K-means clustering

Data set \( \{ x_1, \ldots, x_n \} \), \( x_i \in \mathbb{R}^d \)

Partition the data set into \( K \) clusters

Goal: to find an assignment of data points to clusters and a set of vectors \( \{ \mu_k \} \), such that the sum of the squares of the distances of each data point to its closest vector \( \mu_k \), is a minimum.

\[ x_n \rightarrow r_{nk} \in \{0,1\}, \quad k = 1, \ldots, K \] (1-of-K coding scheme)

\[ r_{nk} = \begin{cases} 1, & x_n \text{ in cluster } k, \\ 0, & \text{otherwise}. \end{cases} \]

Minimize \( J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \mu_k \|^2 \) \( \{ \mu_k \} \) \( \forall k \)

Two-stage optimization

E step \( r_{nk} = \begin{cases} 1, & \text{if } k = \text{arg-min } \| x_n - \mu_k \|^2 \\ 0, & \text{otherwise} \end{cases} \)

M step \( \mu_k = \frac{\partial J}{\partial \mu_k} = 0 \Rightarrow -2 \sum_{n=1}^{N} r_{nk} (x_n - \mu_k) = 0 \)

\( \mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \) (cluster mean)

Examples of K-means clustering

- Image segmentation
- Vector quantization

Sequential update

\[ \mu_k^{\text{new}} = \mu_k^{\text{old}} + \lambda_n \frac{r_{nk}}{\mu_k^{\text{old}}} \nabla J(\mu_k) \]

\[ = \mu_k^{\text{old}} + \lambda_n (-2 r_{nk} (x_n - \mu_k^{\text{old}})) \]

\[ = \mu_k^{\text{old}} + \eta r_{nk} (x_n - \mu_k^{\text{old}}) \]

Robustness to Outliers? \( L^2 \)-norm is not robust to outliers

K-medoids

\[ J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} V(x_n, \mu_k) \]

\( V(x_n, \mu_k) \) is a general dissimilarity measure
Mixture of Gaussians

Generative Models

\[ p(Z_k = 1) = \pi_k \]

\[ 0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1 \]

\[ Z \] uses a 1-of-\( K \) coding scheme

\[ p(Z) = \sum_{k=1}^{K} \pi_k Z_k \]

\[ p(Z_k = 1) = \pi_k \]

Since \( p(x \mid Z_k = 1) \) is assumed to be a Gaussian,

\[ p(x \mid Z_k = 1) = \mathcal{N}(x \mid \mu_k, \Sigma_k) \]

Considering \( Z_k \) as a selector:

\[ p(x \mid Z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) Z_k \]

the marginal distribution of \( x \)

\[ p(x) = \sum_{Z} p(x, Z) = \sum_{Z} p(x \mid Z_k) p(Z_k) \]

\[ = \sum_{Z} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) Z_k = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \]

sum over all possible states of \( Z \in \{(0, \ldots, 0), \ldots, (0, \ldots, 0)\} \)

Maximum Likelihood

Given \( N \) observations \( \{x_1, \ldots, x_N\} \), we want to find \( \mu_k, \Sigma_k, \pi_k \) that maximize the likelihood function

\[ L = \prod_{n=1}^{N} p(x_n \mid \{\pi_k\}, \{\mu_k\}, \{\Sigma_k\}) \]

\[ = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \]

\[ = \sum_{k=1}^{K} \prod_{n=1}^{N} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \]

\[ \uparrow \]

maximize it w.r.t. \( \mu_k, \Sigma_k, \pi_k \)
Singualrities in the Likelihood Function of Mixtures of Gaussians

\[ N(x_n | \mu_j, \sigma_j^2 | I) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{1}{2} (x_n - \mu_j)^2 \right\} \]

one Gaussian collapses onto a specific data point
the other takes care of the remaining data points

For a single Gaussian model, such a situation would not happen.

Suppose the Gaussian collapses onto a specific data point
these data points have zero likelihoods and thus the overall likelihood goes to zero.

Maximum Likelihood \[ \frac{\partial L}{\partial \mu_k} = 0, \quad \frac{\partial L}{\partial \sigma_k^2} = 0, \quad \frac{\partial L}{\partial \Sigma_k} = 0 \]

Recall \[ N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]

\[ \begin{align*}
\frac{\partial L}{\partial \mu_k} &= \sum_{n=1}^{N} \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} (x_n - \mu_k) \\
\frac{\partial L}{\partial \Sigma_k} &= \sum_{n=1}^{N} \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} (x_n - \mu_k)(x_n - \mu_k)^T - \frac{1}{2} \Sigma_k
\end{align*} \]

Plugging these derivatives into \[ \frac{\partial L}{\partial \mu_k}, \frac{\partial L}{\partial \Sigma_k} \]
we have

\[ O = \sum_{n=1}^{N} \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} 2^{-1} (x_n - \mu_k) \]

Multiplying both sides by \[ \Sigma_k \]
we obtain

\[ \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \quad \text{and} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \]

\[ \gamma(z_{nk}) = \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} \]
Next, we set \( \frac{\partial L}{\partial \Sigma_k} \) to zero to get \( \Sigma_k \)

\[
0 = \frac{\partial L}{\partial \Sigma_k} = \frac{3}{2 \Sigma_k} \sum_{n=1}^{N} \beta_n \left( \sum_{k=1}^{K} n_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right) \\
= \sum_{n=1}^{N} \frac{n_k}{\Sigma_k} \frac{2}{3 \Sigma_k} \mathcal{N}(x_n | \mu_k, \Sigma_k) \\
= \sum_{n=1}^{N} \frac{n_k}{\sum_{j=1}^{K} n_j} \mathcal{N}(x_n | \mu_j, \Sigma_j)
\]

\[
\frac{\partial \mathcal{N}}{\partial \Sigma} = \frac{1}{(2\pi)^{k/2}} \left( \frac{\partial \mathcal{N}}{\partial \Sigma} \right)^{-1} \exp \left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \\
+ \frac{1}{(2\pi)^{k/2}} \left| \Sigma \right|^{-1/2} \frac{\partial \mathcal{N}}{\partial \Sigma} \exp \left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \\
= -\frac{1}{2} \Sigma^{-1} \mathcal{N} - \Sigma^{-2} (x-\mu)^T (x-\mu)^T \mathcal{N}
\]

\[
\frac{\partial}{\partial \Sigma} \exp \left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\} = \exp \left\{ \right\} (-\Sigma^{-2})(-\frac{1}{2} (x-\mu)^T (x-\mu)^T)
\]

\[
\frac{\partial}{\partial \Sigma} \text{Tr}(x^T A x) = \frac{\partial}{\partial \Sigma} \text{Tr}(x^T A x) = \left( \frac{\partial}{\partial \Sigma} x^T A \right) x^T = x x^T
\]

\[
0 = \frac{3}{2 \Sigma_k} \sum_{n=1}^{N} \beta_n \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \\
= \sum_{n=1}^{N} \frac{n_k}{\sum_{j=1}^{K} n_j} \mathcal{N}(x_n | \mu_j, \Sigma_j) \left( \frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \Sigma_k^{-2} (x_n \mu_k (x_n - \mu_k)^T) \right)
\]

by multiplying both sides by \( 2 \Sigma_k \)

\[
\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} y(x_n - \mu_k) (x_n - \mu_k)^T
\]

\[
\hat{\gamma} = \sum_{n=1}^{N} \sum_{k=1}^{K} n_k \mathcal{N}(x_n | \mu_k, \Sigma_k) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right)
\]

\[
\frac{\partial \hat{\gamma}}{\partial n_k} = 0
\]

given by the constraint

\[
0 = \sum_{n=1}^{N} n_k \mathcal{N}(x_n | \mu_k, \Sigma_k) + \lambda
\]

\[
\downarrow \text{ multiplying by } n_k, \text{ summing over } k
\]

\[
0 = \sum_{n=1}^{N} 1 + \sum_{k=1}^{K} n_k \lambda
\]

\[
\downarrow \text{ summing over } n_k
\]

\[
\lambda = -N
\]
\[ \pi_{nk} \cap N = \sum_{n=1}^{N} \frac{\pi_{nk} \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_{nj} \mathcal{N}(x_n | \mu_j, \Sigma_j)} \]

\[ \pi_{nk} = \frac{N_{nk}}{N} \]

\[ N_{nk} = \sum_{n=1}^{N} \gamma(z_{nk}) \]

**SUMMARY**

**E step:**

\[ \gamma(z_{nk}) = \frac{\pi_{kn} \mathcal{N}(x_n | \mu_n, \Sigma_n)}{\sum_{j=1}^{K} \pi_{nj} \mathcal{N}(x_n | \mu_j, \Sigma_j)} \]

**M step:**

\[ \mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \]

\[ \Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k^{\text{new}}) (x_n - \mu_k^{\text{new}})^T \]

\[ \pi_k^{\text{new}} = \frac{N_k}{N} \]

where

\[ N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \]