Representation learning : Learn useful representation

1) Dimensionality Reduction

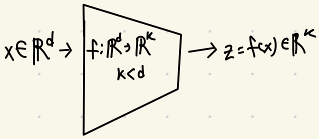
2) Clustering / Compression

3) Generative learning

Unsupervised Learning : Extract useful information ☆ Data is about variance

---

Dimensionality Reduction - Big Idea



1) Generally speaking, we need f to preserve as much information about x as possible

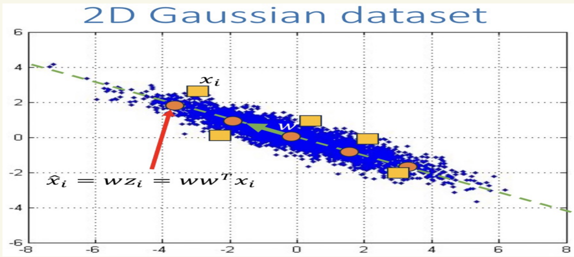
$\|z^{(i)} - z^{(j)}\|^2 \approx \|x^{(i)} - x^{(j)}\|^2$

$f(x) = w^T x$

$z = w^T x^{(i)}$ s.t $\|w\| = 1$

① Find w s.t maximizes the variance of the projected data

② Find w s.t minimizes the reconstruction error



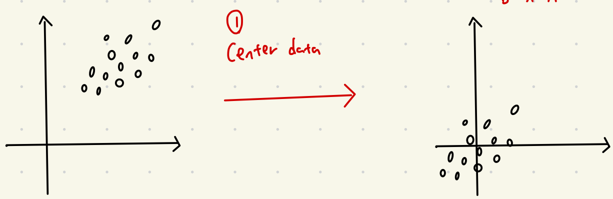
① = ②

What's $w^T x$? : Reduced data

Principal Component Analysis (PCA)

Given data : X

$B = X - \bar{X}$



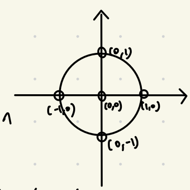
Centering data : Subtract mean of each variable from the dataset

② Covariance Matrix Σ : $\Sigma = B^T B$

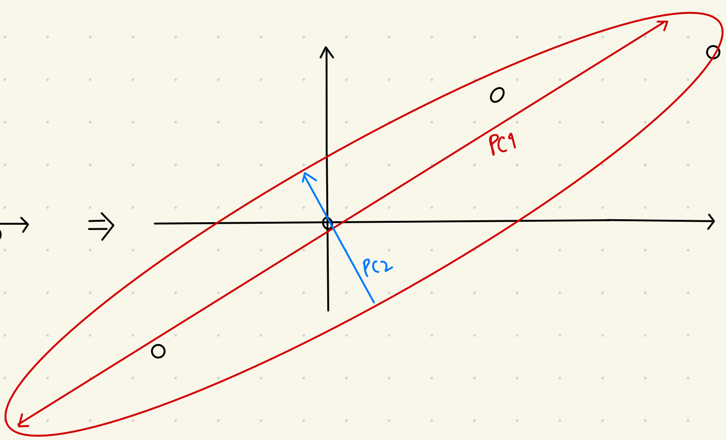
$$\Sigma = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}$$

Ex)

Let $\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$, then



$\Rightarrow$



$(x,y) \to (9x+4y, 4x+3y)$

$(0,0) \to (0,0)$

$(1,0) \to (9,4)$

$(0,1) \to (4,3)$

$(-1,0) \to (-9,-4)$

$(0,-1) \to (-4,-3)$

③ Compute eigenvectors & eigenvalues $Sv = \lambda v$

④ Sort eigenvectors w.r.t eigenvalues in descending order (The eigenvalue indicate the variance of the Projected data)

⑤ Take top k eigenvectors

How these steps are related to finding w that maximizes variance of projected data?

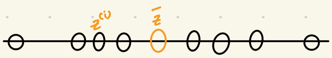
1) The reduced data is $\{z^{(i)} = w^T x^{(i)}\}_{i=1}^{N}$

2) Calculate the mean of the reduced data

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z^{(i)} = \frac{1}{N}\sum_{i=1}^{N} w^T x^{(i)} = w^T\left(\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\right) = w^T\bar{x}$$

3) Calculate the variance of the reduced data

$$Var = \frac{1}{N}\sum_{i=1}^{N}(z^{(i)} - \bar{z})^2 = \frac{1}{N}\sum_{i=1}^{N}(w^T x^{(i)} - w^T\bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(w^T(x^{(i)} - \bar{x})\right)^2$$



$$= \frac{1}{N}\sum_{i=1}^{N} w^T(x^{(i)} - \bar{x})(x^{(i)} - \bar{x})w$$

$$= w^T\underbrace{\left(\frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \bar{x})(x^{(i)} - \bar{x})\right)}_{S}w = w^T S w$$

Hence, maximizing the variance can be written as

$$\max_w w^T S w \quad \text{s.t } \|w\| = 1 \Rightarrow \max_w w^T S w \quad \text{s.t } w^T w = 1$$

Use Lagrangian function

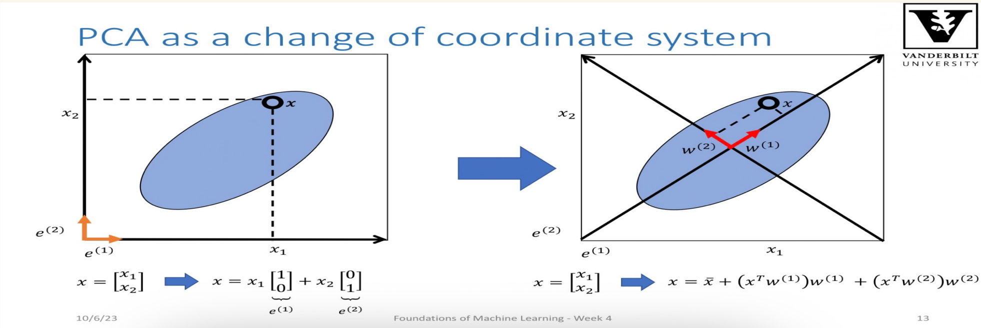
$$\mathcal{L}(w, a) = w^T S w + a(1 - w^T w) = w^T S w + a - a w^T w$$

$$\frac{\partial \mathcal{L}}{\partial w} = 2Sw - 2aw = 0$$

$$= Sw - aw = 0$$

$= Sw = aw$: At $\frac{\partial \mathcal{L}}{\partial w} = 0$, we derive to equation for getting eigen vectors & eigen values

★ $S = \frac{1}{N}\sum_{i=1}^{N}(\underbrace{x^{(i)} - \bar{x}}_{\text{Centering data}})^2$



From Vanderbilt University Lecture note

Reconstruction : $\bar{X} + ww^T x^{(i)} - ww^T\bar{X}$

$$= \bar{X} + wz^{(i)} - w\bar{z}$$