Supervised Learning.

Unsupervised Learning (3:10 ~ 4:30)

Reinforcement Learning

Recommend System.

⋮

Unsupervised Learning. : Clustering Algorithm.

Dimension Reduction.

anormal detection.

Density Estimation.

Clustering Algorithm. K-means clustering

hierarchical clustering.

⋮

Dimensional Reduction PCA.

T-SNE.

NMF

⋮

* K-means Clustering.

Notation. n : dimension of data.
k : number of cluster.

X = data set

$= \{ x_1, x_2, \cdots x_n \}$ where $x_i \in \mathbb{R}^n$.

C_i = i-th cluster, $C_i \subset X$. & $C_i \cap C_j = \phi$.
& $X = C_1 \cup \cdots \cup C_k$.

c_i = center of cluster C_i

Example. $n=2$
$k=3$.

$X = \{ (0,0), (2,0), (3,1), (4,3), (5,2) \}$.

$C_1 = \{ (0,0), (2,0) \}$

$C_2 = \{ (3,1), (5,2) \}$

$C_3 = \{ (4,3) \}$

$\Rightarrow$

$c_1 = (1,0)$

$c_2 = (4, 1.5)$

$c_3 = (4, 3)$

Want to find C_i

which minimize $\displaystyle\sum_{i=1}^{K} \sum_{x \in C_i} \| x - c_i \|^2$

Denote Cost $= \displaystyle\sum_{i=1}^{K} \sum_{x \in C_i} \| x - c_i \|^2$.

In the above example.

$$
\begin{aligned}
\text{Cost} &= \left((0-1)^2 + (0-0)^2 \right) + \left((2-1)^2 + (0-0)^2 \right) \\
&\quad + \left((3-4)^2 + (1-1.5)^2 \right) + \left((5-4)^2 + (2-1.5)^2 \right) \\
&\quad + \left((4-4)^2 + (3-3)^2 \right). \\
&= \star
\end{aligned}
$$

For fixed K, n.

How do we find the $C_1 \cdots C_K$. ?.

전수조사 → ?!

Proposed Algorithm

Step 1. Set the initial $C_1 \cdots C_k$ Randomly.

Step 2. 각 data point를 가까운 Center point를 갖는 cluster C_i에 분배.

Step 3. 클러스터 중심 재설정.

Step 4. Cluster C_i 들이 변하지 않을때까지 (C_i 위치에 차이가 없을때까지.)
step2, step3 을 반복한다.

In the example.

$$X = \{ (0,0), (2,0), (3,1), (4,3), (5,2) \}.$$

Let $K=2$, $\quad c_1 = (0,0)$

$c_2 = (2,0)$.

$$\|x_1 - c_1\|^2 = 0. \quad \Rightarrow \quad x_1 \in C_1.$$

$$\|x_1 - c_2\|^2 = 4.$$

$$\|x_2 - c_1\|^2 = 4 \quad \Rightarrow \quad x_2 \in C_2$$

$$\|x_2 - c_2\|^2 = 0$$

$$\vdots$$

$$\Rightarrow \quad x_3, x_4, x_5 \in C_2.$$

$\|x_1 - c_1\|^2 = 0 \quad \Rightarrow \quad x_1 \in C_1$

$\|x_2 - c_1\|^2 = 4$

$\|x_2 - c_2\|^2 = 2.25 + 2.25 = 4.5 \quad \Rightarrow \quad x_2 \in C_1$

$\|x_3 - c_1\|^2 = 9 + 1 = 10$

$\|x_3 - c_2\|^2 = 0.25 + 0.25 = 0.5 \quad \Rightarrow \quad x_3 \in C_2.$

$\|x_4 - c_1\|^2 = 16 + 9 = 25.$

$\|x_4 - c_2\|^2 = 0.25 + 2.25 = 2.5 \quad \Rightarrow \quad x_4 \in C_2$

$\|x_5 - c_1\|^2 = 25 + 4 = 29$

$\|x_5 - c_2\|^2 = 2.25 + 0.25 = 2.5 \quad \Rightarrow \quad x_5 \in C_2.$

$C_1 = \{(0,0), (2,0)\}.$

$C_2 = \{(3,1), (4,3), (5,2)\}.$

$c_1 = (1,0) \qquad c_2 = (4,2)$

$$\|x_1 - c_1\|^2 = 1 \quad \Rightarrow \quad x_1 \in C_1$$

$$\vdots$$

$$x_2 \in C_1$$

$$x_3 \in C_2$$

$$x_4 \in C_2$$

$$x_5 \in C_2$$

$$\therefore \quad X = \{(0,0), (2,0), (3,1), (4,3), (5,2)\}.$$

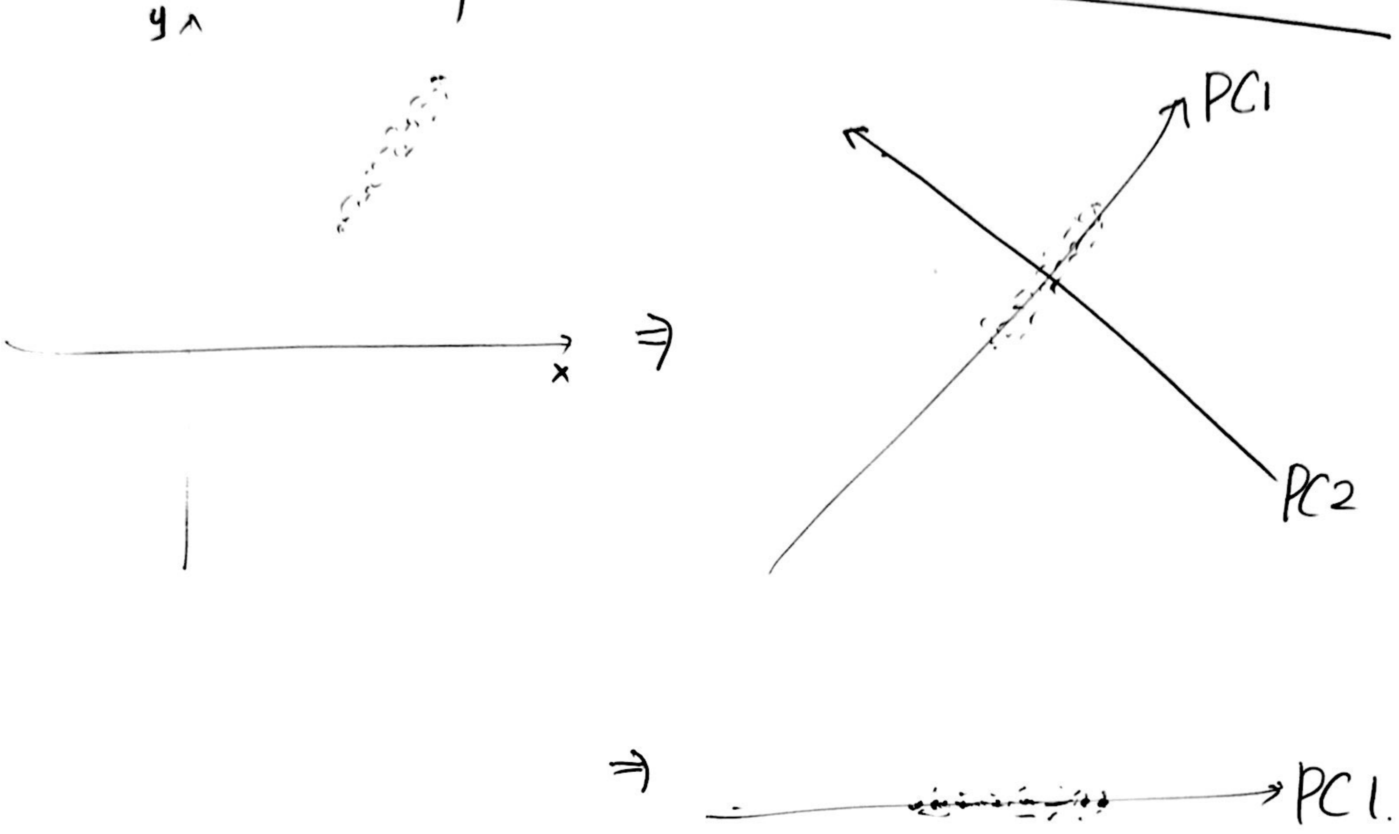
$$= C_1 \cup C_2$$

$$= \{(0,0), (2,0)\} \cup \{(3,1), (4,3), (5,2)\}.$$

Dimensional Reduction.

Principal Component Analysis.

Summary



데이터가 주어져 있다. X.

공분산 행렬 Cov(X)를 구한다.

Cov(X)의 eigenpair를 찾는다.

높은 eigenvalue 값을 갖는 eigenvector를 몇개 선택한다.

그 eigenvector로 data를 다시 표현한다.

1. Matrix 이해

2. Eigenpair 이해

3. Covariance Matrix 이해

4. PCA 이해.

---

1. Matrix 이해.

<u>Definition</u>

A linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$
is a map $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that the following hold:

1. $T(x+y) = T(x) + T(y) \quad \forall x, y \in \mathbb{R}^n$
2. $T(cx) = cT(x)$ for some x, $\forall c \in \mathbb{R}$.

Theorem.

Let $\{v_1, v_2, \cdots, v_n\}$ be a basis on $\mathbb{R}^n$

and $\{u_1, \cdots, u_n\}$ be a arbitrary vectors in $\mathbb{R}^m$

Then there exists a unique linear mapping $F: \mathbb{R}^n \to \mathbb{R}^m$

Such that $F(v_i) = u_i$

Consider standard basis on $\mathbb{R}^2$.

$\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \}$.

and arbitrary vector in $\mathbb{R}^2$

$\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \}$

$$\Rightarrow F = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

$\because F\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$ & the uniqueness.

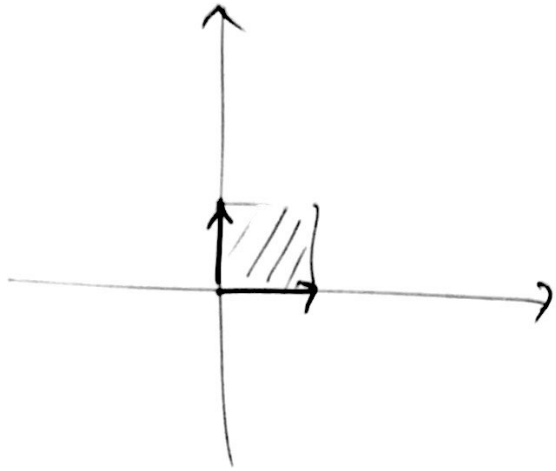
$F\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix}$

Note. Matrix as a linear transformation.

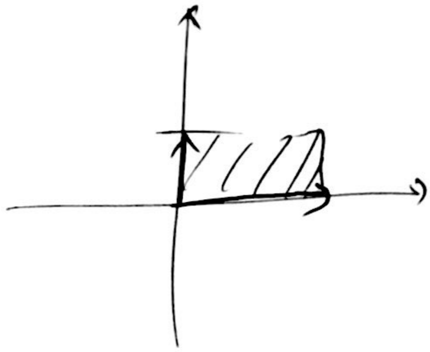
send e_i to i-th column vector.

$T: \mathbb{R}^2 \to \mathbb{R}^2$ 의 종류.

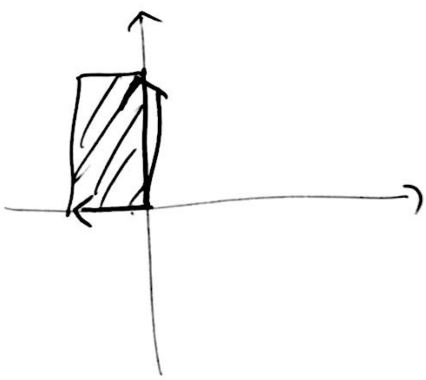
(a) $T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.



(b) $T = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$



(c) $T = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$



⋮

Note.
늘리고 돌리고 대칭이동하고 등등의 변환을 만든다.

Note.
e_1 e_2 를 잡고 늘리는 어떤 변환으로 바라볼수 있다.

matrix가 어떤 선형변환인지 알기위해서.

$\{e_1, e_2, \cdots e_n\}$을 어디에 대응시키는지 보면된다.

각 e_i들을 잡아서 늘리거나 늘리는등의 작업으로 만들어지는 변환.

으로 볼 수 있겠다.

이 insight를 보다 분명히 설명해주는것이 eigenpair 이다.

Definition

For given $n \times n$ matrix A,
we can define eigenpair as a solution of the following.

$$A\mathbb{x} = \lambda \mathbb{x} \quad \text{for some } \mathbb{x} \in \mathbb{R}^n.$$

we called λ is eigenvalue.
$\mathbb{x}$ is eigenvector.
$(\lambda, \mathbb{x})$ is eigenpair.

Note.

$$Ax = \lambda x \Rightarrow A(\alpha x) = \lambda(\alpha x).$$

i.e. (λ, x) is eigenpair

$\Rightarrow$ $\cancel{E\alpha}$ $(\lambda, \alpha x)$ is eigenpair.

We may assume that.

for a eigenvector x,

$\|x\| = 1$.

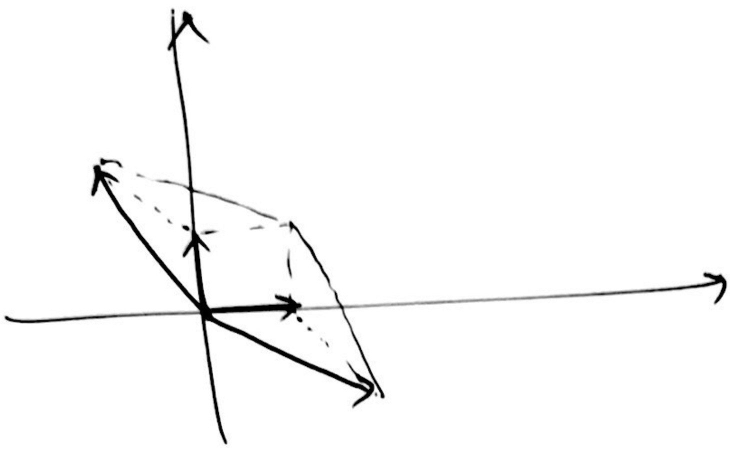
Note.

eigenvalue 와 eigenvector 는.

선형변환으로서의 행렬을 통해 회전이 일어나지 않는 방향과

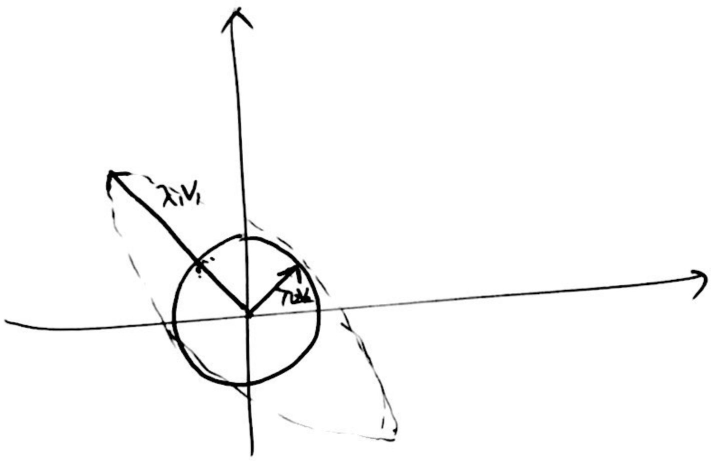
그 방향으로의 scaling 값을 의미한다.

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$



$\lambda_1 = 3, \quad v_1 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

$\lambda_2 = 1, \quad v_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$



Theorem

Linear Transformation

send disk to ellipsoid.

Spectral Theorem

A real $n \times n$ symmetric matrix
has n- orthogonal eigenvector with real eigenvalue.

---

3. Covariance Matrix.

Definition

Covariance of the variables X and Y
is given by $Cov(X, Y) = \sum_{i=1}^{n} (X^{(i)} - \bar{x})(Y^{(i)} - \bar{y})$

where $X^{(i)}, Y^{(i)}$ is the i-th ~~variables~~ data.
$\bar{x}, \bar{y}$ is the mean of X, Y.

Definition.

Covariance matrix S for given data set X.
is defined by

$$S = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_m) \\ \vdots & & \vdots \\ Cov(X_m, X_1) & \cdots & Cov(X_m, X_m) \end{bmatrix}$$

Note S is a symmetric matrix.

Note 공분산이 높다 $\Longleftrightarrow$ 두 변수가 같이 증가,
같이 감소하는 경향이 있다.

Note S는 e_i들을 각각 변수 X_i가
다른 변수와 상관이 없는 정도로 보낸다.

Note eigenvector of S는
이러한 Linear transformation의 장축, 단축 등의 역할을 한다.

---

4. PCA 이해

1. 데이터가 주어져 있다. X.
2. 공분산 행렬 Cov(X)를 구한다.
3. Cov(X)의 eigenpair를 찾는다.
4. 높은 eigenvalue 값을 갖는 eigenpair를 몇개 선택한다.
5. data X를 해당하는 eigenvector를 이용하여 다시 표현한다.

⇒
- 2-dimensional graph화 가능.
- data의 새표현으로 모델의 정확도 향상 기대.

:

?