 같은 방정식으로 나타낼 수 있다.

$a_{11}x_1 + a_{12}x_2 = 5$

$a_{21}x_1 + a_{22}x_2 = 10$

이런 Ax와 같은 Form은 방정식의 1차항까지만 표현할 수 있기에, 2차항을 표현하기 위해서 Quadradic Form을 사용하게 된다.

$$ x^TAx =
\begin{bmatrix}
x_1 & x_2\
\end{bmatrix}
\begin{bmatrix}
a_{11} & a_{12}\
a_{21} & a_{22}
\end{bmatrix}
\begin{bmatrix}
x_1 \
x_2
\end{bmatrix}$$

위의 행렬 연산은 다음과 같은 방정식으로 나타낼 수 있다.

$a_{11}x^2 + a_{12}x_1x_2+ a_{21}x_1x_2 + a_{22}(x_2)^2$

Positive Definite Matrix 정의

행렬에 대하여 Quadradic Form을 적용했을 때,

- 모든 x에 대하여 양수면 Positive Definite Matrix ($x^TAx > 0$)
- 0까지는 허용하면 Positive Semi Definite Matrix이다 ($x^TAx \geq 0$)
- 모든 EigenValue이 양수이다.

Matrix의 Positive (Semi)Definite 여부에 따라 머신러닝의 최적화에 대한 대략적인 그림을 알 수 있다.

머신러닝 또는 딥러닝의 훈련은, Loss Function을 사용하고, Gradient Descent를 통해 Loss가 최소가 되는 방향을 찾는 방식이다. 만일, 이 Loss Function을 직접 계산하여 그려보지 않아도 특성을 알수 있다면 매우 효율적일 것이다.

예를 들어 가장 많이 사용하는 Loss Function인 MSE를 예를 들어보자. 이는 Mean 'Squared' Error 이므로 Quadradic Form이며, 따라서 Positive Definite 여부를 판단할 수 있다.

문제

$A = \begin{bmatrix}
7 & 2 \
2 & 1
\end{bmatrix}$ 은 Positive Definite인가?

$x^TAx =
\begin{bmatrix}
x_1 & x_2\
\end{bmatrix}
\begin{bmatrix}
7 & 2\
2 & 1
\end{bmatrix}
\begin{bmatrix}
x_1 \
x_2
\end{bmatrix} = 7x^2+4x_1x_2+x_2^2$이므로

$7x_1^2+4x_1x_2+x_2^2$ > 0 (except x_1, x_2 = 0)여부를 살피면 Positive Definite 여부를 알 수 있다.

먼저 1차 미분 = 0이 되는 지점을 찾는다.
$\frac{\partial f(x_1, x_2)}{\partial x_1} = 14x_1 + 4x_2= 0$

$\frac{\partial f(x_1, x_2)}{\partial x_2} = 4x_1 + 2x_2= 0$

다음으론 2차 미분의 부호를 확인한다. 이는 부호가 양수면 함수가 극소점을, 음수면 극대점을 가짐을 알려주기 때문에 확인해야 한다.

$\frac{\partial^2 f(x_1, x_2)}{\partial x_1 \partial x_2} = 4 > 0 \Rightarrow$ 극소점을 갖는다!

즉 $7x_1^2+4x_1x_2+x_2^2$는 아래로 볼록한 함수이며, 해당 극소점이 0이 아니면 Positive Definite, 극소점이 0이면 Semi Positive Definite가 된다.

따라서 위의 1차 미분=0이 되는 지점을 살펴보면,

- $\frac{\partial f(x_1, x_2)}{\partial x_1} = 14x_1 + 4x_2= 0$
  ex) $x_1 = 2, x_2 =-7$:
  이를 $7x_1^2+4x_1x_2+x_2^2$에 삽입하면, 21 > 0

- $\frac{\partial f(x_1, x_2)}{\partial x_2} = 4x_1 + 2x_2= 0$
  ex) $x_1 = 1, x_2 =-2$:
  이를 $7x_1^2+4x_1x_2+x_2^2$에 삽입하면, 3 > 0

최종적으로 극소점에서 0이상이므로 해당 함수는 모든 x에 대하여 0보다 크게 되어, 행렬 A는 Positive Definite라고 할 수 있다.