

3.1 (a) i) $\underset{C}{\operatorname{argmin}} \sum_{j=1}^k \sum_{x \in G_j} \|x - \mu_j\|^2$

$$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^k \sum_{x \in G_j} (\|x\|^2 + \frac{1}{|G_j|} \|\sum_{x \in G_j} x'\|^2 - \frac{2}{|G_j|} x^T (\sum_{x \in G_j} x'))$$

$$\Rightarrow \underset{C}{\operatorname{argmin}} \left[\sum_{j=1}^k \sum_{x \in G_j} \|x\|^2 + \sum_{j=1}^k \sum_{x \in G_j} \frac{1}{|G_j|} \|\sum_{x \in G_j} x'\|^2 - \sum_{j=1}^k \sum_{x \in G_j} \frac{2}{|G_j|} x^T (\sum_{x \in G_j} x') \right]$$

$$\Rightarrow \underset{C}{\operatorname{argmin}} \left[\sum_{j=1}^k \sum_{x \in G_j} \|x\|^2 - \sum_{j=1}^k \frac{1}{|G_j|} \|\sum_{x \in G_j} x'\|^2 \right]$$

$$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^k (\sum_{x \in G_j} \|x\|^2 - \frac{1}{|G_j|} \sum_{x, x' \in G_j} x^T x')$$

$$\operatorname{argmin}_C \sum_{j=1}^k \frac{1}{2|G_j|} \sum_{x, x' \in G_j} \|x - x'\|^2$$

$$\Rightarrow \operatorname{argmin}_C \sum_{j=1}^k \frac{1}{2|G_j|} \sum_{x, x' \in G_j} (\|x\|^2 + \|x'\|^2 - 2x^T x').$$

$$\Rightarrow \operatorname{argmin}_C \sum_{j=1}^k \frac{1}{2|G_j|} \left(\sum_{x \in G_j} \sum_{x' \in G_j} \|x\|^2 + \sum_{x \in G_j} \sum_{x' \in G_j} \|x'\|^2 - 2 \sum_{x, x' \in G_j} x^T x' \right)$$

$$\Rightarrow \operatorname{argmin}_C \sum_{j=1}^k \left(\sum_{x \in G_j} \|x\|^2 - \frac{1}{|G_j|} \sum_{x, x' \in G_j} x^T x' \right)$$

So the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster.

ii) Based on (i), $\underset{C}{\operatorname{argmin}} \sum_{j=1}^k \sum_{x \in G_j} \|x - \mu_j\|^2 \Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^k (\sum_{x \in G_j} \|x\|^2 - \frac{1}{|G_j|} \sum_{x, x' \in G_j} x^T x')$

For a given data set, $\sum_{j=1}^k \sum_{x \in G_j} \|x\|^2$ is a constant, so the problem equivalent to

$$\operatorname{argmax}_C \sum_{j=1}^k \sum_{x \in G_j} \frac{1}{|G_j|} x^T x \quad , \text{and define } |C| \text{ is the size of data } X.$$

$$\operatorname{argmax}_C \sum_{i=1}^k \sum_{j=1}^k |C_i||G_j| \|\mu_i - \mu_j\|^2$$

$$\Rightarrow \operatorname{argmax}_C \sum_{i=1}^k \sum_{j=1}^k |C_i||G_j| (\|\mu_i\|^2 + \|\mu_j\|^2 - 2\mu_i^T \mu_j)$$

$$\Rightarrow \operatorname{argmax}_C \left[2 \sum_{i=1}^k \sum_{j=1}^k |C_i||G_j| \|\mu_i\|^2 - 2 \sum_{i=1}^k \sum_{j=1}^k |C_i||G_j| \mu_i^T \mu_j \right]$$

$$\Rightarrow \operatorname{argmax}_C \left[2|C| \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x, x' \in G_i} x^T x' - 2 \sum_{i=1}^k \sum_{j=1}^k (\sum_{x \in G_i} x)^T (\sum_{x \in G_j} x') \right]$$

$$\Rightarrow \operatorname{argmax}_C 2|C| \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x, x' \in G_i} x^T x'$$

$$\Rightarrow \operatorname{argmax}_C \sum_{j=1}^k \sum_{x \in G_j} \frac{1}{|G_j|} x^T x'$$

And

$$\sum_{i=1}^k \sum_{j=1}^k (\sum_{x \in G_i} x)^T (\sum_{x \in G_j} x') \\ = (\sum_{i=1}^k \sum_{x \in G_i} x)^T (\sum_{j=1}^k \sum_{x \in G_j} x') \\ = \left\| \sum_{x \in G} x \right\|^2$$

which is a constant

So the k-means clustering problem is equivalent to maximizing the between-cluster sum of squares.

(b) i. Using Lloyd's algorithm, there are two steps in each iteration.

① In the relabel step, $c_{\text{new}}^{(i)} = \arg \min_j \|x^{(i)} - \mu_j\|^2$.

For any x in \mathcal{X} , exists $\|x - \mu_{c_{\text{new}}^{(i)}}\|^2 \leq \|x - \mu_{c_{\text{old}}^{(i)}}\|^2$.

Thus after relabeling, J does not increase.

② In the re-center step, by solve the problem

$$\min_{\mu_j} \sum_{i=1}^m \|x^{(i)} - \mu_j\|^2, \text{ the solution is } \mu_j = \frac{\sum_{i=1}^m \mathbb{1}\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^m \mathbb{1}\{c^{(i)}=j\}}$$

$$\text{there exists } \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}^{\text{new}}\|^2 \leq \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}^{\text{old}}\|^2$$

Because in every sub-step of Lloyd's algorithm's steps, the distortion J does not increase, it can prove that J does not increase in each step of Lloyd's algorithm.

ii. The algorithm always converges.

Proof: Based on (i), since every step the distortion decrease, Lloyd's algorithm will never move to a new configuration of cluster assignments unless the new configuration has a lower distortion. Since there are finite number of possible assignments, each with a corresponding unique ~~minimum~~ minimum distortion with regard to $\{\mu_j\}_{j=1}^k$. The Lloyd's algorithm will converge after a finite number of steps, when no changes to the cluster assignment will decrease the distortion, the ~~fix~~ $\{\mu_j\}_{j=1}^k$ also don't change.

3.2 (a) i. $U^T \text{Cov}(x) U = U^T E[(x - E[x])(x - E[x])^T] U$ (b). Define $X = [x^{(1)} - \hat{\mu}, \dots, x^{(m)} - \hat{\mu}]$, then

$$= E[U^T (x - E[x])(x - E[x])^T U]$$
$$= E[\| (x - E[x])^T U \|^2] \geq 0$$

$$\hat{C} = X X^T$$
$$\text{rank}(\hat{C}) = \text{rank}(X).$$

There exists $\sum_{i=1}^m (x^{(i)} - \hat{\mu}) = 0$, so the ~~column~~ vectors are not linear independent.

$$\text{rank}(X) \leq \min(d, m-1)$$

If \hat{C} is non-singular, $\text{rank}(\hat{C}) = d$.

$$\min(d, m-1) \geq d$$

$$m-1 \geq d$$

$$m \geq d+1$$

So the minimum value of m is $(d+1)$ such that \hat{C} is non-singular.

$$3.3 (a) \underset{u: u^T u=1}{\operatorname{argmin}} \|x^{(i)} - (x^{(i)T} u) u\|^2$$

$$= \underset{u: u^T u=1}{\operatorname{argmin}} (\|x^{(i)}\|^2 - 2(x^{(i)T} u) u^T x^{(i)} - (u^T x^{(i)})^2 u^T u)$$

$$= \underset{u: u^T u=1}{\operatorname{argmin}} (\|x^{(i)}\|^2 - 2(u^T x^{(i)})^2 + (u^T x^{(i)})^2)$$

$$= \underset{u: u^T u=1}{\operatorname{argmin}} (\|x^{(i)}\|^2 - (u^T x^{(i)})^2)$$

$$\Rightarrow \underset{u: u^T u=1}{\operatorname{argmax}} (x^{(i)T} u)^2$$

$$(C). \min_{u_1, \dots, u_n: u_i^T u_j = \delta_{ij}} \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \sum_{j=1}^n (x^{(i)T} u_j) u_j\|^2$$

$$= \min_{u_1, \dots, u_n: u_i^T u_j = \delta_{ij}} \frac{1}{m} \sum_{i=1}^m \left\| \sum_{j=1}^n (x^{(i)T} u_j) u_j \right\|^2$$

$$= \min_{u_1, \dots, u_n: u_i^T u_j = \delta_{ij}} \sum_{j=1}^n u_j^T \left(\frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \right) u_j$$

$$= \min_{u_1, \dots, u_n: u_i^T u_j = \delta_{ij}} \sum_{j=1}^n u_j^T \Sigma u_j$$

Generalized Lagrange function of this problem is

$$L(u_1, \dots, u_n) = \sum_{j=1}^n u_j^T \Sigma u_j - \sum_{j=1}^n \sum_{i \neq j} \beta_{ji} (u_j^T u_i - \delta_{ij})$$

Left multiply u_i (for any $i \neq j$) in both sides pt. ①

$$2u_i^T \Sigma u_j - 2\beta_{ji} u_i^T u_j - \beta_{ji} u_i^T u_i = 0$$

$$\frac{\partial L}{\partial u_j} = 2 \sum_{i \neq j} u_i^T \Sigma u_j - 2\beta_{ji} u_j - \sum_{i \neq j} \beta_{ji} u_i = 0 \quad ①$$

Bring $\beta_{ji} = 0$ in to ①,

$$\sum_{i \neq j} u_i^T \Sigma u_j = \beta_{jj} u_j$$

which means u_j is a eigenvector and β_{jj} is the corresponding eigenvalue.

$$\sum_{j=1}^n u_j^T \Sigma u_j = \sum_{j=1}^n u_j^T \sum_{i \neq j} u_i = \sum_{j=1}^n \beta_{jj}$$

$$\min_{u_1, \dots, u_n: u_i^T u_j = \delta_{ij}} \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \sum_{j=1}^n (x^{(i)T} u_j) u_j\|^2$$

$$\Rightarrow \min \sum_{j=1}^n \beta_{jj} = \sum_{j=1}^n \lambda_j$$

So, if want to minimize the residual of projection, must choose the $(n-k)$ eigenvectors with the smallest $(n-k)$ eigenvalues, which are $\lambda_{k+1}, \dots, \lambda_n$.

$$\min_{u_1, \dots, u_k: u_i^T u_j = \delta_{ij}} \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \sum_{j=1}^k (x^{(i)T} u_j) u_j\|^2 = \sum_{i=k+1}^n \lambda_i$$

(b) Based on (a).

$$\begin{aligned} u^* &= \underset{u: u^T u=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^m (x^{(i)T} u)^2 \\ &= \underset{u: u^T u=1}{\operatorname{argmax}} u^T \left(\frac{1}{m} \sum_{i=1}^m x^{(i)T} x^{(i)T} \right) u \\ &= \underset{u: u^T u=1}{\operatorname{argmax}} u^T \Sigma u \end{aligned}$$

Generalized Lagrange function of this problem is

$$L(u) = \frac{1}{m} \sum_{i=1}^m (x^{(i)T} u)^2 - \beta(u^T u - 1)$$

$$\frac{\partial L}{\partial u} = 2 \sum_{i=1}^m u_i - 2\beta u = 0 \Rightarrow \sum_{i=1}^m u_i = \beta u$$

So u^* is an eigenvector of Σ and β is the corresponding eigenvalue.

$$\frac{1}{m} \sum_{i=1}^m (x^{(i)T} u^*)^2 = u^* \Sigma u^* = \beta u^T u^* = \beta$$

Thus, minimizing the residual of projections is equivalent to finding the eigenvector with the largest eigenvalue.

3.4

Define $X = [x^{(1)}, \dots, x^{(m)}]^T, x \in \mathbb{R}^{md}, x^{(i)} \in \mathbb{R}^d$.

The covariance matrix is

$$\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} = \frac{1}{m} X^T X$$

Its eigenvectors are the principal axes and eigenvalues are principle component variances.

At the same time,

$$\operatorname{eig}(\Sigma) = \operatorname{eig}(X^T X) = \operatorname{eig}(XX^T)$$

The matrix XX^T has the same eigenvalues, and its eigenvectors are principle components.

In the linear kernel scenario, $K(x_i, x_j) = x_i^T x_j$

Define $X' = [\phi(x^{(1)}), \dots, \phi(x^{(m)})]^T$

$$(X' X'^T)_{ij} = \phi(x^{(i)})^T \phi(x^{(j)}) = K(x_i, x_j) = x_i^T x_j$$

$$\therefore X' X'^T = X X^T \Rightarrow \operatorname{eig}(X' X'^T) = \operatorname{eig}(\Sigma)$$

which means the linear kernel PCA has the same eigenvalues and eigenvectors as conventional PCA.

Hence the principle components will not change.

$$3.5 (a). Av_i = U\Sigma V^T v_i$$

$$= (U\Sigma)(V^T v_i)$$

$$= [\delta_1 u_1 \ \delta_2 u_2 \ \cdots \ \delta_r u_r] \begin{bmatrix} v_1^T v_i \\ v_2^T v_i \\ \vdots \\ v_r^T v_i \end{bmatrix}$$

$$= [\delta_1 u_1 \ \delta_2 u_2 \ \cdots \ \delta_r u_r] \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}_r$$

$$= \sum_{j=1}^r \delta_j u_j v_j^T v_i$$

$$= \delta_i u_i$$

$$A^T u_i = V \Sigma^T U^T u_i$$

$$= (V\Sigma^T)(U^T u_i)$$

$$= [\delta_1 v_1 \ \cdots \ \delta_r v_r] \begin{bmatrix} u_1^T u_i \\ u_2^T u_i \\ \vdots \\ u_r^T u_i \end{bmatrix}$$

$$= \sum_{j=1}^r \delta_j v_j u_j^T u_i$$

$$= \delta_i v_i$$

$$(b). \|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2}$$

$$= \max_x \frac{\|U\Sigma V^T x\|_2}{\|x\|_2} \quad \dots \dots \text{because } U \text{ is orthonormal}$$

$$= \max_x \frac{\|\Sigma V^T x\|_2}{\|x\|_2}$$

$$= \max_x \frac{\|\Sigma V^T x\|_2}{\|V^T x\|_2} \quad \dots \dots \text{because } V \text{ is orthonormal}$$

$$\text{Define } y = V^T x$$

$$\|A\|_2 = \max_y \frac{\|\Sigma y\|_2}{\|y\|_2}$$

$$= \max_y \sqrt{\delta_1^2 y_1^2 + \cdots + \delta_n^2 y_n^2}$$

$$= \delta_1$$