Unsupervised Learning

Task a) Dimensionality Reduction

b) Clustering

c) Density Estimation $\longleftrightarrow$ Generative Model

---

① PCA Principal Component Analysis

Big Idea: Reduced Representation via linear projection (transformation)

Dataset $\mathcal{D} = \{(x_i)\}_{i=1}^N \quad x_i \in \mathbb{R}^d$

1. Maximize Variance Formulation

$z_i = u^T x_i \longrightarrow$ **Interpretation**: the coordinate after projecting to u.

$\downarrow$

variance (estimator) of $\{z_1, \dots, z_n\}$ $\quad (\mathrm{Var}[z] = \mathbb{E}[(z - \mathbb{E}[z])^2])$

$$= \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^2, \quad \bar{z} = \frac{1}{N} \sum_{i=1}^N u^T x_i = u^T \bar{x}$$

$$= \frac{1}{N} \sum_{i=1}^N u^T (x_i - \bar{x})(x_i - \bar{x})^T u$$

Therefore, one realization to calculate variance is:

$$\mathcal{D} = \{x_i\}_{i=1}^N \xrightarrow{x_i - \bar{x}} \tilde{\mathcal{D}} = \{\tilde{x}_i\}_{i=1}^N \longrightarrow S = \frac{1}{N} \sum_{i=1}^N \tilde{x}_i \tilde{x}_i^T$$

$\Rightarrow$ variance of $\{z_1, z_2, \dots, z_n\}$ $\quad z_i = u^T x_i$

$$= u^T S u$$

1-st principal component

Maximize variance $\Leftrightarrow \begin{cases} \max\limits_{u} & u^T S u \\ \text{s.t} & \|u\|_2^2 = 1 \end{cases} \Leftrightarrow$ u: eigenvector corresponds to the largest eigenvalue

to calculate the 2-nd principal component, the formulation is:

$$\begin{cases} \max\limits_{u} & u^T S u \\ \text{s.t} & \|u\|_2^2 = 1 \;\&\; u^T u_1 = 0 \end{cases} \longrightarrow$$ u: eigenvector corresponds to the second largest eigenvalue

Thus, given m (# of Principal Components we need), we can achieve

$\longrightarrow$ ① $Z_m = X U_m$ $\quad \begin{cases} U_m = (u_1, \dots, u_m) \in \mathbb{R}^{d \times m} \\ X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix} \in \mathbb{R}^{N \times d} \end{cases}$

m principal component scores $\rightarrow Z_m$ $\longrightarrow$ 本质: Basis transformation

Interpretation (the projection coordinates for $x_1, \dots, x_N$ towards direction $u_1, \dots, u_m$)

② $U_m \longrightarrow$ projection direction

Interpretation: for the given principal component, the contribution for each dimension (raw).

2. Minimize Re-construct Error Formulation

given a set of orthonormal basis $\{u_1, \dots, u_d\}$

then $x_i = \sum_{j=1}^{d} (u_j^T x_i) \cdot u_j \qquad \hat{x}_i = \sum_{j=1}^{m} \beta_{ij} u_j$

Our aim is: choose $\{\beta_{ij}\}$ & $\{u_1, \dots, u_d\}$

So that we can minimize $\frac{1}{N} \sum_{i=1}^{N} \| x_i - \tilde{x}_i \|_2^2$

Calculation: $\frac{1}{N} \sum_{i=1}^{N} \| x_i - \tilde{x}_i \|_2^2$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{j=1}^{m} (\alpha_{ij} - \beta_{ij})^2 + \sum_{j=m+1}^{d} \alpha_{ij}^2 \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} (\alpha_{ij} - \beta_{ij})^2 + \frac{1}{N} \sum_{i=1}^{N} \sum_{j=m+1}^{d} \alpha_{ij}^2$$

Here, $\alpha_{ij} = u_j^T x_i$

Analysis: ① $\beta_{ij} = u_j^T x_i \qquad i = 1, 2, \dots, N \quad j = 1, 2, \dots, m$

when u_j is determined

② $\{\hat{u}_1, \dots, \hat{u}_d\} = \underset{u_1, \dots, u_d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=m+1}^{d} u_j^T x_i x_i^T u_j$

$$= \underset{u_1, \dots, u_d}{\operatorname{argmin}} \sum_{j=m+1}^{d} u_j^T S u_j$$

$\{u_1, \dots, u_d\}$ is orthonormal basis

$$\Updownarrow$$

$$\begin{cases} u_i^T u_j = 0 & i \neq j \\ u_i^T u_i = 1 & i = 1, 2, \dots, d \end{cases}$$

$\Rightarrow$ one choice of $\{\hat{u}_1, \dots, \hat{u}_d\}$ is:

the set of eigenvector of S, ordered in decreasing eigenvalue

Re-construction : $x \longmapsto \sum_{j=1}^{m} (u_j^T x) \cdot u_j = \sum_{j=1}^{m} z_j u_j = U_m \cdot z$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix} \longmapsto X (u_1 \cdots u_m) \begin{pmatrix} u_1^T \\ \vdots \\ u_m^T \end{pmatrix}$$

$$= X U_m \cdot U_m^T$$

$$Z_m = \begin{pmatrix} z_1^T \\ \vdots \\ z_N^T \end{pmatrix}$$

(coordinates) principal component scores

② Kernel PCA ☆ ⟶ not so convenient

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}$$

For PCA, our interest is : $S = \frac{1}{N} \sum_{i=1}^{N} x_i \cdot x_i^T = \frac{1}{N} X^T X$

In Kernel PCA Framework, $x \in \mathbb{R}^d \longmapsto \phi(x) \in \mathbb{R}^D$

Then. Algo becomes :

a) $\Phi_{ij} := \phi_j(x_i) \in \mathbb{R}^{N \times D}$

b) Centralized $\Phi = \Phi - \frac{1}{N} \mathbb{1}_N \mathbb{1}_N^T \Phi$

c) $S_\phi = \frac{1}{N} \Phi^T \Phi$

d) Eigenvalue / vector Decomposition

$$\begin{cases} U_m = (u_1, \ldots, u_m) & u_i \in \mathbb{R}^D \\ \lambda_m = (\lambda_1, \ldots, \lambda_m) \end{cases}$$

e) $Z_m = \Phi U_m$

③ PCA ⟶ Whitening

Whitening

Whitening refers to: $X \longrightarrow X' = XU\Lambda^{-\frac{1}{2}}$

- mean 0
- covariance I

Check: Covariance of X'

$$= \frac{1}{N} X'^T X'$$

$$= \frac{1}{N} \Lambda^{-\frac{1}{2}} U^T X^T X U \Lambda^{-\frac{1}{2}}$$

$$= \Lambda^{-\frac{1}{2}} U^T S U \Lambda^{-\frac{1}{2}}$$

$$= \Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}}$$

$$= I$$

Note:

$S \cdot U = U \cdot \Lambda$

$\Rightarrow U^T S U = \Lambda$

④ Auto-Encoder ⟶ Encoder-Decoder Architecture

Motivation

$X \xrightarrow{\text{Encoder}} Z_m = XU_m \xrightarrow{\text{Decoder}} X_m = Z_m \cdot U_m^T$

$Z_m \in \mathbb{R}^{N \times m}$ → dimensionality reduction

→ lossy encoder-decoder algorithm

Generalization

$x \in \mathbb{R}^d \xrightarrow[T_{enc}^{\theta}: \mathbb{R}^d \to \mathbb{R}^m]{\text{encoder}} z_m \in \mathbb{R}^m \xrightarrow[T_{dec}^{\phi}: \mathbb{R}^m \to \mathbb{R}^d]{\text{decoder}} x_m \in \mathbb{R}^d$

q-dim q-dim

optimization $\min_{\theta,\phi} \frac{1}{N} \sum_{i=1}^{N} \| x_i - T_{dec}^{\phi}(T_{enc}^{\theta}(x_i)) \|_2^2$

$$\begin{cases} T_{enc}^{\theta}(x) = A \cdot \sigma(Wx+b) & A \in \mathbb{R}^{m \times q} \quad W \in \mathbb{R}^{q \times d} \\ T_{dec}^{\phi}(x) = B\,\sigma(Vx+c) & B \in \mathbb{R}^{d \times q} \quad V \in \mathbb{R}^{q \times m} \end{cases}$$

---

⑤ K-means Clustering

$\mathcal{D} = \{x_i\}_{i=1}^{N}$ $\quad x_i \in \mathbb{R}^d$

Aim: Partition into K groups (K clusters) → { similar within groups; different between groups }

(assignment)

→ { 1. Partition $\{a_{ij}\}_{i=1,j=1}^{N,K}$ $\quad a_{ij} = \begin{cases} 1, & x_i \to \text{cluster } j \\ 0, & \text{o/w} \end{cases}$

2. Centroid $\{z_1, \dots, z_K\}$

1. Formulation :

$$\min_{a_{ij}, z_j} \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{K} a_{ij} \| x_i - z_j \|_2^2$$

{ $A \in \mathbb{R}^{N \times K}$ → assignment matrix

$Z = (z_1, \dots, z_K) \in \mathbb{R}^{d \times K}$ → centroid matrix

→ NP-Hard → intractable

2. Trade-off → greedy algorithm

iterative algorithm

→ a) suppose we know Z

then update the assignment matrix A

by $a_{ij} = \begin{cases} 1, & j = \underset{j'}{\operatorname{argmin}} \|X_i - Z_{j'}\|_2^2 \\ 0, & \text{o/w} \end{cases}$

→ re-assignment

→ b) suppose we know A

then we update the centroid matrix Z

by $Z_k = \frac{\sum_{i=1}^{N} a_{ik} X_i}{\sum_{i=1}^{N} a_{ik}}$

→ re-center

3. Remarks:

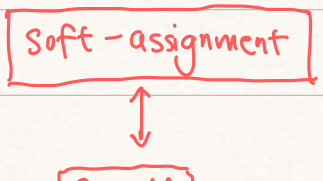
① This algorithm guarantees to converge (in finite steps)

② This algorithm cannot guarantee to converge to global optima

③ $J_{k+1} \leq J_k$

④ K-means → hard-assignment $A \leftrightarrow a_{ij}$

$a_{ij} = \begin{cases} 1 & x_i \to \text{class } j \\ 0 & \text{o/w} \end{cases}$



GMM

---

6. GMM → Gaussian Mixture Model

↓

Density Estimation Model, more appropriate for the cases that

☆ Data Lying Roughly In-between 2 classes (difficult cases)

① Model: (Gaussian Mixture)

$$X \sim GMM(\pi_k, z_k, \Sigma_k)$$

$$\rightarrow p_X(x) = \sum_{k=1}^{K} \pi_k p_G(x; z_k, \Sigma_k)$$

$$= \sum_{k=1}^{K} \pi_k \cdot (2\pi)^{-\frac{d}{2}} |\det \Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - z_k)^T \Sigma_k^{-1} (x - z_k)\right\}$$

☆☆

② How to make use of this model?

Assume we already have the GMM density $p_X(x)$

then our interest is: $p(r_k = 1 \mid x)$ for $k = 1, 2, \ldots, K$

Calculation $p(r_k = 1 \mid x) = \dfrac{p(x, r_k = 1)}{p(x)}$

$$= \frac{\pi_k p_G(x; z_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l p_G(x; z_l, \Sigma_l)}$$

↓

when we have a data point $x \in \mathbb{R}^d$

we can calculate $p(r_1=1|x) \dots\dots p(r_K=1|x)$

① Soft-assignment

② Consider the component of x { 20% 娱乐, 30% 音乐, 50% 学习

⇒ the 'responsibility' of component 1 for explaining the outcome of x

may have some interesting interpretation

③ How to learn the parameters? ← MLE

$$\max_{\theta} \log \prod_{i=1}^{N} p_{GMM}(x_i; \pi_k, z_k, \Sigma_k) \qquad \theta = \{ [\pi_k]_{k=1}^{K}, [z_k]_{k=1}^{K}, [\Sigma_k]_{k=1}^{K} \}$$

$$= \max_{\theta} \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{K} \pi_j p_G(x_i; z_j, \Sigma_j) \right\} := \ell(\theta)$$

a) Brute-Force (Optimization)

$$① \rightarrow \frac{\partial \ell}{\partial z_\ell} = \sum_{i=1}^{N} \frac{\pi_\ell \frac{\partial p_G}{\partial z_\ell}(x_i; z_\ell, \Sigma_\ell)}{\sum_{j=1}^{K} \pi_j p_G(x_i; z_j, \Sigma_j)}$$

$$= \sum_{i=1}^{N} \frac{\pi_\ell \cdot (-\Sigma_\ell^{-1}(x_i - z_\ell)) \cdot p_G(x_i; z_\ell, \Sigma_\ell)}{\sum_{j=1}^{K} \pi_j p_G(x_i; z_j, \Sigma_j)}$$

$$= (-\Sigma_\ell^{-1}) \sum_{i=1}^{N} a_{i\ell} (x_i - z_\ell)$$

$$② \rightarrow \frac{\partial \ell}{\partial \Sigma_\ell} = \sum_{i=1}^{N} \frac{\pi_\ell \frac{\partial p_G}{\partial \Sigma_\ell}(x_i; z_\ell, \Sigma_\ell)}{\sum_{j=1}^{K} \pi_j p_G(x_i; z_j, \Sigma_j)}$$

$$p_G(x_i; \mu_\ell, \Sigma_\ell) = (2\pi)^{-\frac{d}{2}} |\det(\Sigma_\ell)|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (x_i - \mu_\ell)^T \Sigma_\ell^{-1} (x_i - \mu_\ell) \right\}$$

$$= (2\pi)^{-\frac{d}{2}} \det(\Sigma_\ell)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x_i-\mu_\ell)^T \Sigma_\ell^{-1}(x_i-\mu_\ell)\right\}$$

$$\frac{\partial P_G}{\partial \Sigma}(x_i;\mu_\ell,\Sigma_\ell) = (2\pi)^{-\frac{d}{2}}\left[-\tfrac{1}{2}\det(\Sigma_\ell)^{-\frac{3}{2}}\frac{\partial \det(\Sigma_\ell)}{\partial \Sigma_\ell}\exp\left\{-\tfrac{1}{2}(x_i-\mu_\ell)^T\Sigma_\ell^{-1}(x_i-\mu_\ell)\right\}\right.$$
$$\left. + \det(\Sigma_\ell)^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(x_i-\mu_\ell)^T\Sigma_\ell^{-1}(x_i-\mu_\ell)\right\}\frac{\partial(-\frac{1}{2}(x_i-\mu_\ell)^T\Sigma_\ell^{-1}(x_i-\mu_\ell))}{\partial \Sigma_\ell^{-1}}\cdot\frac{\partial \Sigma_\ell^{-1}}{\partial \Sigma_\ell}\right]$$

$$\boxed{\frac{d\det(A)}{dA} = \det(A)\cdot(A^{-1})^T} = (2\pi)^{-\frac{d}{2}}\left[-\tfrac{1}{2}\det(\Sigma_\ell)^{-\frac{3}{2}}\det(\Sigma_\ell)(\Sigma_\ell^{-1})\exp(*)\right.$$
$$\left. + \det(\Sigma_\ell)^{-\frac{1}{2}}\exp(*)\left(-\tfrac{1}{2}(x_i-\mu_\ell)(x_i-\mu_\ell)^T\right)(-\Sigma_\ell^{-2})\right]$$

$$= (2\pi)^{-\frac{d}{2}}\left(-\tfrac{1}{2}\right)\det(\Sigma_\ell)^{-\frac{1}{2}}\exp(*)\left[\Sigma_\ell^{-1} - (x_i-\mu_\ell)(x_i-\mu_\ell)^T\Sigma_\ell^{-2}\right]$$

$$= -\tfrac{1}{2}P_G(x_i;z_\ell,\Sigma_\ell)\left[\Sigma_\ell^{-1} - (x_i-z_\ell)(x_i-z_\ell)^T\Sigma_\ell^{-2}\right]$$

$$\frac{\partial \ell}{\partial \Sigma_\ell} = \sum_{i=1}^{N}\frac{\pi_\ell\frac{\partial P_G}{\partial \Sigma_\ell}(x_i;z_\ell,\Sigma_\ell)}{\sum_{j=1}^{K}\pi_j P_G(x_i;z_j,\Sigma_j)}$$

$$= -\tfrac{1}{2}\sum_{i=1}^{N} a_{i\ell}\left[\Sigma_\ell^{-1} - (x_i-z_\ell)(x_i-z_\ell)^T\Sigma_\ell^{-2}\right]$$

---

③ $\longrightarrow$ $\mathcal{L}(\theta;\lambda) = \ell(\theta) + \lambda\left(1-\sum_{\ell=1}^{K}\pi_\ell\right)$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial \pi_\ell} = \sum_{i=1}^{N}\frac{P_G(x_i;z_\ell,\Sigma_\ell)}{\sum_{j=1}^{K}\pi_j P_G(x_i;z_j,\Sigma_j)} - \lambda$$

$$= \sum_{i=1}^{N} a_{i\ell}\cdot\frac{1}{\pi_\ell} - \lambda$$

To conclude :

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial z_\ell} = 0 \Rightarrow (-\Sigma_\ell^{-1})\sum_{i=1}^{N} a_{i\ell}(x_i - z_\ell) = 0 \\ \frac{\partial \mathcal{L}}{\partial \Sigma_\ell} = 0 \Rightarrow \sum_{i=1}^{N} a_{i\ell}\left[\Sigma_\ell^{-1} - (x_i-z_\ell)(x_i-z_\ell)^T\Sigma_\ell^{-2}\right] = 0 \\ \frac{\partial \mathcal{L}}{\partial \pi_\ell} = 0 \Rightarrow \sum_{i=1}^{N} a_{i\ell}\frac{1}{\pi_\ell} - \lambda = 0 \end{cases}$$

$$
\begin{cases}
\mu_l = \dfrac{\sum_{i=1}^{N} a_{il}\, x_i}{\sum_{i=1}^{N} a_{il}} \\
\Sigma_l = \dfrac{1}{\sum_{i=1}^{N} a_{il}} \sum_{i=1}^{N} (x_i - \mu_l)(x_i - \mu_l)^T \\
\pi_l = \dfrac{\sum_{i=1}^{N} a_{il}}{N}
\end{cases}
$$

b) EM Framework

Recap: objective is $\rightarrow \max_{\theta} \ell_N(X \mid \theta)$

$$= \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{K} \pi_j P_G(x_i \mid \mu_j, \Sigma_j) \right\}$$

$$= \sum_{i=1}^{N} \log \mathbb{E}_{A_i \sim \pi} [L(x_i, A_i \mid \theta)]$$

$$= \log \mathbb{E}_{A \sim \pi} [L(X, A \mid \theta)]$$

EM Framework

$$\Rightarrow \quad \theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{A \sim p(\cdot \mid X, \theta^{(t)})} [\log L(X, A \mid \theta)]$$

$$\rightarrow \log L(X, A \mid \theta) = \sum_{i=1}^{N} \log L(x_i, A_i \mid \theta)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} a_{ij} \log P_G(x_i \mid \mu_j, \Sigma_j)$$

$$\rightarrow \mathbb{E}_{A \sim p(\cdot \mid X, \theta^{(t)})} [\log L(X, A \mid \theta)] \xrightarrow{\arg\max} \theta^{(t+1)} \longrightarrow \text{M-step}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{E}_{A_i \sim P(\cdot \mid x_i, \theta^{(t)})} [a_{ij}] \; P_G(x_i \mid \mu_j, \Sigma_j)$$

$$\rightarrow \mathbb{E}_{A_i \sim P(\cdot \mid x_i, \theta^{(t)})} [a_{ij}]$$

$$= P(x_i \in \text{cluster } j \mid \theta^{(t)}, x_i)$$

$$= \frac{\pi_j^{(t)} P_G(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} P_G(x_i \mid \mu_k^{(t)}, \Sigma_k^{(t)})}$$

→ E-step

11.27

GMM Summary

Big idea!

Rmk: Actually it is a common way since this can be just viewed as 'parameter estimation'

↳ [e.g.] estimate $N(\mu, \sigma^2)$ from samples

1. Maximize log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log \{ f(x_i \mid \theta) \}$$

$$\theta := \{ (\pi_k, \mu_k, \Sigma_k) \}_{k=1}^{K}$$

$$= \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k P_g(x_i; \mu_k, \Sigma_k) \right\}$$

Issue: $\ell(\theta)$ is highly non-convex w.r.t (π_k, μ_k, Σ_k)

↓

Conventional GD may break down (but still can try)

2. 3 Solutions
   - a) GD
   - b) Necessary Condition $\nabla_\theta \ell(\hat{\theta}) = 0$
   - c) EM Framework

* "Augmented Lagrangian Multiplier" → method for constrainted optimization problem

2a) ALM method etc.

Recap: $\ell(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k p_g(x_i; \mu_k, \Sigma_k) \right\}$

From Calculation, we have:

$$\begin{cases} \frac{\partial \ell}{\partial \mu_\ell} = (-\Sigma_\ell^{-1}) \sum_{i=1}^{n} \gamma_{i\ell} (x_i - \mu_\ell) \\ \frac{\partial \ell}{\partial \Sigma_\ell} = -\frac{1}{2} \sum_{i=1}^{n} a_{i\ell} \left[\Sigma_\ell^{-1} - (x_i - \mu_\ell)(x_i - \mu_\ell)^T \Sigma_\ell^{-2} \right] \\ \frac{\partial \mathcal{L}}{\partial \pi_\ell} = \sum_{i=1}^{n} a_{i\ell} \frac{1}{\pi_\ell} - \lambda \end{cases}$$

Therefore, roughly speaking, we can apply GD Framework:

$$\mu_\ell \leftarrow \mu_\ell - \eta \frac{\partial \ell}{\partial \mu_\ell}$$

$$\Sigma_\ell \leftarrow \Sigma_\ell - \eta \frac{\partial \ell}{\partial \Sigma_\ell}$$

→ there are Constraint on $(\pi_1, \dots, \pi_K)$

$\Rightarrow \boxed{\sum_{i=1}^{K} \pi_i = 1}$

2b) Solve $\nabla_\theta \ell(\hat{\theta}) = 0$ ⟶ Non-linear Equation

$$\begin{cases} \frac{\partial \ell}{\partial \mu_\ell} = (-\Sigma_\ell^{-1}) \sum_{i=1}^{n} \gamma_{i\ell} (x_i - \mu_\ell) = 0 \\ \frac{\partial \ell}{\partial \Sigma_\ell} = -\frac{1}{2} \sum_{i=1}^{n} a_{i\ell} \left[\Sigma_\ell^{-1} - (x_i - \mu_\ell)(x_i - \mu_\ell)^T \Sigma_\ell^{-2} \right] = 0 \\ \frac{\partial \mathcal{L}}{\partial \pi_\ell} = \sum_{i=1}^{n} a_{i\ell} \frac{1}{\pi_\ell} - \lambda = 0 \Leftrightarrow \pi_\ell = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i\ell} \end{cases}$$

Issue: No closed-form solution

↓

use iterative method to solve this non-linear system:

$$\begin{cases} \mu_\ell = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}} \sum_{i=1}^{n} \gamma_{i\ell} x_i \\ \Sigma_\ell = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}} \sum_{i=1}^{n} \gamma_{i\ell} (x_i - \mu_\ell)(x_i - \mu_\ell)^T \\ \pi_\ell = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i\ell} \end{cases}$$

where $\gamma_{i\ell} := \frac{\pi_\ell p_g(x_i; \mu_\ell, \Sigma_\ell)}{\sum_{k=1}^{K} \pi_k p_g(x_i; \mu_k, \Sigma_k)}$

$$\begin{cases} \mu_\ell^{(t+1)} = \dfrac{1}{\sum_{i=1}^{n} \gamma_{i\ell}^{(t)}} \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} x_i \\ \Sigma_\ell^{(t+1)} = \dfrac{1}{\sum_{i=1}^{n} \gamma_{i\ell}^{(t)}} \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} (x_i - \mu_\ell^{(t)})(x_i - \mu_\ell^{(t)})^T \\ \pi_\ell^{(t+1)} = \dfrac{1}{n} \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \end{cases}$$

Rmk: This actually corresponds to the EM Framework Update.

→ Therefore, this can be viewed as another perspective of EM:

⇒ MLE

↓

first-order Necessary Condition

↓

Solution of Non-linear Equation (system)

↓

Iterative Solving Method

Diagram

2C) EM Framework → the most successful one

Recap: $$l(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k P_g(x_i; \mu_k, \Sigma_k) \right\}$$

likelihood of complete observation

$$L(x_i, A_i = \ell \mid \theta) = P_g(x_i \mid \mu_\ell, \Sigma_\ell)$$

$$= \sum_{i=1}^{n} \log \left\{ \mathbb{E}_{A_i \sim \theta} [L(x_i \mid A_i, \theta)] \right\}$$

$$= \log \left\{ \mathbb{E}_{A \sim \theta} [L(X \mid A, \theta)] \right\}$$ → re-write

↓

difficult to optimize directly

A: hidden variable

↓

"assignment"

EM Framework

Big idea: instead of optimizing log-likelihood (primal objective) directly, we

1) optimize on "another objective" (easy to deal with)

2) the solution of "another objective" is still better with respect to the log-likelihood (primal objective) (discuss later)

Realization:

A: hidden variable → "assignment"

1) "another objective" is:

$$\mathbb{E}_{A\sim\theta^{(t)},X}[\log L(X, A|\theta)]$$

Recap: the "primal objective" is: $\log\{\mathbb{E}_{A\sim\theta}[L(X, A|\theta)]\}$

in EM Framework, $\theta^{(t+1)} = \arg\max \mathbb{E}_{A\sim\theta^{(t)},X}[\log L(X, A|\theta)]$

2) Recap:

we actually want $\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \ell(\theta)$

From the Property of EM Algorithm,

we can show that $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$

$\mathbb{E}_{A\sim\theta^{(t)},X}[\log L(X, A|\theta)]$ ← "another objective"

☆ interpretation: the solution from solving "another objective" is actually better in the sense of "primal objective"

$\ell(\theta) = \log\{\mathbb{E}_{A\sim\theta}[L(X|A, \theta)]\}$

Q1: What is the solution of $\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{A \sim \theta^{(t)}, X}[\log L(X, A \mid \theta)]$

↓

Answer:

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{A \sim \theta^{(t)}, X}[\log L(X, A \mid \theta)]$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{A \sim \theta^{(t)}, X}\left[\sum_{i=1}^{n} \log L(X_i, A_i \mid \theta)\right]$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{A \sim \theta^{(t)}, X}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik} \log\{\pi_k p_g(x_i; \mu_k, \Sigma_k)\}\right]$$

$$\left[\text{here, } a_{ik} := \begin{cases} 1, & x_i \text{ belongs to cluster } k \\ 0, & \text{o/w} \end{cases}\right]$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{A_i \sim \theta^{(t)}, x_i}[a_{ik}] \log\{\pi_k p_g(x_i; \mu_k, \Sigma_k)\}$$

$$\mathbb{E}_{A_i \sim \theta^{(t)}, x_i}[a_{ik}] := \gamma_{ik}^{(t)} \text{ (for the agreement)}$$

$$= \mathbb{E}_{A_i \sim \theta^{(t)}}[\mathbb{1}\{x_i \text{ belongs to cluster } k\} \mid X_i = x_i]$$

$$= P_{\theta^{(t)}}(x_i \text{ belongs to cluster } k \mid X_i = x_i)$$

$$= \frac{\pi_k p_g(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j p_g(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}$$

Q2: what is $\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^{(t)} \log\{\pi_k p_g(x_i; \mu_k, \Sigma_k)\}$?

Answer : $\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \log \{ \pi_k (2\pi)^{-\frac{d}{2}} (\det(\Sigma_k))^{-\frac{1}{2}} \exp\{ -\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \}$

$$\propto \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \left[\log \pi_k - \frac{1}{2} \log \{\det(\Sigma_k)\} - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right]$$

$$:= \underset{\theta \in \Theta}{\operatorname{argmax}} \; J(\theta)$$

① $\frac{\partial J}{\partial \mu_\ell} = \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \Sigma^{-1} (x_i - \mu_\ell)$

② $\frac{\partial J}{\partial \Sigma_\ell} = \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \cdot \left[-\frac{1}{2} \Sigma_\ell^{-1} - \frac{1}{2} (x_i - \mu_\ell)(x_i - \mu_\ell)^T \cdot (-\Sigma_\ell^{-2}) \right]$

$= \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \left[-\frac{1}{2} \Sigma_\ell^{-1} + \frac{1}{2} (x_i - \mu_\ell)(x_i - \mu_\ell)^T \Sigma_\ell^{-2} \right]$

③ $\frac{\partial \mathcal{L}}{\partial \pi_\ell} = \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \cdot \frac{1}{\pi_\ell} - \lambda$

$$\Rightarrow \begin{cases} \frac{\partial J}{\partial \mu_\ell} = 0 \Rightarrow \mu_\ell^{(t+1)} = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}^{(t)}} \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} \cdot x_i \\ \frac{\partial J}{\partial \Sigma_\ell} = 0 \Rightarrow \Sigma_\ell^{(t+1)} = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}^{(t)}} \sum_{i=1}^{n} \gamma_{i\ell}^{(t)} (x_i - \mu_\ell^{(t+1)})(x_i - \mu_\ell^{(t+1)})^T \\ \frac{\partial \mathcal{L}}{\partial \pi_\ell} = 0 \Rightarrow \pi_\ell^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{i\ell}^{(t)}}{n} \end{cases}$$

$\theta^{(t+1)} := \{ \pi_\ell^{(t+1)}, \mu_\ell^{(t+1)}, \Sigma_\ell^{(t+1)} \}_{\ell=1}^{K}$

#

To conclude, although 2b) and 2c) share the similar formula, the idea is totally different!

2b) → lecture note

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ l(\theta)$$

($\hat{\theta} \neq \hat{\theta}_{MLE}$ may happen) ↓ 這一步很粗糙 (necessary condition)

$\hat{\theta}$ may not be the exact solution since this is only necessary condition!

$$\nabla_\theta l(\hat{\theta}) = 0$$

↓

solution of Non-linear Equation System

↓

Iterative method for solving Equation System

⇓

direct method in my personal perspective

↓

directly solve $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ l(\theta)$

2C) → EM

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ l(\theta)$$

↓ difficult to solve directly

(曲线救国) ↘

consider $\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ J(\theta, \theta^{(t)})$

terminology: $J(\theta, \theta^{(t)})$ is auxilary function of log-likelihood $l(\theta)$

⇒ $\theta^{(t+1)}$ is exact solution since $J(\theta, \theta^{(t)})$ can be good enough w.r.t θ

($l(\theta)$ can be very bad!)

⇓

indirect method since it will not solve $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ l(\theta)$ directly

☆ Instead, it trys to find one $\theta^{(L)}$ such that it can maximize $l(\theta)$ as much as possible!

12.5 More EM Framework

1. **Observation**

log-likelihood → $\log p(X|\theta)$

$$= \log p(X, Z|\theta) - \log p(Z|X, \theta)$$

$$= \sum_z q(z) \log \frac{P(x, z|\theta)}{q(z)} - \sum_z q(z) \cdot \log \frac{P(z|x,\theta)}{q(z)}$$

$$= \sum_z q(z) \log \frac{P(x, z|\theta)}{q(z)} + \sum_z q(z) \log \frac{q(z)}{p(z|x,\theta)}$$

$$= \underbrace{\mathcal{L}(q, \theta)} + \underbrace{KL(q \,\|\, p_z)}$$

$$\geq \mathcal{L}(q, \theta) \qquad \boxed{p_z := p(\cdot|x, \theta)}$$

To conclude:

$$\ell(\theta) = \log P(x|\theta)$$
$$\geq \sum_z q(z) \log \frac{P(x, z|\theta)}{q(z)}$$

$$:= \mathcal{L}(q, \theta)$$

when $q = p_z = p(z|x,\theta)$. then $\ell(\theta) = \mathcal{L}(q, \theta)$

2. Consider the following thing:

① $\underset{q}{\operatorname{argmax}}\, \mathcal{L}(q, \theta) = \underset{q}{\operatorname{argmin}}\, KL(q \,\|\, p_z) = p_z$

② $\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q, \theta) = \underset{\theta}{\operatorname{argmax}} \sum_z q(z) \log \frac{P(x, z|\theta)}{q(z)}$

$$= \underset{\theta}{\operatorname{argmax}} \sum_z q(z) \log P(x, z|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{z \sim q}[\log P(x, z|\theta)]$$

3. 2 Diagrams for illustration:

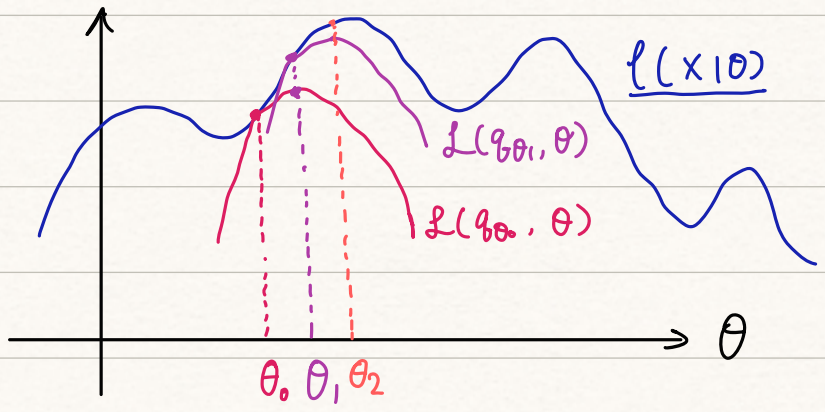
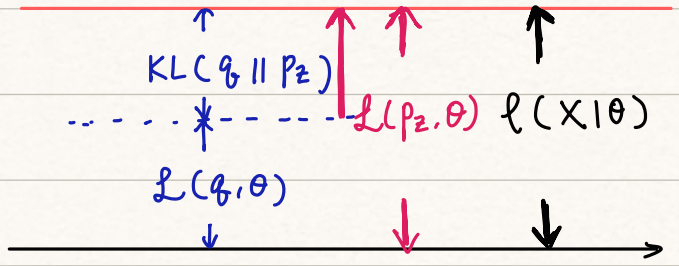
① 

Diagram 1

$q_{\theta_0} := P(\cdot \mid X, \theta_0)$ $q_{\theta_1} := P(\cdot \mid X, \theta_1)$

② $\underset{q}{\operatorname{argmax}}\ \mathcal{L}(q, \theta) = p_z$



$\underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(q, \theta) = \theta^{(new)}$