Outline: PCA

Clustering

$$\min \sum_{i=1}^{n} \| x_i - \mu - A\theta_i \|_2^2$$

$\mu \in \mathbb{R}^d$

$A = [u_1, \dots, u_k]$ orthonormal

$\theta_i \in \mathbb{R}^k$

if $k=1 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\theta_j = u_1^T(x_j - \mu)$

$u_1 = \arg\max \frac{1}{n}\sum (u_1^T(x_i - \bar{x}))^2$

$u_1^T u_1 = 1$

$$\max_{u_1} \frac{1}{n}\sum_{i} (u_1^T(x_i - \bar{x}))^2 + \lambda \cdot (u_1^T u_1 - 1)$$

$$u_1^T \underbrace{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T}_{S} u_1 + \lambda(u_1^T u_1 - 1)$$

$$\frac{\partial}{\partial u_1}(\ ) = 2 \cdot S u_1 + \lambda \cdot 2 u_1 = 0 \Rightarrow S u_1 = -\lambda u_1$$

General Case ($k \geq 1$) Solution $\mu = \bar{x}$

$\theta_i = A^T(x_i - \bar{x})$ $A = [u_1, \dots, u_k]$ ← Eigenvectors of the sample covariance

Define $\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}$, $\tilde{x}_i = x_i - \bar{x}$

$S = \frac{1}{n}\tilde{X}^T\tilde{X}$

$$\min \sum_{i=1}^{n} \| x_i - \mu - A\theta_i \|_2^2$$

$A = U\Sigma V^T$ SVD

$A^+ = V\Sigma^{-1}U^T$ pseudo-inverse

$= (A^TA)^{-1}A^T$

$\Rightarrow AA^+ = U\Sigma V^T V \Sigma^{-1} U^T = UU^T$

1) Eliminate θ_i

$$\theta_i^+ = (A^TA)^{-1}A^T(x_i - \mu) = A^+(x_i - \mu)$$

$I - AA^+ = I - UU^T$

$(I - AA^+)^2 = (I - UU^T)(I - UU^T)$

$= I - UU^T$

2) Eliminate μ

$$\sum_i \|x_i - \mu - AA^+(x_i-\mu)\|_2^2 = \sum_i \|(I-AA^+)(x_i-\mu)\|_2^2$$
$$= \sum_i (x_i-\mu)^T \cdot (I-AA^+)^T (I-AA^+)(x_i-\mu)$$

$$\frac{\partial}{\partial \mu}(\ ) = \sum_i 2\cdot(I-AA^+)\cdot(x_i-\mu) = 2(I-AA^T)\left(\sum(x_i) - n\cdot\mu\right)$$

$$\mu^+ = \frac{1}{n}\sum x_i.$$

3) Optimize A: $\min \sum_i \|(I-\underset{\sim}{P}AA^+)(x_i-\mu)\|_2^2$ where $\tilde{x}_i = x_i - \bar{x}$.

$$= \min_{P:\text{ projection}} \|\tilde{X}^T - P\tilde{X}^T\|_F^2 \qquad \left(\|B\|_F = \sqrt{\sum_{ij} B_{ij}^2}\right) = \sqrt{\sum \|B_i\|_2^2}$$

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} \qquad \frac{1}{n}\tilde{X}^T\tilde{X} = \begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_n \end{bmatrix} \cdot \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} = \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{x}_i^T$$

$$P\tilde{X}^T = P\cdot\begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_n \end{bmatrix} = \begin{bmatrix} P\tilde{x}_1, & P\tilde{x}_2, & \cdots & P\tilde{x}_n \end{bmatrix}$$
$$\tilde{X}^T - P\tilde{X}^T = \begin{bmatrix} \tilde{x}_1 - P\tilde{x}_1, & \cdots, & \tilde{x}_n - P\tilde{x}_n \end{bmatrix}$$

Theorem: Eckart-Young Let Y have rank $r \geq k$, $Y = U\Sigma V^T$ is the SVD

$$\arg\min_{\substack{Z\in\mathbb{R}^{d\times n} \\ \text{rank}(Z)=k}} \|Y-Z\|_F = U\Sigma_k V^T$$ where Σ_k is Σ with $\sigma_{k+1}, \cdots, \sigma_r$ set to zero.

Choose $A = [u_1, \cdots, u_k]$ where u_i are eigenvectors of $\tilde{X}\tilde{X}^T$ (right-singular vectors of $\tilde{X}$)

Proof:

$$\|Y-Z\|_F = \|U\Sigma V^T - Z\|_F = \|U^T\cdot(U\Sigma V^T - Z)\|_F$$
$$= \|\Sigma V^T - U^T Z\|_F$$
$$= \|\Sigma - \underbrace{U^T Z V}_{N}\|_F$$

$$\|Y-Z\|_2^2 = \|\Sigma - N\|_2^2$$

$$= \sum_{i=1}^{r} (\sigma_i - N_{ii})^2 + \sum_{i>r} N_{ii}^2 + \sum_{i \neq j} N_{ij}^2$$

Choose $N_{ii} = \sigma_i$ for $i = 1, \dots, k$ and all others $N_{ij} = 0$

$Z = U \Sigma_k V^T \Rightarrow U^T Z V = \Sigma_k$

(Eigen) Face Recognition

- Face similarity
  - in the reduced space
  - insensitive to lighting expression, orientation
- Projecting new "faces"
  - everything is a face

new face

projected to eigenfaces

K-means

Clustering

Let $x_1, \dots, x_n \in \mathbb{R}^d$

Goal: Partition $\{x_1, \dots, x_n\}$ into disjoint subsets called "clusters" such that the points in the same cluster are more similar to each other

Cluster map $C: \{1, \dots, n\} \to \{1, \dots, k\}$

k is the number of clusters.

k-means Criterion

$$W(C) = \sum_{l=1}^{k} \sum_{i:\, C(i)=l} \| x_i - \bar{x}_l \|_2^2$$

where $\bar{x}_l = \frac{1}{n_l} \sum_{j:\, C(j)=l} x_j$, $n_l = \#\{ i : C(i) = l \}$

k-means Algorithm minimizes $W(C)$ approximately

Combinatorial optimization problem (hard)

iterative and sub-optimal algorithm

Initialize $m_1, \dots, m_k \in \mathbb{R}^d$

Repeat

For $i = 1, \dots, n$

$C(i) = \arg\min_l \| x_i - m_l \|_2$

End

For $l = 1, \dots, k$

$m_l = \frac{1}{|\{i : C(i) = l\}|} \sum_{i:\, C(i)=l} x_i$

End

Until clusters don't change

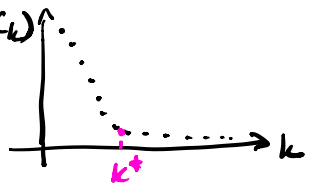
Also known as Lloyd-Max algorithm used in quantization.

Initialization: Common to randomly initialize. Also common to run the k-means algorithm several times and take the run with min $W(C)$

Model selection

How to set k ?

Simple



Idea: If $k < k^*$ $W(C_k) - W(C_{k+1})$ will be relatively large

If $k \geq k^*$ $W(C_k) - W(C_{k+1})$ will be relatively small

This suggests choosing the knee of the curve.

- kmeans ++

kernel k-means : $$W(C) = \sum_{l=1}^{k} \sum_{i : C(i)=l} \| \Phi(x_i) - \mu_l \|_2^2$$

where $$\mu_l = \frac{1}{n_l} \cdot \sum_{j : C(j)=l} \Phi(x_j)$$

$$\| \Phi(x_i) - \mu_l \|_2^2 = \| \Phi(x_i) \|_2^2 - 2 \langle \Phi(x_i), \mu_l \rangle + \| \mu_l \|_2^2$$

$$= \| \Phi(x_i) \|_2^2 - 2 \cdot \underbrace{\left\langle \Phi(x_i), \frac{1}{n_l} \cdot \sum_{j : C(j)=l} \Phi(x_j) \right\rangle}_{\sum_{j : C(j)=l} \langle \Phi(x_i), \Phi(x_j) \rangle} + \left\| \frac{1}{n_l} \cdot \sum_{j : C(j)=l} \Phi(x_j) \right\|_2^2$$

$K_{ij} = k(x_i, x_j)$