Dimensionality Reduction

D features

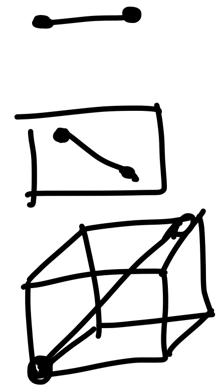
N data points

$$\underbrace{X}_{\text{data matrix}} \quad N \times D \xrightarrow{\text{dimensionality reduction}} Z$$

D' features

N data points

usually $D' << D$



Reasons

1. to reduce computational complexity
2. to reduce storage complexity
3. to reduce data acquisition cost
4. to increase robustness
5. to increase interpretability
6. to enable visualization (when $D' = 2$ or $D' = 3$)

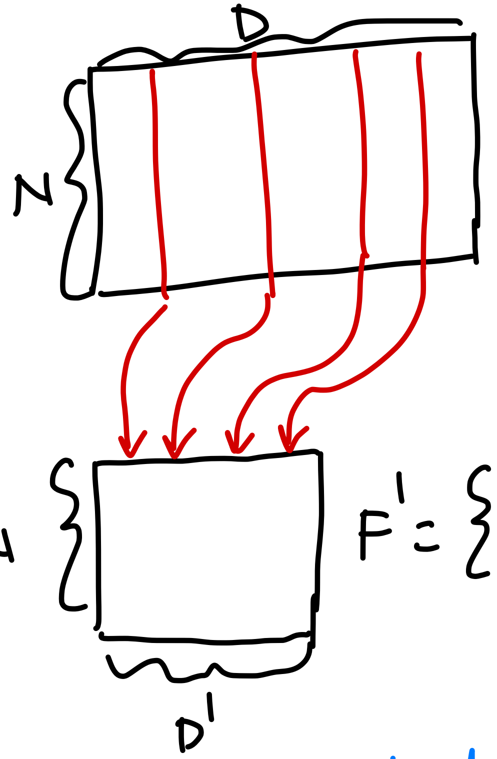
⋮

Feature Selection

$\mathcal{X} = \{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^D$

We will select a subset of

$\{1, 2, \dots, D\}$



$F = \{1, 2, \dots, D\}$

$F' = \{$ selected column indices $\}$

# of possible subsets

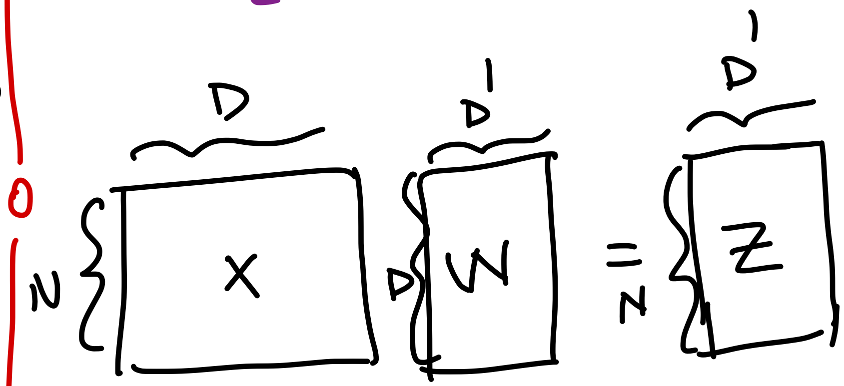
$$= 2^D \underbrace{- 1}_{\text{empty set}} \underbrace{- 1}_{\text{full set}}$$

Feature Extraction

$\mathcal{X} = \{x_i\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^D$

$x_i \in \mathbb{R}^D \longrightarrow z_i \in \mathbb{R}^{D'}$

$$\underbrace{z_i}_{[D' \times 1]} = \underbrace{W^T}_{[D' \times D]} . \underbrace{x_i}_{[D \times 1]} \qquad z_i = \Phi(x_i)$$



$W \Rightarrow$ model parameters

W

$$\begin{bmatrix} 1 & +2 \\ -2 & +2 \\ +1 & -3 \\ -1 & -1 \end{bmatrix} \qquad \begin{bmatrix} 1 & -1 \\ -2 & +2 \\ +1 & -1 \\ -1 & +1 \end{bmatrix}$$

① Forward Selection

this is calculated on a set of data points not included in training.

- $F = \emptyset$ (empty set)
- At each iteration, find the best new feature to be added to F'.

$$d^* = \arg\min_d \text{Error}(F' \cup d)$$

- Add d^* to F' if $\text{Error}(F' \cup d) < \text{Error}(F')$

$F' = \emptyset$

$t=1 \Rightarrow$ 1 2 3 4 5 6 $\Rightarrow F' = \{1\}$

$t=2 \Rightarrow$ $\times$ (1,2) (1,3) (1,4) (1,5) (1,6) $\Rightarrow F' = \{1,4\}$

if $\text{Error}(\{1,4\}) < \text{Error}(\{1\}) \Rightarrow$ YES

$t=3 \Rightarrow$ $\times$ (1,4,2) (1,4,3) $\times$ (1,4,5) (1,4,6) $\Rightarrow \boxed{F' = \{1,4,5\}}$

if $\text{Error}(\{1,4,5\}) < \text{Error}(\{1,4\}) \Rightarrow$ YES

$t=4 \Rightarrow$ $\times$ (1,4,5,2) (1,4,5,3) $\times$ $\times$ (1,4,5,6) STOP

if $\text{Error}(\{1,4,5,2\}) < \text{Error}(\{1,4,5\}) \Rightarrow$ NO

② Backward Elimination

- $F' = F$ (full set)
- At each iteration, find the best feature to be removed from F'. $d^* = \arg\min_d \text{Error}(F'/d)$
- Remove d^* from F' if $\text{Error}(F'/d) < \text{Error}(F')$

$F' = \{1, 2, 3, 4, 5, 6\}$

$t = 1 \Rightarrow$ $\{2,3,4,5,6\}$ $\{1,3,4,5,6\}$ $\{1,2,4,5,6\}$ $\{1,2,3,5,6\}$ $\{1,2,3,4,6\}$ $\{1,2,3,4,5\}$

if $\text{Error}(\{1, 3,4,5,6\}) < \text{Error}(\{1,2,3,4,5,6\}) \Rightarrow$ YES

$$\boxed{F' = \{1,3,4,5,6\}}$$

$t = 2 \Rightarrow$ $\{3,4,5,6\}$ × $\{1,4,5,6\}$ $\{1,3,5,6\}$ $\{1,3,4,6\}$ $\{1,3,4,5\}$

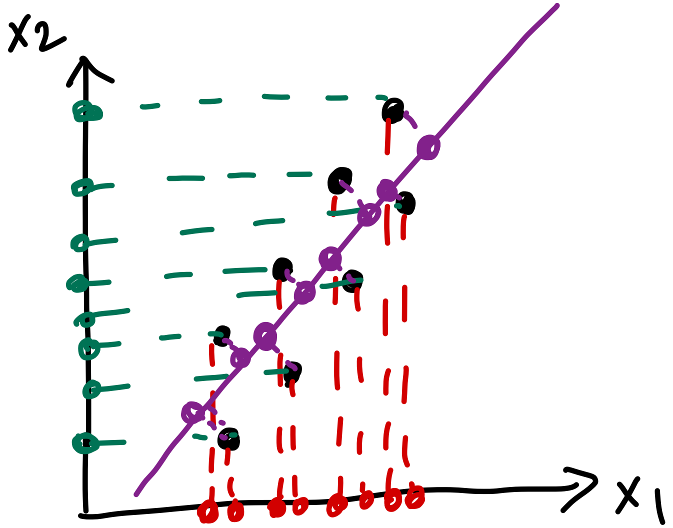
if $\text{Error}(\{1,4,5,6\}) < \text{Error}(\{1,3,4,5,6\}) \Rightarrow$ NO

STOP

Principal Component Analysis (PCA)

- PCA is a feature extraction algorithm

$z = W^T . x \qquad x \in \mathbb{R}^D \qquad z \in \mathbb{R}^{D'} \qquad W \in \mathbb{R}^{D \times D'}$



We would like to find the direction that maximizes the variance.

$$VAR(z) = VAR(W^T . x)$$

$$= \underbrace{W^T}_{1\times 2} . \underbrace{VAR(x)}_{2\times 2} . \underbrace{W}_{2\times 1} = 1\times 1$$

$$= W^T . \Sigma_x . W$$

Σ_x → Covariance matrix of original data points

$$\boxed{\begin{aligned} &VAR(2x) \\ &= 4VAR(x) \end{aligned}}$$

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

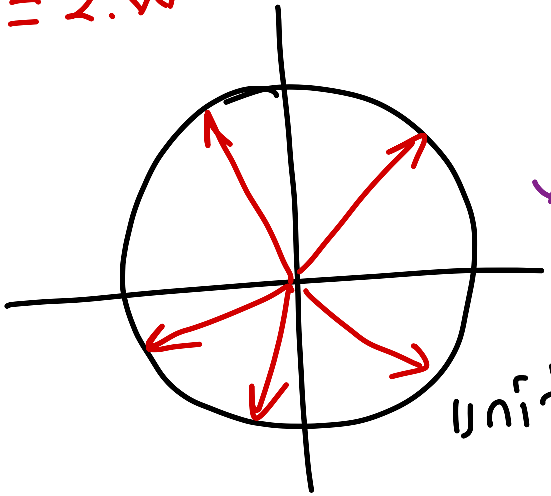
maximize $w^T.\Sigma_x.w$

assume w^* is the optimal solution

$\tilde{w} = 2.w^*$

$(\tilde{w})^T.\Sigma_x.\tilde{w}$

$4.w^{*T}.\Sigma_x.w^*$



unit circle

$\|w\|_2 = 1$

$w_1^2 + w_2^2 = 1$

maximize $w^T.\Sigma_x.w$

subject to $\|w\|_2^2 = 1$] α

$$L_p = w^T.\Sigma_x.w - \alpha(\|w\|_2^2 - 1)$$

$$= w^T.\Sigma_x.w - \alpha.(w^T.w - 1)$$

$$\frac{\partial L_p}{\partial w} = \cancel{2}.\Sigma_x.w - \cancel{2}.\alpha.w = 0$$

$$\Downarrow$$

$$\boxed{\Sigma_x.w = \alpha.w}$$

$$Ax = bx$$

matrix, vector, scalar, vector

eigenvector of A

eigenvalue of A

D eigenvalues

$\alpha_1, \alpha_2, \ldots, \alpha_D \Rightarrow \alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_D$

$w^* \Rightarrow$ the eigenvector that corresponds to the largest eigenvalue (α_1) [the first eigenvector]