인공지능수학

(확률과 통계4)

학습내용

1. 마르코프와 체비세프 부등식
2. 가중최소제곱 과 칼만필터
3. 마르코프 행렬 과 마르코프 연쇄

▣ 확률

■ 확률변수 X 에 대한 확률

- 확률과 기대값의 관계를 설명
- 확률은 X 가 a보다 크거나 같을 확률이며, a가 커지면 확률은 작아진다
- 마르코프는 음이 아닌 임의의 확률변수 X에 대해 $X \geq a$ 를 만족하는 확률상의 단순한 한계를 찾았다.

마르코프 부등식

마르코프 부등식은 $X \geq 0$ 을 가정하면 음인 표본은 없다. $X(s) \geq a$의 확률은 $\frac{E(X)}{a} = \frac{X\text{의평균}}{a} = \frac{\overline{X}}{a}$ 보다 작거나 같다.

■ 체비세프 부등식

- 평균과 분산만 알려진 경우 확률분포가 평균에서 벗어난 정도
- 부등식은 음이 아닌 함수만이 아니라 임의의 변수 $X(s)$ 에 적용
- 평균 $\overline{X}$ 로 부터 먼 사건의 확률에 대한 추정치를 제공
- 확률분포의 꼬리를 찾는다.
- 꼬리가 무거운 분포는 일반 확률분포보다 더 높은 대편차 $|X(s) - \overline{X}|$ 를 가진다. $|X - \overline{X}| \geq a$ 일 확률은 a 가 커지면 낮아진다.
- $Y = |X(s) - \overline{X}|^2 \geq a^2,$
- $Y(s) \geq a^2$, $\mathrm{E}(Y) = E((X(s) - \overline{X})^2) = \sigma^2$ 에 마르코프 부등식 적용

체비세프 부등식

임의의 확률분포 X(s) 에 대한 체비세프부등식은
$|X(s) - \overline{X}| \geq a$ 일 확률은 $\frac{\sigma^2}{a^2}$ 보다 작거나 같다.

모멘트와 중심 모멘트

■ 모멘트

- ➲ : 임의의 n에 대해 $m_n = E[x^n]$을 모멘트라함
- ➲ 0번째 모멘트 = 1 … $\sum p_i = 1$ 또는 $\int p(x)dx = 1$
- ➲ 1번째 모멘트 = 평균 = $E(x)$ … $\sum ip_i = m$ 또는 $\int xp(x)dx = m$
- ➲ 2번째 모멘트 (0근방) … $\sum i^2 p_i$ 또는 $\int x^2 p(x)dx = \sigma^2 + m^2 = E(x^2)$
- ➲ 2번째 중심모멘트(m근방) …
  - ◆ $\sum (i-m)^2 p_i = \sigma^2$ 또는 $\int (x-m)^2 p(x)dx = \sigma^2$

n **번째 모멘트** ---- $m_n = \sum i^n p_i$ **또는** $\int x^n p(x)dx$

중심모멘트(m근방) ---- $\mu_n = \sum (i-m)^n p_i$ **또는** $\int (x-m)^n p(x)dx$

n **번째 정규화된 중심모멘트** ---- $= \mu_n / \sigma^n$

- ➲ 3번째 중심모메트 는 분포의 비틀림(왜도, 비대칭도)
- ➲ 네번째 정규모멘트는 뽀족한 정도(첨도, 평균주위에 몰리 정도)
- ➲ P(x)가 평균을 중심으로 대칭이면 홀수 중심모멘트는 0이 될 것이다.
- ➲ 일반적으로 모멘트는 중심에서 멀리 있을수록 커진다.

생성함수 for 이산분포

- **생성함수와 누적생성함수 : 확률등을 계수로 하는 거듭제곱수의 합**
- **이산확률변수 X, 에 대해 $X = n$ 일 확률이 p_n**
  - 확률생성함수 $G(z) = \sum_0^\infty p_n z^n$: x가 음이아닌정수
  - 특성함수 $\emptyset(t) = \sum_0^\infty p_n e^{itn}$: 각각의 확률함수와 일대일 대응하며 기대값과 분산을 알수있음.
    함수이항분포의 특성함수 $(1\text{-}p+pe^{it})^n$
  - 모멘트생성함수 $M(t) = \sum_0^\infty m_n \frac{t^n}{n!}$:
  - 누적생성함수 $K(t) = \sum_0^\infty k_n \frac{t^n}{n!}$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots = \sum_{k=n}^{\infty} \frac{x^k}{k!}$$

$$e^{ix} = \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \ldots = 1 + ix + \frac{-x^2}{2!} + \frac{-ix^3}{3!} + \frac{x^4}{4!} +$$

$$e^{ix} = \cos x + i \sin x$$

- **생성함수와 기대값**
  - $G(z) = E[Z^X],\ \emptyset(t) = E\left[e^{itX}\right],\ M(t) = E[e^{tX}],\ K(t) = \log E[e^{tX}]$
- **베르누이 분포 (p0=1-p, p1 = p)**
  - $P = E[x] = E[x^2] = E[x^3] = \ldots$
  - $M(t) = (1-p) + pe^t$
  - $K(t) = \log(1 - p + pe^t) \rightarrow\rightarrow \frac{dK}{dt} = \left[\frac{pe^t}{1-p+pe^t}\right]_{t=0} = p$

생성함수 for 연속분포

■ 연속분포에서의 생성함수는 덧셈대신 적분을 형태를 사용

➲ $\emptyset(\mathrm{t}) = \int_{-\infty}^{\infty} p(x)e^{itx}dx, \quad M(t) = \int_{-\infty}^{\infty} p(x)e^{tx}dx, \;\; K(t) = \log M(t)$

■ 정규분포에 대한 생성함수

➲ 모멘트생성함수 : $M(t) = e^{\mu t}e^{\sigma^2 t^2/2}$

➲ 누적생성함수 : $K(t) = \mu t + \sigma^2 t^2/2$

➲ 정규분포일때 $k_1 = \frac{dK}{dt} = \mu, \; k_2 = \sigma^2, \; k_3 = \; k_4 = k_5 = \cdots = 0$

중심극한 정리

■ 평균이 m 이고 분산이 σ^2인 n개의 독립표본 $X_1, \dots X_n$에 대해

➲ $S_n=(x_1+x_2+++x_n)/n$ 의 분포는 기대값 μ, 표준편차 $\sigma/\sqrt{N}$

➲ 표준화된 평균 $Z_n = \sum(X_k - m)/\sigma\sqrt{N}$ 을 활용

■ 중심극한정리

$N \to \infty$ 일때 Z_n 의 분포는 평균이 0 이고 분산이 1인 표준정규분포에 가까워 진다.

■ 증명

➲ $Y = \frac{X-m}{\sigma}$ 특성함수를 사용

➲ $E\left(e^{itY}\right) = E[1 + itY - \frac{1}{2}t^2Y^2 + O(t^3)] = 1 + 0 - \frac{1}{2}t^2 + O(t^3)$

➲ $Z_n = (Y_1 + Y_2 + \cdots + Y_n)/\sqrt{N}$

➲ $N \to \infty,\ (E\left[e^{\frac{itY}{\sqrt{N}}}\right])^N = [1 - \frac{1}{2}(\frac{t}{\sqrt{N}})^2 + O(\frac{t}{\sqrt{N}})^3]^N \to e^{-t^2/2}$

➲ 극한 $e^{-t^2/2}$ 는 표준정규분포 $N(1,\sigma)$ 의 특성함수 이다.

합에 대한 체르노프 부등식

- 합 $X = X_1 + X_2 + \cdots + X_n$ 의 평균은 $\overline{X} = \overline{X_1} + \overline{X_2} + \cdots + \overline{X_n}$
- 확률변수가 독립이면 $\sigma^2 = {\sigma_1}^2 + {\sigma_2}^2 + \cdots + {\sigma_n}^2$
- 체비세프 $Prob\left(\left|X - \overline{X}\right| \geq a\right) \leq \frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_n^2}{a^2}$
- 체르노프 : X_i 의 결합독립일때, 각 쌍의 독립이상
  - ➲ 확률곱 $p(x_1, \ldots x_n) = p_1(x_1) \ldots p_n(x_n)$
  - ➲ 평균 $\overline{X}$ 에서 멀리 떨어진 X 의 확률은 급격히 0에 가까워진다.
- 예) n개의 동전중 앞면이 나오는 개수
  - ➲ 앞면이 나올 확률이 $\overline{X_i} = p_i$ ->$X = X_1 + \cdots + X_n$의 평균은 $\overline{X} = p_1 + \cdots + p_n$
- 체르노프 상하한
  - ➲ 상한: $Prob\left(X \geq (1+\delta)\overline{X}\right) \leq e^{-\overline{X}\delta^2/(2+\delta)}$
  - ➲ 하한: $Prob\left(X \leq (1-\delta)\overline{X}\right) \leq e^{-\overline{X}\delta^2/2}$
  - ➲ $\overline{X} = \frac{n}{2}$, $\delta^2 = (4\log n)/n$

■ 평균이 0이 되도록 맞추자

- ➲ $\overline{X} = 0$ 이 되게 X를 중심에 맞춘다
- ➲ 일반적인 체비세프 : $Prob(|X| \geq a) = Prob(X^2 \geq a^2) \leq E[X^2]/a^2$
- ➲ 체르노프 상한 : $Prob(|X| \geq a) = Prob(e^{sX} \geq e^{sa}) \leq E[e^{sX}]/e^{sa}$
- ➲ 체르노프 하한 : $Prob(|X| \leq a) = Prob(e^{-sX} \geq e^{-sa}) \leq E[e^{-sX}]/e^{-sa}$

- ➲ 모멘트생성함수 : $M(s) = E[e^{sX}]$, $M(-s) = E[e^{-sX}]$
  - ◆ S를 조절하여 체비세프를 이끌수 있다.

행렬에 대한 마르코프 부등식

■ 기본

- ➲ 양의 준정부호 : $x^T Ax \geq 0$
- ➲ $A - X$ 가 양의 준정부호 라면 $X \leq A$, 모든 고유값은 0보다 크거나 같다.
- ➲ $X \not\leq A$ 라면 $A - X$ 는 음의 고유값을 갖는다.

■ 마르코프 부등식

- ➲ $X \geq 0$ 가 평균이 $E[X] = \overline{X}$ 인 준정부호 혹은 정부호 랜덤행렬 일때
- ➲ A가 양의 정부호 행렬이면 (마르코프부등식)

$$prob\{X \not\leq A\} = prob\{A - X \text{ 는 양의 준정부호가 아니다}\} \leq \overline{X}A^{-1}\text{의 대각합}$$

- ➲ 증명
  - $A^{\frac{1}{2}}$ 이 A의 양의 정부호 제곱근이면
  - $X \not\leq A$ 이면 $A^{-1/2}XA^{1/2}$ 의 대각합 > 1 이다
  - $A - X$ 가 양의 준정부호가 아니면 음의 에너지 $v^T(A - X)v < 0$ 를 만족하는 v 가 존재
  - $w^T w < w^T A^{-1/2} X A^{1/2} w$ 를 만족하도록 $w = A^{1/2}v$ 라 하면
  - $A^{-1/2}XA^{1/2}$ 의 최대 고유값은 $\lambda_{max} > 1$
  - $\lambda_{max} = max \frac{y^T A^{-1/2} X A^{1/2} y}{y^T y} > 1$
  - $A^{-1/2}XA^{1/2}$ 은 음의 고유값이 없으므로 대각합은 1보다 커진다.
  - $prob\{X \leq A\} \leq E[trace(A^{-\frac{1}{2}}XA^{\frac{1}{2}})] = trace(A^{-\frac{1}{2}}XA^{\frac{1}{2}}) = trace(\overline{X}A^{-1})$

행렬에 대한 체비세프 부등식

■ 체비세프 부등식

- 평균이 0인 랜덤행렬 X 의 체비세프 부등식
- A 가 양의 정부호 행렬이면
- $Prob\{|X| \not\leq A\} < trace(E[X^2]A^{-2})$ 이다
- $A - |X|$가 양의 준정부호 행렬이면 $A^2 - X^2$은 양의 정부호 행렬이 아니다.
- $Prob\{|X| \not\leq A\} \leq Prob\{X^2 \not\leq A^2\} < trace(E[X^2]A^{-2})$ 이다

행렬에 대한 체르노프부등식

- $n \times n$ 크기의 양의 준정부호 행렬 또는 정부호행렬인 확률변수 X_k의 합 Y에 대한 체르노프 부등식
  - ➲ Y 가 평균에서 멀어지려면 평균에서 멀리 있는 X_k가 많이 필요함
  - ➲ 일반적이지 않은 확률은 급격히 0에 가까워진다.
  - ➲ 평균에서 멀리 떨어진 꼬리확률에 대해 매우 작은 한계를 얻음
- 체르노프 부등식
  - ➲ $Y = \sum X_k$ 에 속하는 각 행렬 X_k 가 $0 \le \lambda \le c$ 인 고유값을 갖는다고 가정
  - ➲ μ_{min} 과 μ_{max} 를 평균합 $Y = \sum X_k$의 최소 고유값과 최대 고유값이라하자
  - ➲ 그러면
    - $E[\lambda_{min}(Y)] \ge \left(1 - \frac{1}{e}\right)\mu_{min} - C\log n$
    - $E[\lambda_{max}(Y)] \le (e-1)\mu_{max} + C\log n$
  - ➲ 고유값이 평균으로 부터 멀리 있을 확률은 급격히 0에 가까워짐
    - $Pro\{\lambda_{min}(Y) \le t\mu_{min}\} \le ne^{-(1-t)\mu_{min}/2C}$
    - $Pro\{\lambda_{max}(Y) \ge t\mu_{max}\} \le n(\frac{e}{t})^{t\mu_{max}/C}$ (단, $t \ge e$)

공분산 행렬과 결합확률

■ 선형대수학

➲ 한번에 M개의 서로다른 실험을 할 때 사용하는 계산

■ 공분산

➲ 서로다른 실험의 연관(연결)성을 측정하는 방법

➲ 예)

$$\sigma_{ah} = E[(\text{나이} - \text{평균나이})(\text{키} - \text{평균키})]$$

➲ 위 계산하려면 (나이와 키)쌍에 대한 결합확률 P_{ij}을 알아야 함

◆ 공분산은 $(x - m_1)(y - m_2)$ 의 기대값,

-> 공분산 $\sigma_{12} = \sum\sum_{i,j} P_{ij}(x_i - m_1)\ (y_j - m_2)$

➲ 두개의 실험이 독립이라면 결합확률 $p_{ij} = p_i \times p_j$, 공분산 $\sigma_{ij} = 0$

➲ 각실험이 종속이 아니라면 공분산행렬 V 는 양의 정부호이다.

➲ $z = x + y$ 일때

◆ $E(z) = E(x) + E(y)$

◆ $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$

■ $Z = AX$ 의 공분산 행렬

➲ $z = x + y$ 일 경우를 보면

- $\sigma_z^2 = [1\ 1]\begin{bmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2\end{bmatrix}\begin{bmatrix}1 \\ 1\end{bmatrix}$ -> $\sigma_z^2 = AVA^T$ ($\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$를 참조)

➲ 길이가 K인 벡터일때

- $Z = AX$ 의 공분산 행렬은 $V_Z = AV_XA^T$
- 예) Z = x1+x2+x3 -> $[1,1,1]A[1,1,1]^T$

■ 상관계수

- ➲ 공분산과 상관계수는 독립과 종속을 측정한다.
- ➲ 확률변수 x, y 에 대한 제척도구성(rescaling) 이다
- ➲ 분산이 $\sigma_x^2 = \sigma_y^2 = 1$ 이면 $X = \frac{x}{\sigma_x}, Y = \frac{y}{\sigma_y}$
- ➲ x와 y의 상관관계는 X 와 Y 의 공분산이다.
- ➲ 상관관계
  - ◆ $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{x}{\sigma_x}$ 와 $\frac{y}{\sigma_y}$ 의 공분산 (단, $-1 \le \rho \le 1$)
- ➲ 독립인 확률변수일때는 $\rho_{xy} = 0$
- ➲ $(\rho_{xy})^2 \le \sigma_x^2 \sigma_y^2$, 공분산행렬 V는 최소한 양의 준정부호행렬이다.
- ➲ 상관계수 행렬 R
  - ◆ 대각행렬 $D = diag\left[\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_M}\right]$ 에 대해여 $R = DVD$ 이고 $R_{ii} = 1$ (모든 i)
  - ◆ 만약 공분산 행렬 V가 양의 정부호이면 상관관계 $R = DVD$또한 양의 정부호이다

다변량 정규분포와 가중최소제곱

■ 다변량 정규분포

➲ 확률밀도함수 즉 가우스 함수는 $p(x)$ 는 평균 m 과 분산 σ^2에 의존

- $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2}$

➲ $p(x)$의 그래프는 평균 $x = m$ 이 중심인 종모양의 곡선이고 연속변수 x는 $-\infty$ 에서 $+\infty$ 사이의 값을 가질때, 변수 x가 $m-\sigma$에서 $m+\sigma$ 사이에 있을 확률은 대략 2/3이다.

- $X = (x-m)/\sigma$
- $\int_{-\infty}^{\infty} p(x)dx = 1$이고 $\int_{m-\sigma}^{m+\sigma} p(x)dx = \frac{1}{\sqrt{2}}\int_{-1}^{1} e^{-X^2/2}dX \approx \frac{2}{3}$

➲ 표준정규분포 $N(1,0)$ 은 $p(x) = \frac{1}{2\pi} e^{-x^2/2}$ 을 만족

- $F(a) = \int_{-\infty}^{a} p(x)dx$, $F(0) = \frac{1}{2}$

■ 2차원 정규분포

➲ $M = 2$ 일때 정규확률변수 x, y 가 서로 독립이라면

- 확률밀도함수 $p(x,y) = p(x) \times p(y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-(x-m_1)^2/2\sigma_1^2} e^{-(y-m_2)^2/2\sigma_2^2}$
- 공분산 $\sigma_{12} = 0$, 공분산행렬 V은 대각행렬이다.
- $p(x,y)$ 은 x-지수 와 y-지수 의 합이고, 두 지수는
- $\frac{1}{2}(x-m)^T V^{-1}(x-m)$으로 결합할수 있는 장점

- $-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(y-m_2)^2}{2\sigma_2^2} = -\frac{1}{2}\begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix}$

➲ 독립이 아닌 두 변수 x, y

- M개의 변수 $x = x_1, \dots, x_M$ 사이의 종속성은 $M \times M$ 공분산행렬 V로 설명
- 공분산 역행렬은 $p(x)$ 를 표현하는식의 일부이다.
- 다변량 정규확률분포 $p(x) = \frac{1}{(\sqrt{2\pi})^M \sqrt{\det V}} e^{-(x-m)^T V^{-1}(x-m)/2}$
- $V = Q\Lambda Q^T$
- $X = x - m,\ (x-m)^T V^{-1}(x-m) = X^T Q \Lambda^{-1} Q^T X = Y^T \Lambda^{-1} Y$
- $\mathrm{Y} = Q^T X = Q^T(x-m)$ 은 확률적으로 독립
- X 가 되도록 m을 빼서 변수 $x = (x_1, \dots, x_M)$를 중심에 맞추고나서 변수 $\mathrm{Y} = Q^T X$ 가 되도록 회전하였을때 $p(x)$ 의 적분은 바뀌지 않는다. Λ는 대각행렬이 된다.
- $p(x)$ 의 평균과 분산도 M차원 중적분이다.
- 평균의 벡터 m $\int \dots \int x p(x) dx = (m_1, m_2, \dots) = m$
- 공분산 행렬 V $\int \dots \int (x-m) p(x) (x-m)^T dx = V$

■ 가중최소 제곱

➲ $Ax = b$ 가 비가해 선형방정식이라 하자. 즉 해가 없지만 근사해를 구해보자

➲ 오차 $E = \|b - Ax\|^2$ 를 최소화 하는 근사해 $\hat{x}$ 를 선택

- $N(1,0)$을 따르는 정규분포인 독립확률변수 -> 최소화하는 측정방법이다.
- 오차가 독립적이지 않거나 분산이 같지 않다면 최소화 측정법이 아니다. 이때는 가중최소제곱을 사용해야 한다.
- $E = (b - Ax)^T V^{-1}(b - Ax)$ 가 좋은 오차 측정 방법이다.

➲ 분산이 1 (공분산은 0) 이 되도록 b 의 오차에 가중치를 준다.

- 가중 최소제곱(b의 독립인오차) -> $E = \sum_{i=1}^{m} \frac{(b-Ax)_i^2}{\sigma_i^2}$ 을 최소화
- b_i/σ_i 에 대해 분산이 1이 되도록 방정식을 σ_i 로 나눈다
- $Ax = b$ 가 $V^{-1/2}Ax = V^{-1/2}b, \quad V^{-1/2} = diag(1/\sigma_1, \dots, 1/\sigma_m)$
- 일반적인 최소제곱이 된다.
- $\left(V^{-\frac{1}{2}}A\right)^T (V^{-\frac{1}{2}}A)x = \left(V^{-\frac{1}{2}}A\right)^T V^{-\frac{1}{2}}b$ -> $A^T V^{-1} A\hat{x} = A^T V^{-1} b$

■ 추정된 $\hat{x}$ 의 분산

- ➲ b 의 평균이 0 이면 $\hat{x}$ 의 평균도 0 이다.
- ➲ b 의 분산 V 와 $\hat{x}$ 의 분산 W 의 연결하는 공식
- ➲ $\hat{x}$ 에 대한 분산-공분산 행렬 $W = E[(\hat{x} - x)(\hat{x} - x)^T] = (A^T V^{-1} A)^{-1}$
- ➲ $\hat{x} = Lb$ 는 공분산행렬 LVL^T 이고, $\hat{x} = Lb$는

  가중치 방정식 $A^T V^{-1} A\hat{x} = A^T V^{-1} b$를 풀므로

  $L = (A^T V^{-1} A)^{-1} A^T V^{-1}$ 이고 $LVL^T = (A^T V^{-1} A)^{-1^T}$

마르코프 연쇄

■ 마르코프 연쇄

➲ 시간에 따른 계의 상태 변화를 나타냄. 미래상태의 조건부 확률이 과거가 아닌 현재 상태에 의해서만 결정됨.

➲ 예) 렌터카

- 시카고에 있는 차의 80%는 시카고에 머문다.
- 덴버에 있는 차의 30%는 시카고로 이동한다.
- 시카고에 있는 차의 20%는 덴버로 이동한다.
- 덴버에 있는 차의 70%는 덴버에 머문다.

➲ n 번째 달에서 $n+1$ 번째달까지 차의 이동은 $y_{n+1} = Py_n$

- $y_{n+1} = \begin{bmatrix} \text{시카고차량} \\ \text{덴버차량} \end{bmatrix}_{n+1} = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \begin{bmatrix} \text{시카고차량} \\ \text{덴버차량} \end{bmatrix}_{n} = Py_n$
- n 개월 후의 차량의 분포는 $y_n = P^n y_0$
- $y_0 = \begin{bmatrix} 100 \\ 0 \end{bmatrix}$일때 $y_1 = \begin{bmatrix} 80 \\ 20 \end{bmatrix}$, $y_2 = \begin{bmatrix} 70 \\ 30 \end{bmatrix}$, $y_3 = \begin{bmatrix} 65 \\ 35 \end{bmatrix}$,$\cdots$ $y_\infty = \begin{bmatrix} 60 \\ 40 \end{bmatrix}$
- $y_0 = \begin{bmatrix} 0 \\ 100 \end{bmatrix}$일때 $y_1 = \begin{bmatrix} 30 \\ 70 \end{bmatrix}$, $y_2 = \begin{bmatrix} 45 \\ 55 \end{bmatrix}$, $y_3 = \begin{bmatrix} 52.5 \\ 47.5 \end{bmatrix}$,$\cdots$ $y_\infty = \begin{bmatrix} 60 \\ 40 \end{bmatrix}$
- 모두 같은 극한 분포를 갖는다.

■ 행렬 P 의 고유값과 고유벡터

- ➲ 고유값 $det\begin{bmatrix}0.8-\lambda & 0.3\\ 0.2 & 0.7-\lambda\end{bmatrix} = \lambda^2 - 1.5\lambda + 0.5 = (\lambda-1)(\lambda-0.5) = 0$
- ➲ $\lambda = 1$ 과 $\lambda = 0.5$
- ➲ 고유벡터 $\begin{bmatrix}0.8 & 0.3\\ 0.2 & 0.7\end{bmatrix}\begin{bmatrix}0.6\\ 0.4\end{bmatrix} = \begin{bmatrix}0.6\\ 0.4\end{bmatrix}$, $\begin{bmatrix}0.8 & 0.3\\ 0.2 & 0.7\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1\\ -1\end{bmatrix}$
- ➲ 안전상태 $\begin{bmatrix}0.6\\ 0.4\end{bmatrix}$는 극한분포 이며 $\lambda_1 = 1$, $\lambda_2 = \frac{1}{2}$ 은 매달 안정상태까지의 거리가 $\frac{1}{2}$ 씩 좁혀짐을 의미한다.
- ➲ $P_n = (X\Lambda X^{-1}) \dots (X\Lambda X^{-1}) = X\Lambda^n X^{-1} = \begin{bmatrix}0.6 & 0.6\\ 0.4 & 0.4\end{bmatrix} + \left(\frac{1}{2}\right)^n \begin{bmatrix}0.4 & -0.6\\ -0.4 & 0.6\end{bmatrix}$
- ➲ 이때의 P 가 마르코프 행렬이다.
- ➲ 양의 마르코프 행렬이 되기 위한 조건
  - 모든 성분은 $p_{ij} > 0$ 이고, P의 각 열을 더하면 1이다
  - 그러면 $P^T 1 = 1$ 이다. 행렬 P의 고유값은 $\lambda_1 = 1$ 이고 양수인 고유벡터 $x_1 > 0$

■ 전이확률

- ➲ n 시점의 상태 j 에서 $n+1$ 시점에서의 상태 i 가 될 확률 p_{ij}
- ➲ 전이확률 $p_{ij} = \{x(n) = j$ 일때 $x(n+1) = i$ 가 될 확률$\}$
  - ◆ 확률 p_{ij} 는 n에 의존하지 않는다. 시점에 관계없이 똑같이 적용
  - ◆ 새로운 상태 $x(n+1)$ 의 확률 y_{n+1} 은 현재 상태 $x(n)$ 에만 의존, 과거의 기록은 무시된다.
- ➲ 마르코프 연쇄
  - ◆ 유한마르코프연쇄 : 각 상태 $x(n)$ 은 $1,2,\dots,N$중 하나이다.
  - ◆ 무한마르코프연쇄 : 각 상태 $x(n)$ 은 $1,2,\dots$이다.
  - ◆ 연속마르코프연쇄 : 각 상태 $x(n)$ 은 실수이다.
- ➲ 유한마르코프연쇄
  - ◆ 각 상태의 확률은 $p_{1j}, p_{2j}, \dots, p_{Nj}$로 나타냄
  - ◆ 행렬 P의 j열 $p_{1j} + p_{2j} + \cdots p_{Nj} = 1$
  - ◆ 확률 p_{ij}는 $P = N \times N$으로 전이확률, 전이 행렬이라 함

➲ $y_{n+1} = \begin{bmatrix} Prob\{x(n+1)=1\} \\ \vdots \\ Prob\{x(n+1)=N\} \end{bmatrix} = \begin{bmatrix} P_{11} & \cdots & P_{1N} \\ \vdots & \ddots & \vdots \\ P_{N1} & \cdots & P_{NN} \end{bmatrix} \begin{bmatrix} Prob\{x(n)=1\} \\ \vdots \\ Prob\{x(n)=N\} \end{bmatrix}$

➲ $1^T P = 1^T$,(단, $1^T = N$개 성분이 모두 1인 행백터)

➲ $P^T 1 = 1$,(단, $1 = N$개 성분이 모두 1인 열백터)

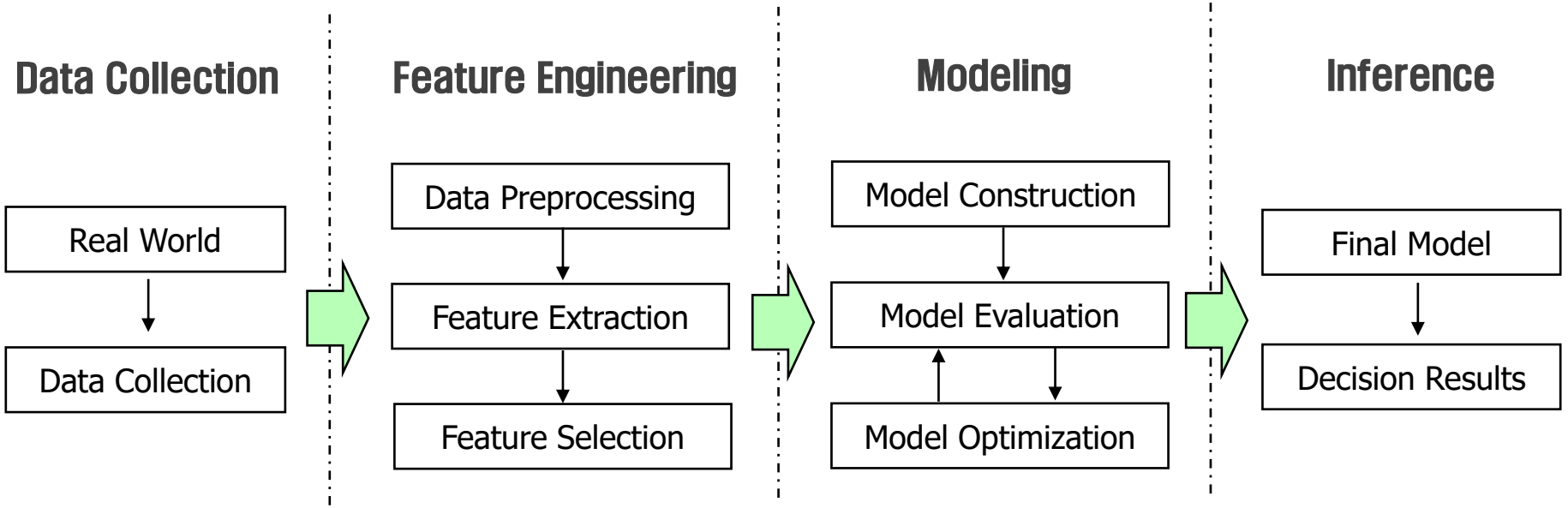
➲ P^T 는 고유값 $\lambda = 1$ 과 고유벡터 $1 = (1,1,\ldots,1)$

머신러닝

■ 머신러닝의 기본적인 흐름

- 머신러닝은 다음과 같은 동작들로 분류할 수 있다.

1) Data collection
2) Feature Engineering
3) Modeling (또는 Learning)
4) Inference

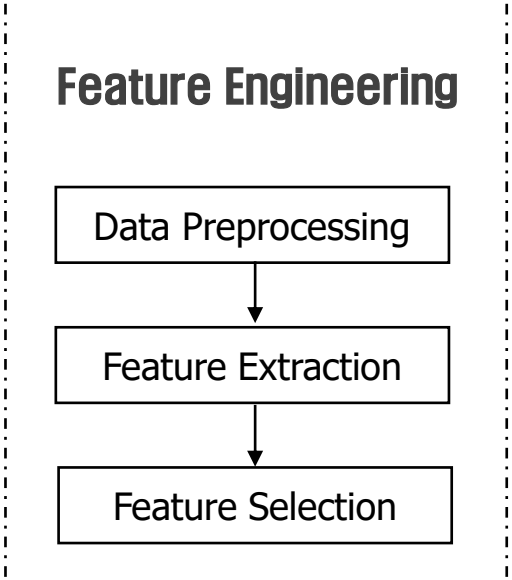


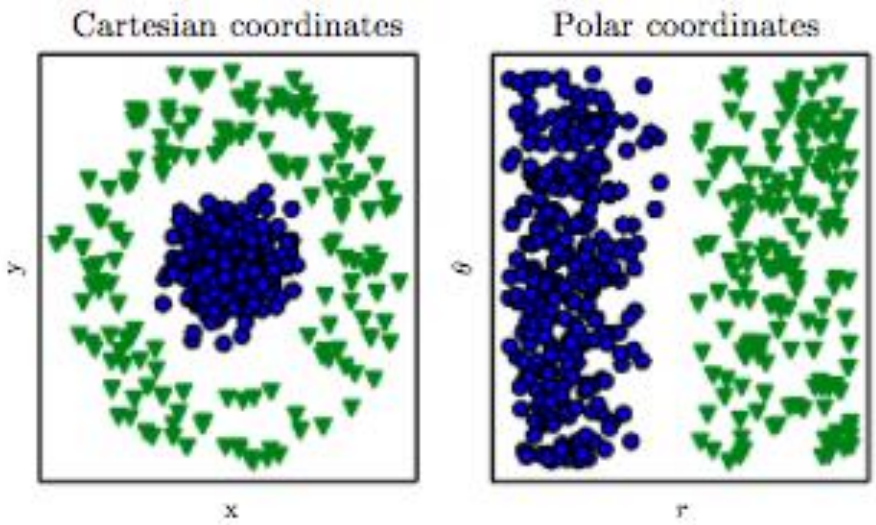
Feature Engineering

□ Feature engineering

- Feature engineering은 다음의 두 가지를 위해 필요한 과정이다.

1. Input 데이터셋에 가장 적합한 머신러닝 모형을 설계하기 위해 필요하다.
2. 머신러닝 학습시 더욱 높은 정확도를 얻기 위해서 필요하다.





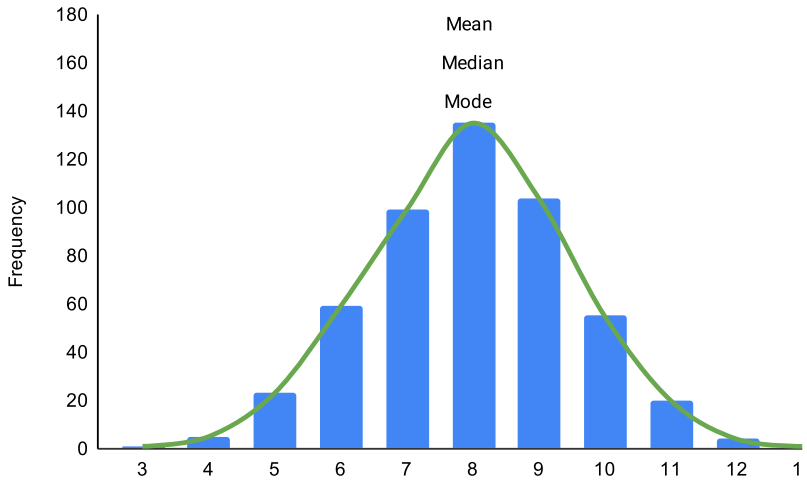
어떤 형태로 정리된 데이터를 사용하는게 더 좋은지는 머신러닝의 모형에 따라 달라질 수 있다.

그렇다면 Feature engineering을 위해 무엇이 필요할까?

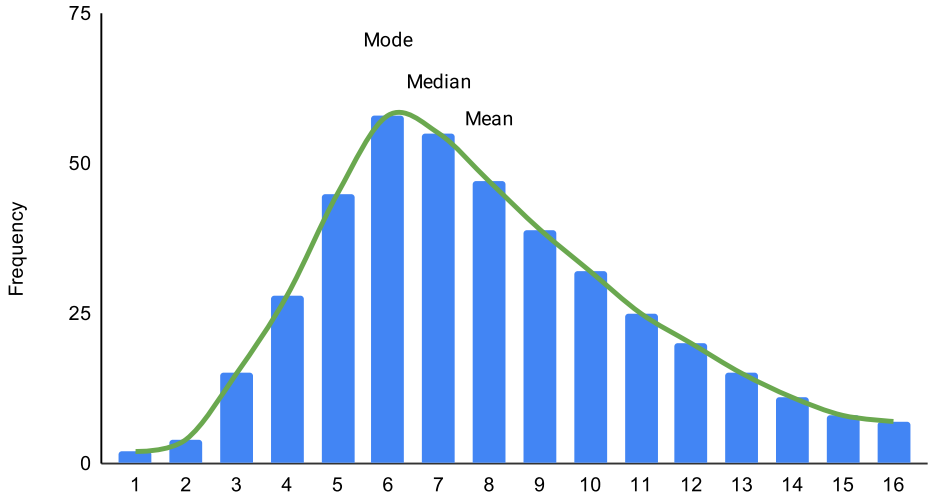
Feature Engineering

■ 이슈 1 : 데이터의 경향성

모든 유의미한 데이터들은 특정 값에 가까운 값들을 갖는 경향이 있다.

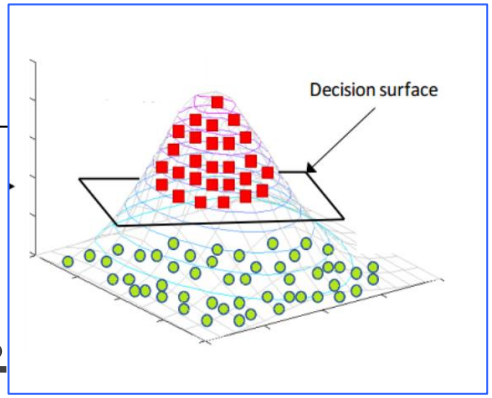


Normal distribution



Skewed distribution

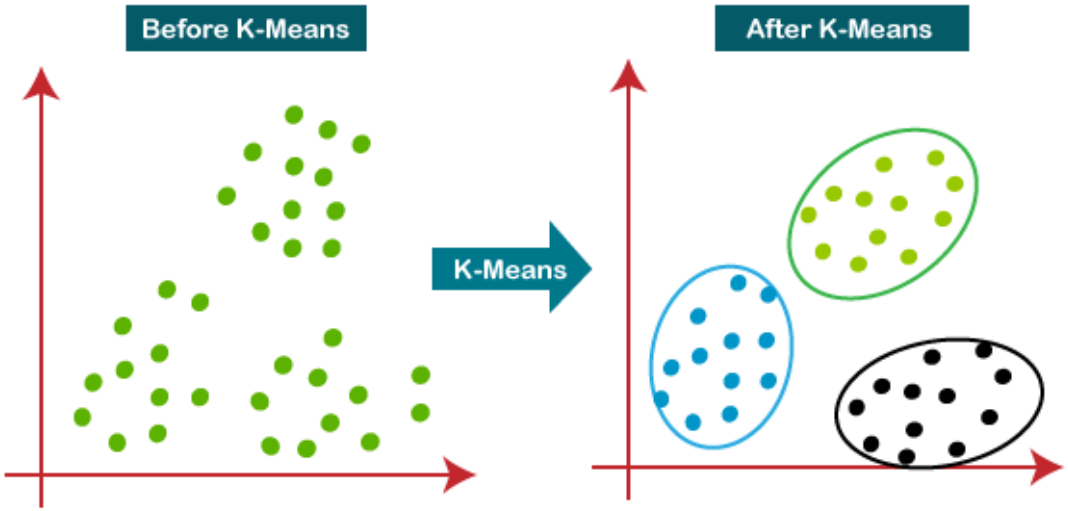
Feature Engineering

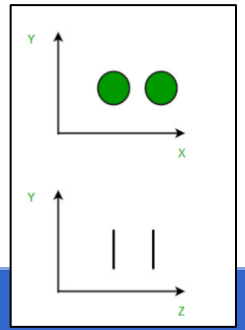


SVM

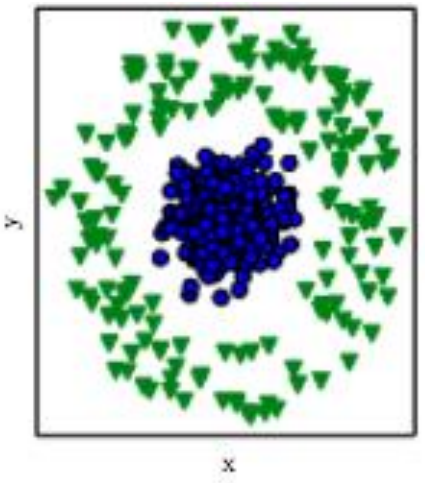
■ 이슈 1 : 데이터의 경향성

- 예를들어 클러스터링을 수행하는 머신러닝의 경우 이러한 데이터의 경향성을 이용하여 수행한다.





PCA



- 하지만 이런 데이터는 잘 분류해내지 못한다.

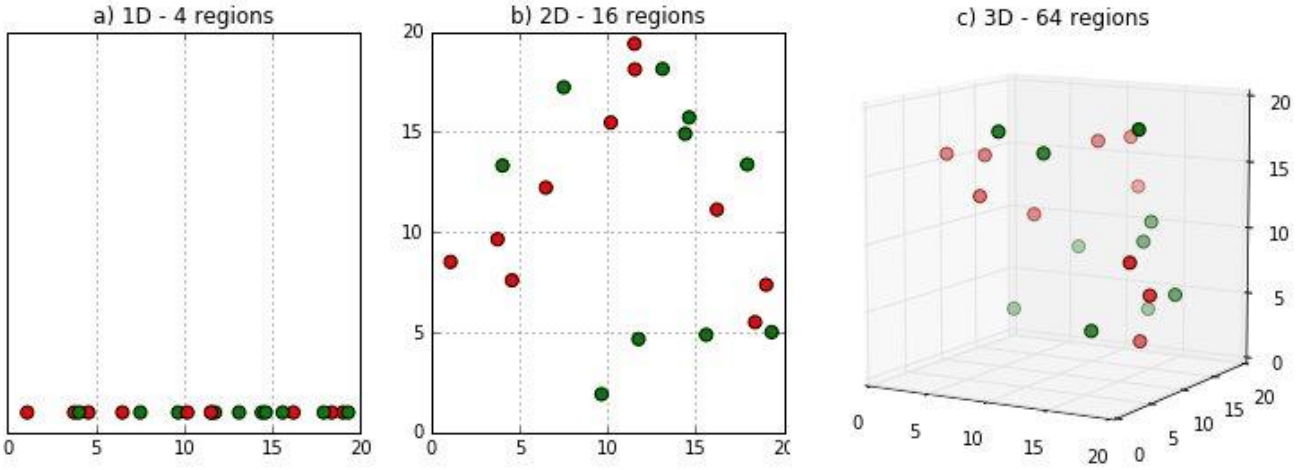
Feature Engineering

■ 이슈 2 : 차원의 저주(The Curse of Dimensionality)

- 고차원의 다변수 데이터를 이용하여 머신러닝을 제작하는 경우,
머신러닝의 성능이 크게 떨어지는 현상들을 통틀어 **차원의 저주**라고 한다.

1) Sparsity problems
데이터의 차원이 커지면, 정확한 클러스터링을 위해 필요한 데이터의 수가 기하급수적으로 커지는 현상

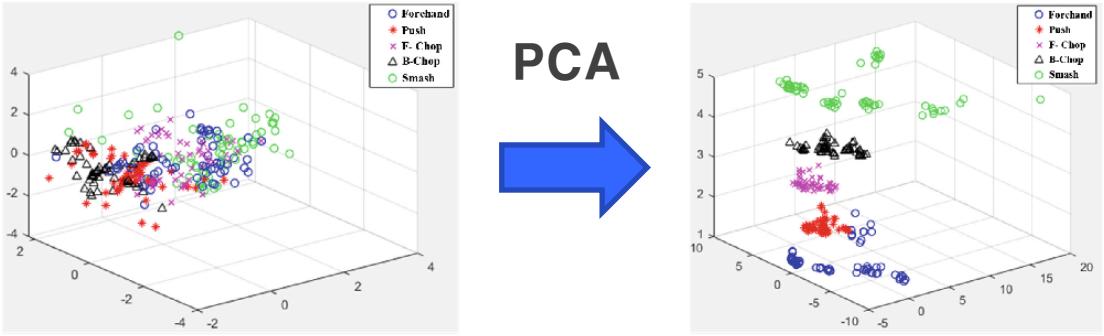
2) Computational problem
데이터의 차원이 커지면, 그만큼 계산에 걸리는 시간도 기하급수적으로 커지는 현상



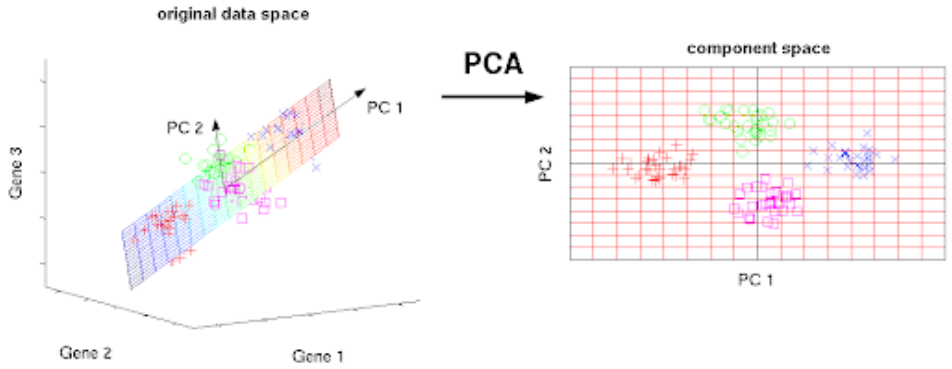
PCA

■ Principal Component Analysis(PCA)

- PCA는 데이터의 경향성을 뚜렷하게 하면서 동시에 데이터의 차원을 낮추는 대표적인 Feature engineering 기법이다.



- PCA는 주어진 고차원의 데이터를, 데이터의 통계적 특성들(분산과 편차)을 유지한 채로 특정 방향의 저차원 축으로 사영시키는 방법
- 전체 데이터의 분포를 잘 표현할 수 있는 차원을 선정하는 알고리즘



PCA

□ Preliminaries for PCA

1) SVD (특이값분해)
2) 공분산행렬

- N : 데이터의 크기
- x_i : 데이터 X의 i번째 값
- 데이터 X의 평균 : μ_x

SVD

$$A = U \Sigma V^T$$

Truncated SVD

$$A_t = U_t \Sigma_t V_t^T$$

평균

$$\mu = \sum_i^N x_i$$

분산

$$V[X] = \frac{\sum_i^N (x_i - \mu)^2}{N}$$

공분산

$$Cov[X,Y] = \frac{\sum_i^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

공분산행렬

$$C_2 = \begin{pmatrix} V[X] & Cov[X,Y] \\ Cov[Y,X] & V[Y] \end{pmatrix}$$

$$C_3 = \begin{pmatrix} V[X] & Cov[X,Y] & Cov[X,Z] \\ Cov[Y,X] & V[Y] & Cov[Y,Z] \\ Cov[Z,X] & Cov[Z,Y] & V[Z] \end{pmatrix}$$

$$\vdots$$

PCA

☐ The process of PCA

1) Input

$m \times n$ 행렬

$$A = \begin{bmatrix} x_1 & \cdots & z_1 \\ \vdots & \ddots & \vdots \\ x_m & \cdots & z_m \end{bmatrix}$$

2) 정규화

$$D = \begin{bmatrix} x_1 - \mu_x & \cdots & z_1 - \mu_z \\ \vdots & \ddots & \vdots \\ x_m - \mu_x & \cdots & z_m - \mu_z \end{bmatrix}$$

3) Truncated SVD

$$D = U \Sigma V^T$$

Select V_t

4) 기저변환
축의 새로운 base선정

$$new\ A' = AV_t^T$$

5) Output

Return A'

PCA

□ PCA의 원리

- 공분산행렬의 특이행렬로 기저변환을 하면, 데이터의 분산을 유지한 채 낮은 차원의 데이터로 변환하는 것이 가능하다.

Input 행렬 $A = \begin{bmatrix} x_1 & \cdots & z_1 \\ \vdots & \ddots & \vdots \\ x_m & \cdots & z_m \end{bmatrix}$, $D = \begin{bmatrix} x_1 - \mu_x & \cdots & z_1 - \mu_z \\ \vdots & \ddots & \vdots \\ x_m - \mu_x & \cdots & z_m - \mu_z \end{bmatrix}$ 에 대하여,

A의 공분산행렬은 다음과 같이 구할 수 있다.

$$C = \frac{1}{m} D^T D = \begin{bmatrix} \frac{\sum_i^m (x_i - \mu_x)^2}{m} & \cdots & \frac{\sum_i^m (x_i - \mu_x)(z_i - \mu_z)}{m} \\ \vdots & \ddots & \vdots \\ \frac{\sum_i^m (x_i - \mu_x)(z_i - \mu_z)}{m} & \cdots & \frac{\sum_i^m (z_i - \mu_z)^2}{m} \end{bmatrix} = \begin{bmatrix} V[X] & \cdots & Cov[X,Z] \\ \vdots & \ddots & \vdots \\ Cov[Z,X] & \cdots & V[Z] \end{bmatrix}$$

PCA

□ PCA의 원리

- 공분산행렬의 특이행렬로 기저변환을 하면, 데이터의 분산을 유지한 채 낮은 차원의 데이터로 변환하는 것이 가능하다.

- 그런데 공분산행렬의 우특이행렬은 D의 우특이행렬과 같다.

2) 정규화

$$D = \begin{bmatrix} x_1 - \mu_x & \cdots & z_1 - \mu_z \\ \vdots & \ddots & \vdots \\ x_m - \mu_x & \cdots & z_m - \mu_z \end{bmatrix}$$

3) Truncated SVD

$$D = U\,\Sigma\,V^T$$

$$\text{Select } V_t$$

여기서 필요한 것은 V뿐이다
따라서 D에 대한 SVD를
구하면 된다.

$$mC = D^T D = (V\sum U^T)(U\sum V^T) = V\,\Sigma^2\,V^T$$

$$C = V\,\Sigma'\,V^T$$

여기서 V는 동일하다

PCA

□ PCA의 원리

- 공분산행렬의 특이행렬로 기저변환을 하면, 데이터의 분산을 유지한 채 낮은 차원의 데이터로 변환하는 것이 가능하다.

4) 기저변환

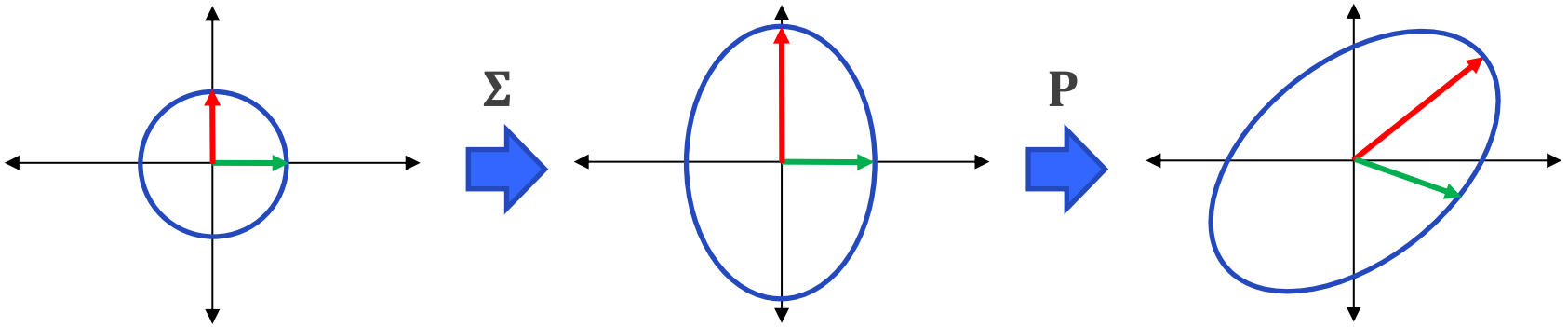
$$new\ A' = AV_t^T$$

■ 기저변환

$$\mathbf{A} = \mathbf{P}\,\mathbf{\Sigma}\,\mathbf{P}^{-1}$$

D의 우특이행렬을 곱하는 것으로 다음 두 효과를 얻을 수 있다.

1) 데이터의 분산이 보존된다.
2) 서로 다른 경향의 데이터들 사이의 거리가 최대로 멀어진다.



PCA

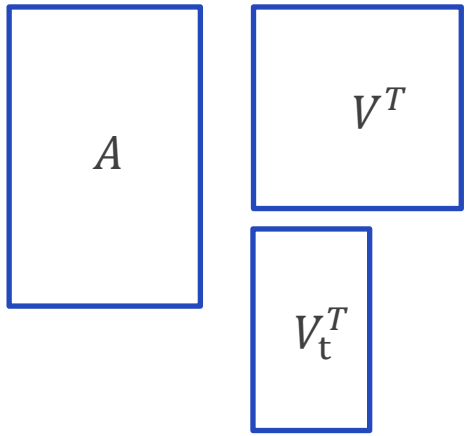
□ PCA의 원리

- 공분산행렬의 특이행렬로 기저변환을 하면, 데이터의 분산을 유지한 채 낮은 차원의 데이터로 변환하는 것이 가능하다.

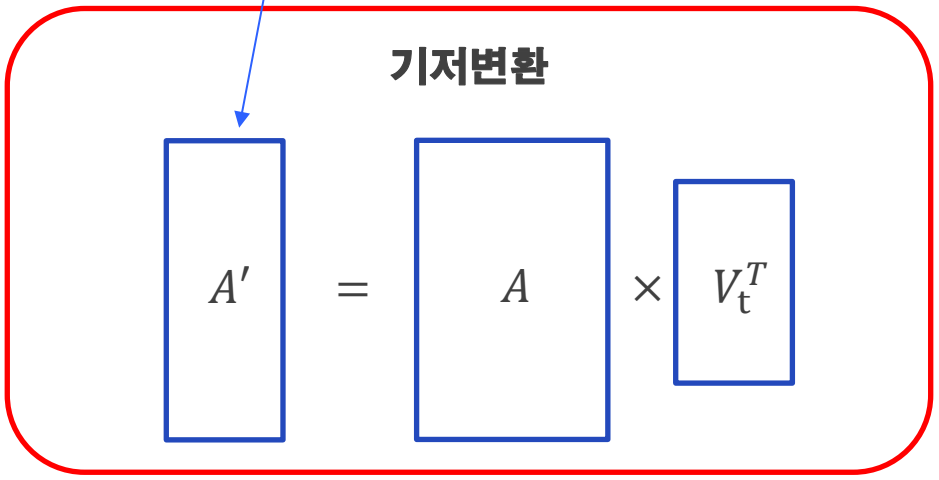
4) 기저변환

$$new\ A' = AV_t^T$$

차원을 줄이기 위해 V^T대신 V_t^T를 이용하여 기저변환을 한다!



차원축소의 결과



PCA

□ PCA의 원리

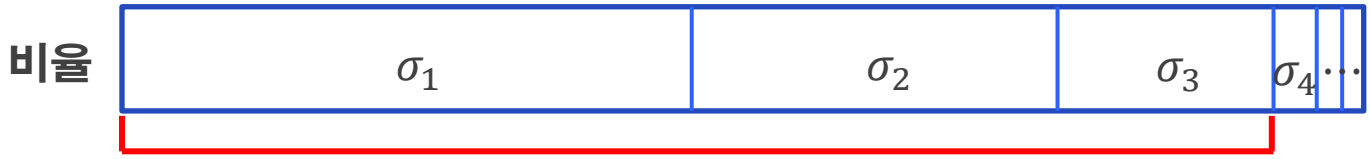
- 공분산행렬의 특이행렬로 기저변환을 하면, 데이터의 분산을 유지한 채 낮은 차원의 데이터로 변환하는 것이 가능하다.

>> 여기서 "t"의 크기는 어떻게 구할까?

만약 고윳값들이 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ 를 만족한다면,

$$A = \sigma_1 x_1 y_1 + \sigma_2 x_2 y_2 + \cdots + \sigma_{n-1} x_{n-1} y_{n-1} + \sigma_n x_n y_n$$

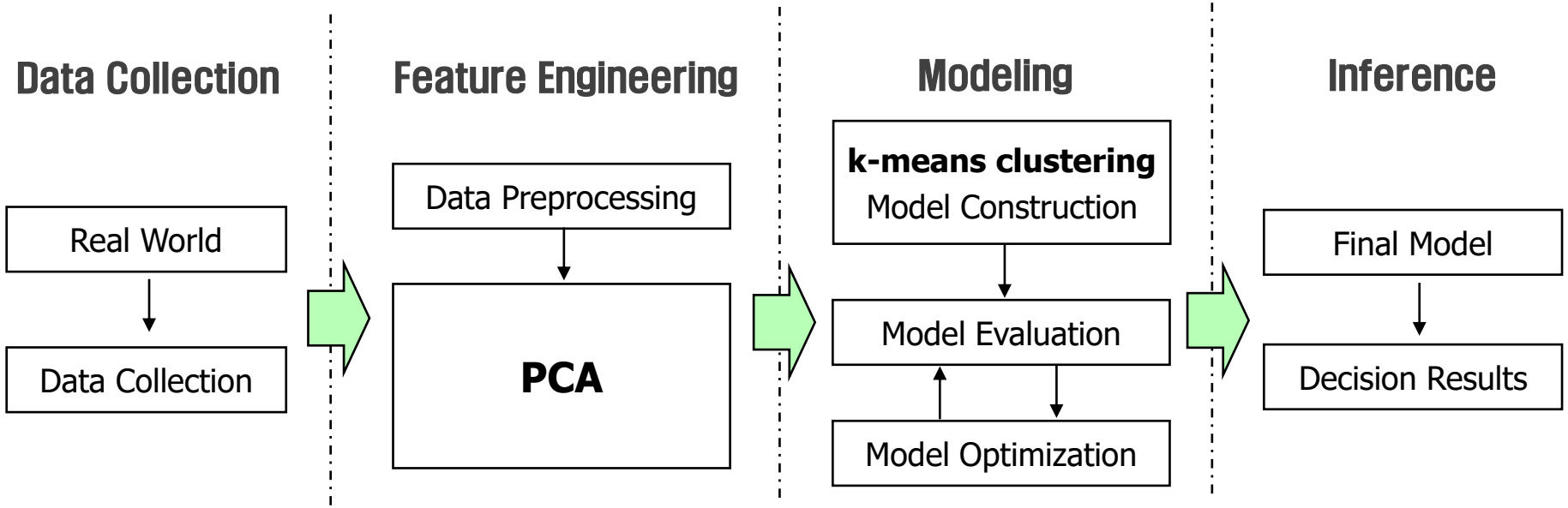
중요함 ← 특징의 중요도 → 무시해도 됨



약 90% 정도를 선택 → $\sigma_1, \sigma_2, \sigma_3$ $(t = 3)$만큼 사용함

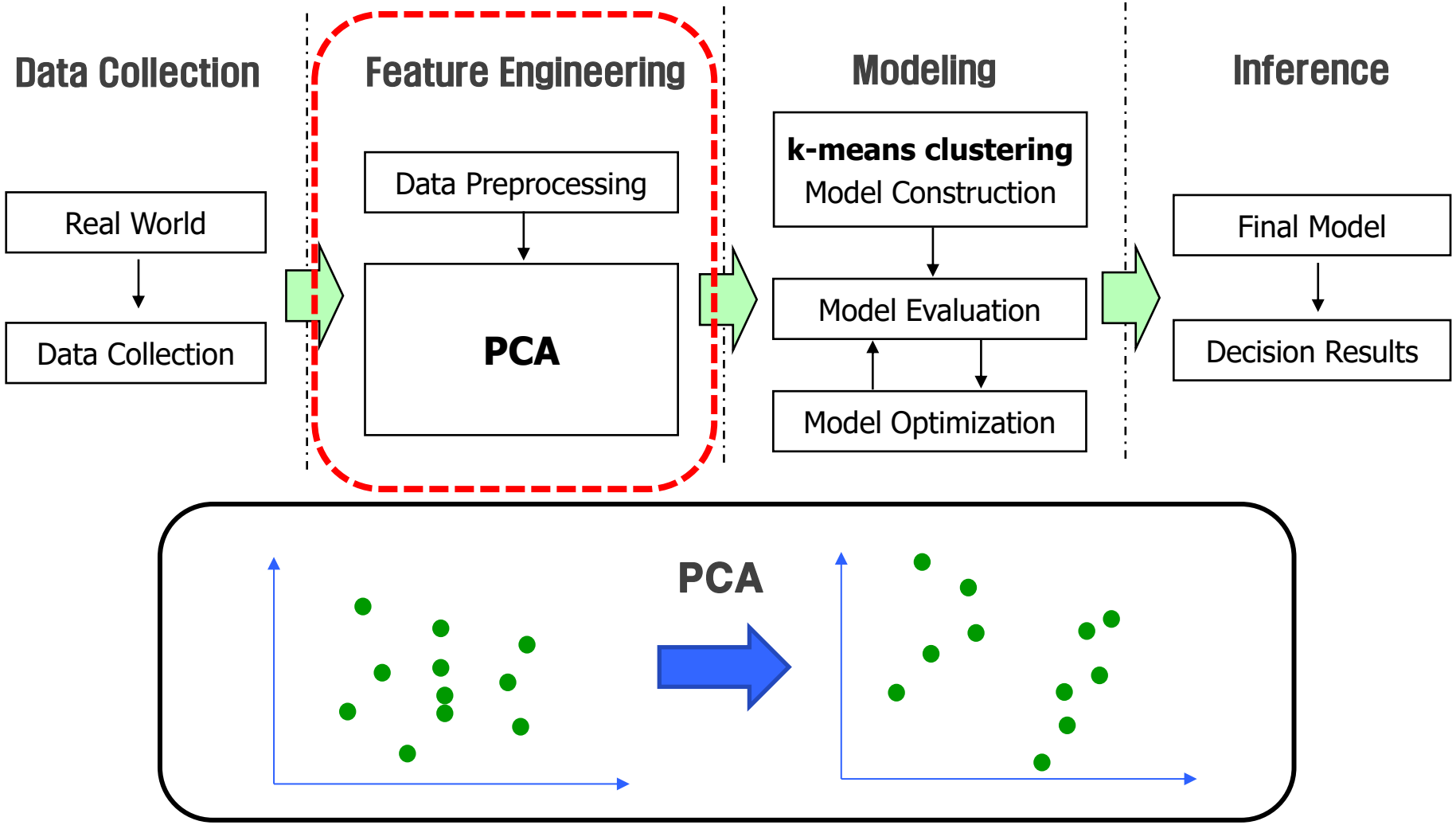
머신러닝

□ PCA와 클러스터링을 이용한 머신러닝 과정 예시



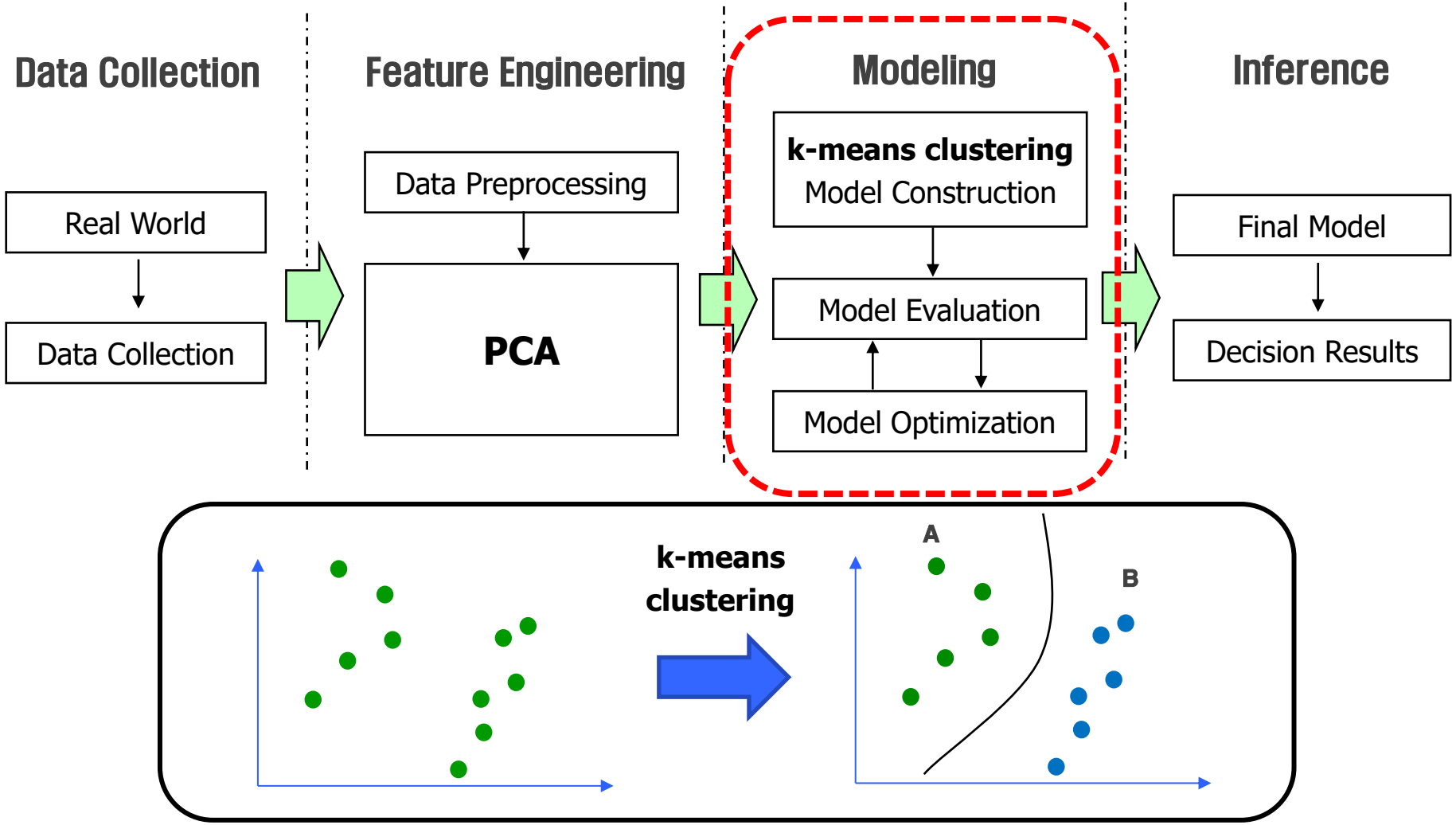
머신러닝

□ PCA와 클러스터링을 이용한 머신러닝 과정 예시



머신러닝

□ PCA와 클러스터링을 이용한 머신러닝 과정 예시



머신러닝

□ PCA와 클러스터링을 이용한 머신러닝 과정 예시