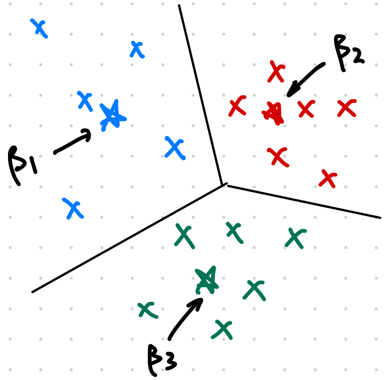
K-means as Low-rank Problem



* given list of n points $x_1, x_2, \dots, x_n \in \mathbb{R}^p$
* assign K centroids $\beta_1, \beta_2, \dots, \beta_K \in \mathbb{R}^p$
* Mininimize distance to closest centroid

$$(1) \qquad \min_{\beta_1,\dots,\beta_k \in \mathbb{R}^p} \sum_{i=1}^{n} \min_{k \in \{1,2,\dots,K\}} \| x_i - \beta_k \|^2$$

$$\min_{\beta_1,\dots,\beta_k \in \mathbb{R}^p} \ \min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \sum_{j \in G_k} \| x_j - \beta_k \|^2 : \underbrace{\overset{\text{disjoint union}}{\bigsqcup_{k=1}^{K}} G_k = \{1,2,\dots,n\}}_{\text{"partition into k parts"}} \right\}$$

Closed-form solution:

$$(2) \qquad \beta_k = \frac{1}{|G_k|} \sum_{j \in G_k} x_j$$

is the mean (or centroid) of k-th cluster.

Lloyd's algorithm (Expectation Maximization)

Alternate b/w centroids & cluster assignment

Underlying optimization problem

(3)
$$\min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \underbrace{\sum_{i\in G_k} \left\| x_i - \frac{1}{|G_k|}\sum_{j\in G_k} x_j \right\|^2}_{\text{Variance of } G_k} : \bigsqcup_{k=1}^{K} G_k = \{1,2,\dots,n\} \right\}$$

$$\boxed{\begin{aligned} \sum_{i\in G_k} \left\| x_i - \frac{1}{|G_k|}\sum_{j\in G_k} x_j \right\|^2 &= \frac{1}{|G_k|} \sum_{i,j\in G_k} \| x_i - x_j \|^2 \\ &= \frac{1}{2}\left\langle D, \frac{1}{|G_k|} \mathbb{1}_{G_k} \mathbb{1}_{G_k}^T \right\rangle \\ \text{where } D_{i,j} = \|x_i - x_j\|^2, &\quad (\mathbb{1}_{G_k})_i = \begin{cases} 1 & i\in G_k \\ 0 & i\notin G_k \end{cases} \end{aligned}}$$

Example

$$\left\| x_1 - \frac{1}{2}(x_1+x_2) \right\|^2 + \left\| x_2 - \frac{1}{2}(x_1+x_2) \right\|^2$$

$$= \left\| \frac{1}{2}x_1 - \frac{1}{2}x_2 \right\|^2 + \left\| \frac{1}{2}x_2 - \frac{1}{2}x_1 \right\|^2$$

$$= \frac{1}{2}\|x_1 - x_2\|^2 = \frac{1}{2}\left\langle D, \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\rangle \quad \square$$

(4)
$$\min_{G_1,\dots,G_K} \left\{ \frac{1}{2}\left\langle D, \sum_{k=1}^{K} \frac{1}{|G_k|} \mathbb{1}_{G_k} \mathbb{1}_{G_k}^T \right\rangle : \bigsqcup_{k=1}^{K} G_k = \{1,2,\dots,n\} \right\}$$

Low-rank change of variables.

(5)

$$\min_{U \in \mathbb{R}_+^{n\times k}} \left\{ \frac{1}{2}\langle D, UU^T\rangle : UU^T 1_n = 1_n,\ U^T U = I_k \right\}$$

Example $(N=3, K=2)$ $G_1 = \{1,2\}$, $G_2 = \{3\}$

$$\sum_{k=1}^{K} \frac{1}{|G_k|} 1_{G_k} 1_{G_k}^T = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}^T$$

$$U \geq 0,\ U^T U = I_2 \iff G_1 \cap G_2 = \emptyset$$

$$UU^T 1_3 = 1_3 \iff G_1 \cup G_2 = \{1,2,3\} \ \square$$

(6)

$$\max_{U \in \mathbb{R}_+^{n\times k}} \left\{ \langle XX^T, UU^T\rangle : UU^T 1_n = 1_n,\ U^T U = I_k \right\}$$

where $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n\times d}$

$$D_{ij} = \|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2\langle x_i, x_j\rangle$$

$$\frac{1}{2}\langle D, UU^T\rangle = \frac{1}{2}\langle 1 d^T + d 1^T - 2X^T X, UU^T\rangle$$

$$= 1^T d - \langle X^T X, UU^T\rangle$$

where $d = \operatorname{diag}(X^T X)$, since $UU^T 1 = 1$.

Theorem The K-means problem

$$(1) \quad \min_{\beta_1,\dots,\beta_K\in\mathbb{R}^p} \sum_{i=1}^{n} \min_{k\in\{1,2,\dots,K\}} \|x_i-\beta_k\|^2$$

is equivalent to the following

$$(6) \quad \max_{U\in\mathbb{R}_+^{n\times K}} \{\langle XX^T, UU^T\rangle : UU^T 1_n = 1_n,\ U^TU = I_K\}$$

Indeed, $(1) = \sum_i \|x_i\|^2 - (6)$ and

$$\beta_k^* = X^T U_k^* / 1^T U_k^*$$

SDP relaxation of K-means

$$\max_{U\in\mathbb{R}^{n\times K}} \{\langle XX^T, UU^T\rangle : UU^T 1_n = 1_n,\ U^TU = I_K,\ U\geq 0\}$$

Let $Z = UU^T \succeq 0$. Then, $U \geq 0 \Rightarrow Z \geq 0$.

$$U^TU = I_K \Rightarrow \mathrm{tr}(U^TU) = \mathrm{tr}(I_K) \Leftrightarrow \mathrm{tr}(UU^T) = K$$

$$\leq \max_{Z\succeq 0} \{\langle XX^T, Z\rangle : Z 1_n = 1_n,\ \mathrm{tr}(Z) = K,\ Z\geq 0\}$$

If relax is tight, recover U^* by $Z^* = U^* U^{*T}$.

Nonneg Matrix Factorization as relaxation of K-Means

$$\max_{U \in \mathbb{R}^{n\times k}} \left\{ \langle XX^T, UU^T \rangle : UU^T 1_n = 1_n,\ U^T U = I_k,\ U \geq 0 \right\}$$

$$\| XX^T - UU^T \|^2 = \| XX^T \|^2 + \underbrace{\| UU^T \|^2}_{= \|U^T U\|^2 = K} - 2 \langle XX^T, UU^T \rangle$$

$$= \text{const} - 2 \langle XX^T, UU^T \rangle$$

$$\min_{U \in \mathbb{R}^{n\times k}} \left\{ \| XX^T - UU^T \|^2 : UU^T 1_n = 1_n,\ U^T U = I_k,\ U \geq 0 \right\}$$

$$\geq \min_{U \in \mathbb{R}_+^{n\times k}} f(U) \triangleq \| XX^T - UU^T \|^2$$

$$\boxed{\begin{array}{l} \text{Projected Gradient descent:} \\ \qquad U \leftarrow \left(U - \alpha \nabla f(U) \right)_+ \end{array}}$$