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

$x \in \mathbb{R}^d \rightarrow \boxed{f : \mathbb{R}^d \rightarrow \mathbb{R}^k,\ k<d} \rightarrow z = f(x) \in \mathbb{R}^k$

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

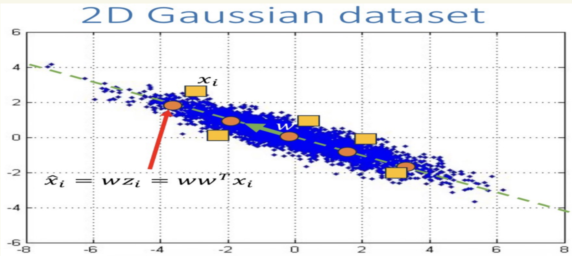
$\| z^{(i)} - z^{(j)} \|^2 \approx \| x^{(i)} - x^{(j)} \|^2$

$f(x) = w^T x$

$z = w^T x^{(i)}$ 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

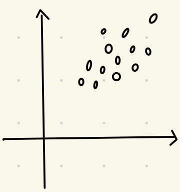


① = ②

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

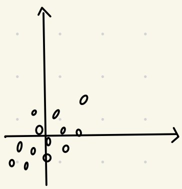
Principal Component Analysis (PCA)

Given data : X



① Center data

$B = X - \bar{X}$



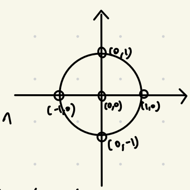
Centering data : Subtract mean of each variable from the dataset

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

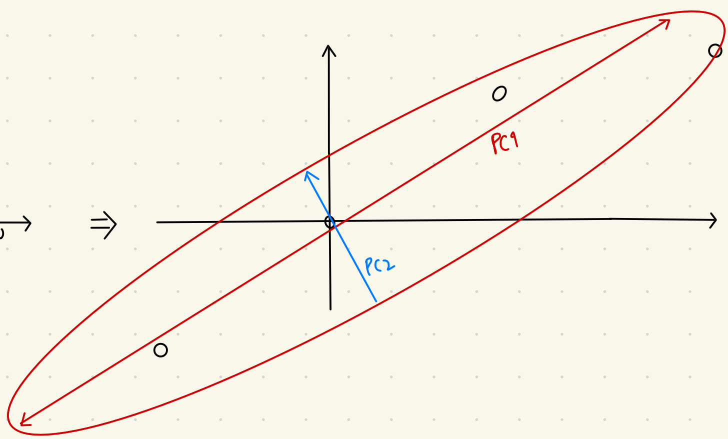
$$\Sigma = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}$$

Ex)

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



$\Rightarrow$



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

$(0,0) \rightarrow (0,0)$

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

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

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

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

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?

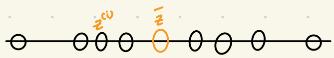
1) The reduced data is $\{z^{(i)} = w^T x^{(i)}\}_{i=1}^{N}$

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

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

Use Lagrangian function

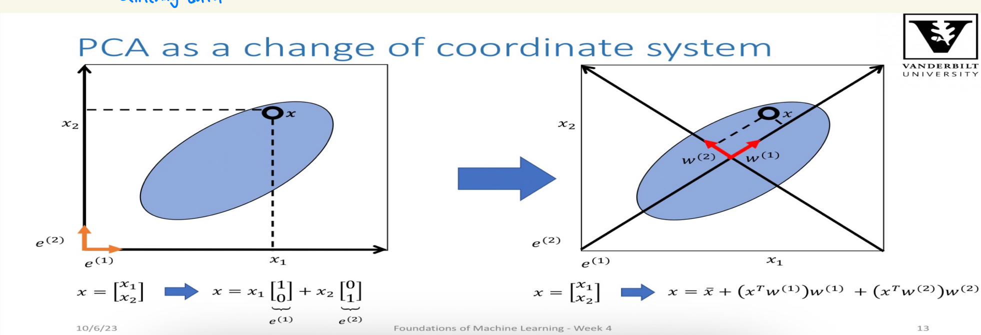
$$L(w, a) = w^T S w + a(1 - w^T w) = w^T S w + a - a w^T w$$

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

$$= Sw - aw = 0$$

$= Sw = aw$: At $\frac{\partial L}{\partial w} = 0$, 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 : $\bar{X} + ww^T x^{(i)} - ww^T\bar{X}$

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

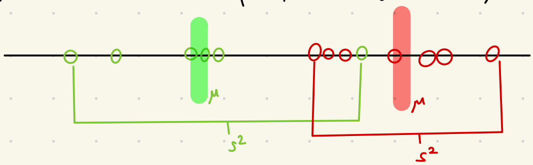
Linear Discriminant Analysis Derivations (LDA) : Supervised dimensionality reduction

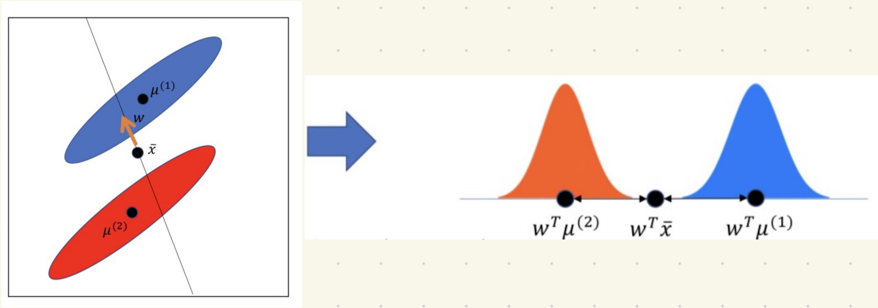
: LDA focuses on maximizing the separability among known categories

How LDA creates a new axis?

1) Maximize the distance between means

2) Minimize the variation (which LDA calls "scatter", and is represented s^2) within each categories.





① Distance between class means * K: # of classes

$$: \frac{1}{K}\sum_{K}\left(w^T\mu^{(k)} - w^T\bar{x}\right)^2 = \frac{1}{K}\sum_{K}\left(w^T\left(\mu^{(k)} - \bar{x}\right)\right)^2$$

$$= w^T\underbrace{\left(\frac{1}{K}\sum_{k}\left(\mu^{(k)} - \bar{x}\right)\left(\mu^{(k)} - \bar{x}\right)^T\right)}_{S_b}w = w^T S_b w$$

② Within class variance

$$: \frac{1}{N}\sum_{k}\sum_{i \in C_k}\left(w^T x^{(i)} - w^T\mu^{(k)}\right)^2$$

$$= w^T\underbrace{\left(\frac{1}{N}\sum_{k}\sum_{i \in C_k}\left(x^{(i)} - \mu^{(k)}\right)\left(x^{(i)} - \mu^{(k)}\right)^T\right)}_{S_w}w = w^T S_w w$$

LDA Objective : $\underset{w}{\operatorname{argmax}} \dfrac{w^T S_b w}{w^T S_w w}$ s.t. $\|w\| = 1$ is equivalent to $\underset{w}{\operatorname{argmax}}\, w^T S_b w$ s.t. $w^T S_w w = 1$

Lagrangian function:

$$\mathcal{L}(w, \lambda) = w^T S_b w - \lambda\left(w^T S_w w - 1\right)$$

$$\nabla_w \mathcal{L} = 2S_b w - 2\lambda S_w w = 0$$

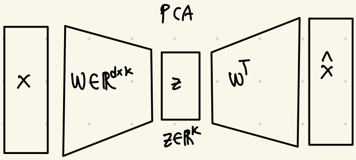
$$\Rightarrow S_b w - \lambda S_w w = 0$$

$$\Rightarrow \left(S_w^{-1} S_b\right) w = \lambda w$$

Hence, finding eigenvector & eigenvalues of $S_w^{-1} S_b$

Auto-Encoder (AE).

PCA vs AE



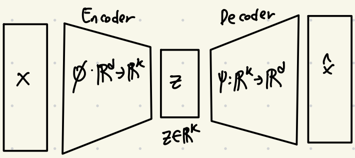
PCA: Linear dimensionality Reduction

- Forward transform : $z = W^T x$
- Inverse transform: $\hat{x} = Wz$

Objective: Maximizing Variance or Minimizing Reconstruction Error (they're equivalent)

$$\min_W \mathbb{E}_x\left[\|X - \hat{X}\|^2\right] = \mathbb{E}_x\left[\|X - WW^T X\|^2\right]$$
$$\text{s.t. } W^T W = I_{k \times k}$$

Auto Encoder

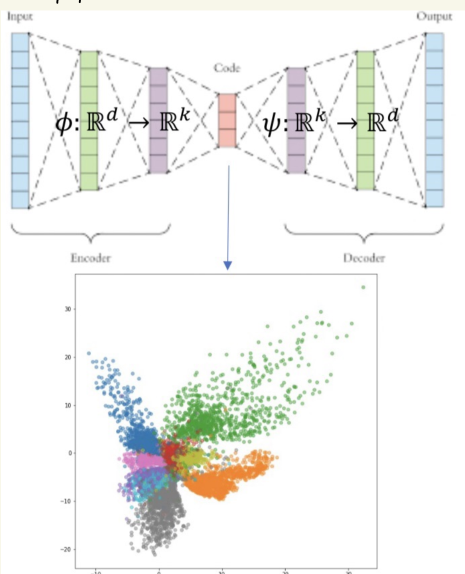


AE: Nonlinear dimensionality Reduction

- Forward transform: $z = \phi(x)$
- Inverse transform : $\hat{x} = \psi(x)$

Objective: Minimizing Reconstruction Error.

$$\min_{\phi, \psi} \mathbb{E}_x\left[\|X - \hat{X}\|^2\right] = \mathbb{E}_x\left[\|X - \psi(\phi(x))\|^2\right]$$



From Vanderbilt University Lecture note

- Reconstruction Loss: $\mathbb{E}_x\left[\|x - \psi(\phi(x))\|^2\right]$
- $\hat{y}$ = Likelihood = $\text{Softmax}(W^T \phi(x))$
- Negative Log-likelihood = $-\sum_i y^{(i)} \log(y^{(i)})$, which is the Cross entropy.