Unsupervised learning

PCA reduction of features, K-means is reduction of data

$\{x_i, y_i\}$
$\{x_i\}$

1) PCA : Principal component analysis

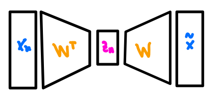
* Dimensionality reduction

$$x = \begin{bmatrix} \\ \\ \end{bmatrix} \in \mathbb{R}^d \xrightarrow{f} \underbrace{z}_{\text{latent variable}} = \begin{bmatrix} \\ \\ \end{bmatrix} \in \underbrace{\mathbb{R}^l}_{\text{latent space}} \quad \text{with } l < d$$

* Mapping f as:
  - Parametric model : $z = f(x; \theta)$
    Preprocessing step for other learning algorithms
  - Non-parametric model :
    $\underbrace{z_n}_{\text{embedding}} = f(x_n) \Leftrightarrow$ Data visualization
    $\{$ order Reduction + simple linear classifier

* Also named discrete Karhunen-Loève transform (KLT)
* Eckart-Young theorem
* Most widely used form of dimensionality reduction
* Idea : Find linear and ortogonal mapping $W \in \mathbb{R}^{d \times l}$ such that $z = W^T x \in \mathbb{R}^l$ is a good approximation of $x \in \mathbb{R}^d$ in the sense that decoding $\tilde{x} = Wz \in \mathbb{R}^d$ is similar to the original data $x \in \mathbb{R}^d$. This is find : $W \in \mathbb{R}^{d \times l}$

$$\min_W \ \mathcal{L}(W) = \frac{1}{N} \sum_{n=1}^{N} \| x_n - \underbrace{W \underbrace{W^T x_n}_{z}}_{\tilde{x}} \|_2^2 \leftarrow \text{Reconstruction error}$$



$W = d \begin{bmatrix} \overset{l}{} \end{bmatrix}$

$$\text{s.t.} \quad W^T W = I$$

$$d \begin{bmatrix} \tilde{x} \end{bmatrix} = \underbrace{\begin{bmatrix} W \end{bmatrix} \begin{bmatrix} W^T \end{bmatrix}}_{\text{low-rank approximation}} \begin{bmatrix} x \end{bmatrix}$$

(with $\underbrace{[W^T][x]}_{z}$)

$\Rightarrow$ In rows format : $[\ \tilde{x}\ ] = [\ x\ ] \begin{bmatrix} W \end{bmatrix} [\ W^T\ ]$

$z \equiv \{$ projected points / principal component / PC scores

* In matrix form : $X = N \begin{bmatrix} \overset{d}{\equiv} \end{bmatrix} \in \mathbb{R}^{N \times d}$; $Z = N \begin{bmatrix} \overset{l}{\equiv} \end{bmatrix} \in \mathbb{R}^{N \times l}$;
  Each sample is now written in rows

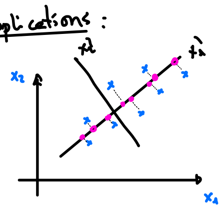
$$\min_W \ \mathcal{L}(W) = \frac{1}{N} \| X - \underbrace{XW}_{Z} W^T \|_F^2 \ ;$$

$$N \begin{bmatrix} \overset{d}{\hat{X}} \end{bmatrix} = N \underbrace{\begin{bmatrix} \overset{d}{X} \end{bmatrix} \begin{bmatrix} \overset{l}{W} \end{bmatrix}}_{Z} \begin{bmatrix} \overset{d}{W^T} \end{bmatrix}$$

$$\min_W \ \mathcal{L}(W) = \frac{1}{N} \| X^T - W \underbrace{W^T X^T}_{Z^T} \|_F^2 \ ;$$

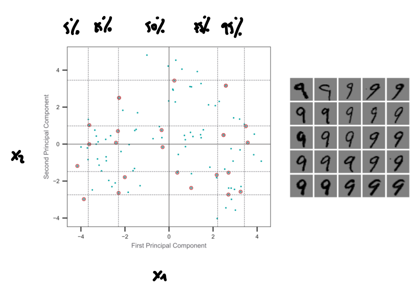
* Two applications :

1)



- "Find direction with maximum variance"
- "Rest of directions have limited info"
- "With one direction captured most part of the info"

2)



$x_1 \equiv$ "Orientation"
$x_2 \equiv$ "thickness"

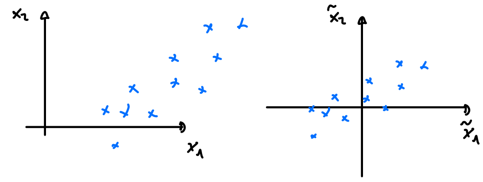
* Canonical form of the data

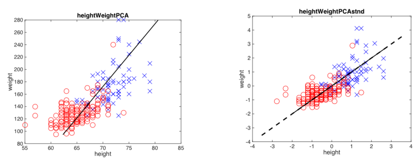
* Data processing:

For each feature:

* Empirical mean: $\mu_j = \frac{1}{N}\sum_i x_{ij}$

* Standard dev: $s_j^2 = \frac{1}{N-1}\sum (x_{ij}-\mu_j)^2$





(a) (b)

* Re-definition: $\hat{x}_{ij} = \frac{(x_{ij}-\mu_j)}{s_j}$; $\hat{\mu}_j = \frac{1}{N}\sum_i \frac{\overbrace{x_{ij}}^{\mu_j}-\mu_j}{s_j} = 0$

* Covariance matrix: $\Sigma_{ik} = \frac{1}{N}\sum_i (x_{ij}-\mu_j)(x_{ik}-\mu_k) = (X-\mu\otimes 1)^T(X-\mu\otimes 1)$; $\hat{\Sigma}_{ik} = \frac{1}{N}\sum_i x_{ij}x_{ik} \Rightarrow \hat{\Sigma} = \frac{1}{N}X^T X$

* Geometric interpretation

$$\min_{W} \mathcal{L}(W) = \frac{1}{N}\|X^T - WW^TX^T\|_F^2 \quad \text{s.t.} \quad W^TW = I \Rightarrow \min_{W} \frac{1}{N}\|(I-WW^T)X^T\|_F^2 \quad \text{s.t.} \quad W^TW = I$$

Note that: $\begin{cases} P_W = WW^T \text{ is a projection} \\ P_W : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d} \\ P_W^2 = P_W P_W = W\underbrace{W^TW}_{I}W^T = P_W \\ P_W^T = P_W \\ P_W^{\perp} = I - P_W = (I-P_W)^T = (P_W^{\perp})^T \end{cases}$

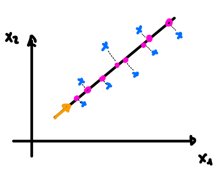
- "Find an orthogonal subspace which its range is described by the columns of W such that the orthogonal projection of the data X^T is minimal"

Recall: $\|B\|_F = B:B = B_{ij}B_{ij} = B_{ij}B_{ik}\delta_{kj} = (B^TB):I = \mathrm{tr}(B^TB)$

$$\arg\min_{W} \frac{1}{N}\|(I-P_W)X^T\|_F^2 \ \text{s.t.}\ W^TW = I = \arg\min_{W} \frac{1}{N}\left[X(I-P_W)^T(I-P_W)X^T\right]:I \ \text{s.t.}\ W^TW = I = \arg\min_{W} \frac{1}{N}\left(X(I-P_W)X^T\right):I \ \text{s.t.}\ W^TW = I$$

$$= \arg\min_{W} \frac{1}{N}\left[XX^T - XP_WX^T\right]:I \ \text{s.t.}\ W^TW = I = \arg\min_{W} -\frac{1}{N}XP_WX^T : I \ \text{s.t.}\ W^TW = I = \arg\max_{W} \frac{1}{N}\left(XP_WX^T\right):I \ \text{s.t.}\ W^TW = I = \arg\max_{W} \frac{1}{N}\left(XWW^TX^T\right):I \ \text{s.t.}\ W^TW = I$$

$$= \arg\max_{W} \frac{1}{N}\|W^TX^T\|_F \ \text{s.t.}\ W^TW = I = \arg\max_{W} \frac{1}{N}\|\overbrace{XW}^{Z}\|_F \ \text{s.t.}\ W^TW = I$$



- "Find an orthogonal subspace which its range is described by the columns of W such that its projection of the data X^T is maximal"

* Statistical interpretation

a) Eigenvalue decomposition of the covariance matrix

$$\arg\max_{W} \frac{1}{N}\|\overbrace{XW}^{Z}\|_F^2 = \arg\max_{W} \Big(W^T \underbrace{\overbrace{\tfrac{1}{N}X^T X}^{\Sigma}}_{\text{covariance matrix when data is centered}} W\Big) : I = \arg\max_{w_{(i)}} \sum_{i=1}^{\ell} w_{(i)}^T \Sigma w_{(i)}$$

s.t. $W^T W = I$ s.t. $W^T W = I$ s.t. $w_{(i)}^T w_{(i)} = 1 \ \forall i$, $w_{(i)}^T w_{(j)} = 0 \ \forall i \neq j$

abuse of notation

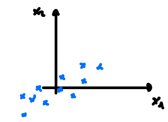
Thus : $\mathcal{L}(w_{(i)}, \lambda_i) = \sum_{i=1}^{\ell} w_{(i)}^T \Sigma w_{(i)} + \lambda_i (1 - w_{(i)}^T w_{(i)}) + \lambda_j w_{(i)}^T w_{(j)}$

KKT : $\begin{cases} [\Sigma + \lambda_i I] w_{(i)} + \lambda_j w_{(j)} = 0 & \forall i \\ w_{(i)}^T w_{(i)} = 1 & \forall i \\ w_{(i)}^T w_{(j)} = 0 & \forall i \neq j \end{cases} \Rightarrow \begin{cases} \Sigma w_{(i)} = \lambda_i w_{(i)} & \forall i \\ w_{(i)}^T w_{(i)} = 1 & \forall i \\ w_{(i)}^T w_{(j)} = 0 & \forall i \neq j \end{cases} \Rightarrow \begin{cases} \Sigma = W \Lambda W^T \\ W^T W = I \end{cases}$ Eigenvalue decomp.

Remark :

- "Find an orthogonal eigenvalue decomposition of the covariance matrix $\Sigma = \frac{1}{N} X^T X$ and select its first ℓ values"

* **Example** : $\Sigma = \frac{1}{N} X^T X = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$



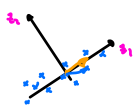
$[W, \Lambda] = \mathrm{eig}(\Sigma)$ is $W = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$; $\Lambda = \begin{bmatrix} 3/2 & \\ & 1/2 \end{bmatrix}$

b) Maximization of the covariance matrix trace

* Note that defining $Z = XW$, its corresponding covariance matrix

$$\tilde{\Sigma} = \frac{1}{N} Z^T Z = \frac{1}{N} W^T X^T X W = W^T \Sigma W = \Lambda = \begin{bmatrix} 3/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

" the covariance matrix becomes diagonal "



* $\arg\max_{W} \frac{1}{N}\|\overbrace{XW}^{Z}\|_F^2$ s.t. $W^T W = I$ $= \arg\max_{W, Z} \mathrm{tr}\left(\frac{1}{N} Z^T Z\right)$ s.t. $W^T W = I$, $XW = Z$

Remark :

- "Find an orthogonal transformation W s.t. the data in the new coordinates system $Z = XW$ provides a maximum trace of the new covariance matrix $\tilde{\Sigma} = \frac{1}{N} Z^T Z$"

* Computing PCA using SVD

- Top ℓ eigendecomposition of the covariance matrix : $\Sigma = W_\ell \Lambda_\ell W_\ell^T$ with $W^T W =$
- Top ℓ svd-decomposition of the data $X = U_\ell S_\ell V_\ell^T$ with $U_\ell^T U_\ell = I$ and $V_\ell^T V = I$

$$\begin{array}{c} \\ N \end{array} \Big[\overset{d}{X} \Big] = N \Big[\overset{\ell}{U_\ell} \Big] \ell \Big[\overset{\ell}{S_\ell} \Big] \ell \Big[\overset{d}{V_\ell^T} \Big]$$

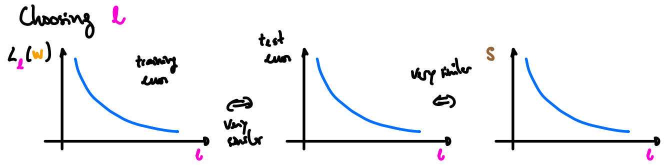
Reconst $\boxed{\begin{array}{l} Z = XW = U_\ell S_\ell \overbrace{\cancel{V_\ell^T V_\ell}}^{I} = U_\ell S_\ell \\ X^T = Z W^T = U_\ell S_\ell V_\ell^T \end{array}}$

- Note that : $\Sigma = \frac{1}{N} X^T X = \frac{1}{N} V_\ell S_\ell U_\ell^T U_\ell S_\ell V_\ell^T = V_\ell \frac{1}{N} (S_\ell)^2 V_\ell^T \Rightarrow \begin{cases} W_\ell = V_\ell \\ \lambda_\ell = S_\ell^2 / N \end{cases}$

* W_ℓ may be obtained with the V_ℓ, this is the Top ℓ svd-decomposition of X

* SVD algorithm more efficient, stable and precise

* Choosing ℓ



2) K-means (Unsupervised)

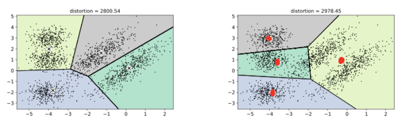
* Motivation: clustering
* Un-labeled data $\Rightarrow$ Finding clusters and clusters centers.
* Method: Choose R number of cluster centers $\Rightarrow$ minimize cluster variance
* Properties: $O(N)$ time + Euclidian distance
* Algorithm: (Lloyd)

a) Labeling (assignment): identify the clusters of each sample given means μ_k

$$z_k = \{ x_p : \| x_p - \mu_k \|^2 \le \| x_p - \mu_\ell \|^2 ; \forall \ell, 1 \le \ell \le K \}$$

Voronoi diagram

$$z_n^* = \arg\min_k d(x_n, \mu_k) = \| x_n - \mu_k \|^2 \; ; \quad \text{Example: } z_n = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \\ 3 \\ 3 \\ 1 \end{bmatrix} \; n \; z_n = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



(a) (b)

b) Means recalculation. For each k ; $\mu_k = \frac{1}{N_k} \sum_{n \in z_k} x_n$ given clusters

Note that

$$z^T z = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & & & \\ & 2 & & \\ & & 2 & \\ & & & 1 \end{bmatrix} = \sum N_k \delta_{ik}$$

Note that:

$$z^T X = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}}_{z} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ & \\ x_1^{(N)} & x_2^{(N)} \end{bmatrix} = \sum_{n \in z_k} x_n = N_k \mu_k$$

* Optimization problem

$$\min_{M, z \in \{0,1\}} J(M, z) \quad \text{s.t.} \quad G(z) = 0$$

Cost: $J(M, z) = \| X - z M^T \|_F^2 = \left\| \underbrace{\begin{bmatrix} X \end{bmatrix}}_{X \in \mathbb{R}^{N \times D}} - \underbrace{\begin{bmatrix} z \end{bmatrix}}_{z \in \{0,1\}^{N \times K}} \underbrace{\begin{bmatrix} M^T \end{bmatrix}}_{M \in \mathbb{R}^{D \times K}} \right\|^2$

Consts: $G(z) = z \cdot 1_K - 1_N$; $\begin{bmatrix} z \end{bmatrix} \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}_{1_K} = \underbrace{\begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}}_{1_N}$

$$J(M, z) = \| X - z M^T \|_F^2 = \sum_{n,d} \Big(X_{nd} - \sum_k z_{nk} M_{dk} \Big)^2$$

$$\| A \|_F = A : A = \sum_{ij} A_{ij} A_{ij} = \sum_{ij} (A_{ij})^2$$

Alternate direction method

a) Take M constant

separable problem

$$\min_z J(M, z) \quad \text{s.t.} \quad G(z) = 0 \;\Rightarrow\; \begin{cases} J(M, z) = \| X - z M^T \|_F^2 = \sum_{n,d} \big(X_{nd} - \sum_k z_{nk} M_{dk} \big)^2 \\ G(z) = z \cdot 1_K - 1_N = \sum_k z_{nk} 1_k - 1 \quad \forall n \end{cases}$$

$$\Rightarrow \forall n, d \; \begin{cases} \min_{z_k \in \{0,1\}} \big(X_{nd} - \sum_k z_{nk} M_{dk} \big)^2 \\ \text{s.t.} \; \sum_k z_{nk} 1_k - 1 = 0 \end{cases}$$

$$\forall n, d \; \begin{cases} \min_{z_k \in \{0,1\}} \Big(x_{nd} - \begin{bmatrix} M_d \end{bmatrix} \begin{bmatrix} z \end{bmatrix} \Big)^2 \\ \text{s.t.} \; \begin{bmatrix} z \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 1 \end{cases}$$

• Try all K and take the min Expensive if K is large!

$$\left. \begin{array}{l} K = 1 \; ; \; z_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ K = 2 \; ; \; z_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ \vdots \\ k = K \; ; \; z_K = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{array} \right\} \Rightarrow \min_k J(z_k, M)$$

$$\min_X \frac{1}{2} (B_{ij} - A_{ik} X_{kj})(B_{ij} - A_{i\ell} X_{\ell j})$$

$$\frac{\partial J}{\partial X_{mn}} = \frac{1}{2} \Big[(B_{ij} - A_{ik} X_{kj})(-A_{i\ell} \delta_{\ell m} \delta_{jn}) + (\overbrace{-A_{ik} \delta_{km}}^{A_{im}} \delta_{jn})(B_{ij} - A_{i\ell} X_{\ell j}) \Big]$$

$$= \frac{1}{2} \big[-B_{in} A_{im} + A_{ik} X_{kn} A_{im} - A_{im} B_{in} + A_{im} A_{i\ell} X_{\ell n} \big]$$

$$= \frac{1}{2} \big[-2 B_{in} A_{im} + 2 A_{im} A_{i\ell} X_{\ell n} \big] = 0 \Rightarrow \boxed{(A_{im} A_{i\ell}) X_{\ell n} = A_{im} B_{in}}$$

b) Take z constant

separable problem

$$\min_M J(M, z) \quad \text{s.t.} \quad \cancel{G(z) = 0} \quad \Big\} \quad \min_M J(M, z) = \| X - z M^T \|_F^2$$

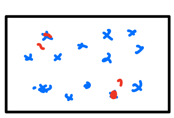
$$\hookrightarrow \underbrace{(z^T z)}_{\sum N_k \delta_{ik}} M^T = z^T X \Rightarrow M_{kd} = \frac{1}{N^{(k)}} z_{nk} X_{nd} = \mu_k$$

* Initial value: K-means++

* K-means problem is non-convex
* Much dependent on initial point
* Use an heuristic for initialization. At iteration t make the probability proportional to the distance $\Rightarrow$ The further to all μ, the most likely to be selected.

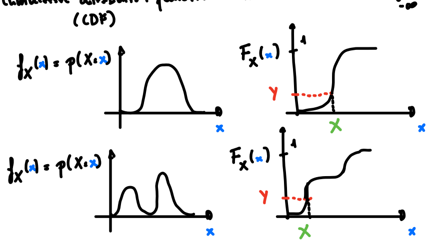
$$p(\mu_t = x_n) = \frac{D_{t-1}(x_n)}{\sum_{j=1}^{N} D_{t-1}(x_j)} \quad \text{with}$$

$$D_t(x) = \min_{k=1 \dots t-1} \| x - \mu_k \|_2^2 \; ;$$ Squared distance of x to the closest existing centroid. ;



* Sample a probability distribution

Cumulative distribution function (CDF): $F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt$



The sample X is such that $F_X(X) = Y$ where Y is a random sample from a uniform distribution $U(0,1)$

$$X = F_X^{-1}(Y)$$