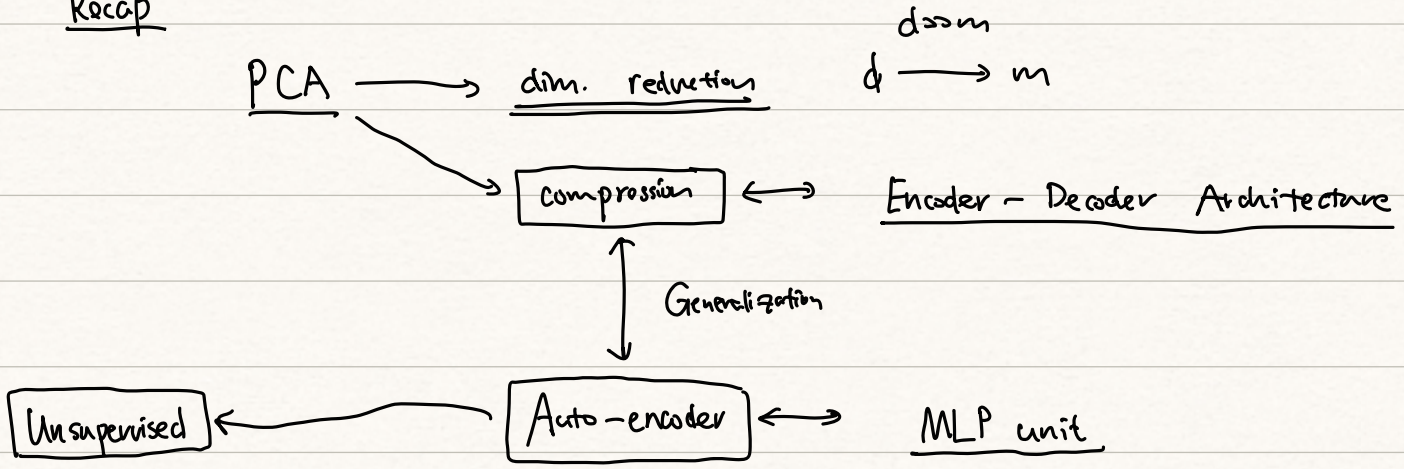
DSA5105 lec7 · clustering + density estimation

Recap



Unsupervised → cluster: K-means

→ { DE, cluster } → GMM

① cluster

a) Defn → ① $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{D}_i$

② within same group, similar

different groups, dissimilar → similarity metric

b) Why important?

① information retrieve ↔ search → match cluster

② Genetics

③ Market Research → 用户 → partition ← target doctor

④ Social Network Analysis ← Graph

聚类

Spectral Clustering

⑤ Anomaly Detection ⟶ focus on outliers

C) Algo ⟶ K-means

1) Re-assignment ← min dist

2) centroid ← aver

← Idea (repeat)

⟶ Mathematically formulate

① Assignment Matrix $R \in \mathbb{R}^{N \times K}$

$$r_{n,k} = \begin{cases} 1, & x_n \text{ is assigned to cluster } k \\ 0, & \text{o/w} \end{cases}$$

property ⟶

- $R \cdot \mathbb{1}_K = \mathbb{1}_N$ ⟷ 行和为 1
- $R^T \cdot \mathbb{1}_N =$ (# of points in each cluster)

② Minimize some loss

$$J(R, Z) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \| x_i - z_j \|_2^2$$

R: assignment, Z: Centroid

each row (i-th), there is only one non-zero entry

$$= \sum_{i=1}^{N} r_{i,r(i)} \| x_i - z_{r(i)} \|_2^2$$

Combinatorial optimization problem $\leftrightarrow$ partition (infeasible)

partition $\leftrightarrow$ R

partition $\rightarrow$ NP-HARD

approximation (simplify the problem)

3 Step-wise optimization (greedy algorithm) $\rightarrow$ ☆ Idea: trade-off between { feasibility, loyalty }

assume we have R, then try to find Z (vice versa)

Analysis

① given Z, find R

$$L(R, Z) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{k} r_{ij} \| x_i - z_j \|_2^2$$

assume centroids are fixed

$$= \frac{1}{2N} \sum_{i=1}^{N} \| x_i - z_{j(i)} \|_2^2$$

$$\Rightarrow \boxed{j(i) = \underset{k \in [K]}{\operatorname{argmin}} \| x_i - z_k \|_2^2}$$

$r_{i,j(i)} = 1$ determineds the assignment matrix R

② given R, find Z. → intuitively, we should choose average within each cluster

assume that assignment is fixed

$$L(R, Z) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{k} r_{ij} \| x_i - z_j \|_2^2$$

Pf1 →

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

$$\hat{z}_j = \frac{\sum_{i=1}^{N} r_{ij}\, x_i}{\sum_{i=1}^{N} r_{ij}}$$

↓

average within cluster

Pf 2 → calculate $\nabla_z L(R, Z)$ directly

$$\frac{\partial L(R,Z)}{\partial z_\ell} = \sum_{i=1}^{N} r_{i\ell}\,(x_i - \hat{z}_\ell) = 0$$

$$\Rightarrow \hat{z}_\ell = \frac{\sum_{i=1}^{N} r_{i\ell}\, x_i}{\sum_{i=1}^{N} r_{i\ell}} := N_\ell$$

④ Initialization → generate Z^0 randomly

↓

have some Guideline to modify the algorithm

↓

K-means ++

⑤ ☆ Guarantee to converge with finite steps !

↓

⇒ although we may not achieve the optimal solution

→ highly likely not to

↓

depends on the initialization

⑥ Limitation

a) Boundary points ⟶ lack of uncertainty

↓

one data can only belong to one cluster (hard assignment)

⇓

Motivation for GMM

↳ soft assignment

---

② GMM

$$\sum_{k=1}^{K} p(x, z=z_k) = \sum_{k=1}^{K} p(z=z_k) \cdot p(x \mid z=z_k)$$

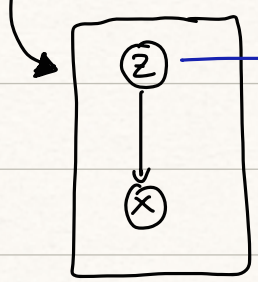
Model: $p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathbb{P}(x \mid C_k)$ → Gaussian Model $\mathbb{P}(x \mid \mu_k, \Sigma_k)$

↳ is actually a distribution → Sampling (Monte Carlo)

① from $\pi = (\pi_1, \ldots, \pi_K) \rightarrow c$

② from $X \mid C_k = c \rightarrow$ Gaussian

Graphic Model



(Ident) hidden variable → can apply EM to make approximation

$$\ln = \sum_{i=1}^{n} \log p(x_i \mid \theta) \qquad \theta = \{ \{\pi_k\}_{k=1}^{K}, \{\mu_k, \Sigma_k\}_{k=1}^{K} \}$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \cdot p(x_i \mid C_k)$$ → difficult to optimize ↓ Cross-term

$$= \sum_{i=1}^{n} \log \mathbb{E}_{z \sim \pi} p(x \mid z)$$

Preparation for EM

→ Posterior Dist for π_k

$$\Rightarrow \mathbb{P}(z = z_k \mid x) = \frac{\mathbb{P}(x \mid C_k) \cdot \mathbb{P}(C_k)}{\sum_k \mathbb{P}(x \mid C_k) \cdot \mathbb{P}(C_k)}$$

↳ improvement compared with hard assigment K-means

⇒ we are able to say:

↳ for one data point x, its component in each class → means a lot in application of cluster

☆ ↓

we can say, for one user:

{ 50% read books
  50% watch TV

something like that!

☆ 补全完了

Optimization $\quad \theta = \{ \mu_k, \Sigma_k, \pi_k \}_{k=1}^{K}$ $\quad P_{Gauss}(x_i | \mu_k, \Sigma_k)$

→ Directly solve MLE! (not EM Framework) $\quad = \frac{1}{(\sqrt{2\pi})^d} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right)$

$$\underline{ll} := \ell_n(\theta) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i | \mu_k, \Sigma_k) \right\}$$

↓

log-likelihood

① $$\frac{\partial ll}{\partial \mu_\ell} = \sum_{i=1}^{N} \frac{\pi_\ell \cdot \frac{\partial}{\partial \mu_\ell} P_{Gauss}(x_i | \mu_\ell, \Sigma_\ell)}{\left\{ \sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i | \mu_k, \Sigma_k) \right\}}$$

$$= \sum_{i=1}^{N} \frac{\pi_\ell \cdot \left(-\Sigma_\ell^{-1} \cdot (x_i - \mu_\ell) \right) \cdot P_{Gauss}(x_i | \mu_\ell, \Sigma_\ell)}{\sum_{k=1}^{K} \pi_k P_{Gauss}(x_i | \mu_k, \Sigma_k)}$$

not closed-form !!

$$= \left(-\Sigma_\ell^{-1}\right) \cdot \sum_{i=1}^{N} r_{i\ell} \cdot (x_i - \mu_\ell) \qquad r_{i\ell} = P(C_\ell | x_i)$$

posterior dist.

⇒ $$\mu_\ell = \frac{\sum_{i=1}^{N} r_{i\ell} \cdot x_i}{\sum_{i=1}^{N} r_{i\ell}}$$

Soft-assigument to cluster ℓ

② $S_\ell := \Sigma_\ell^{-1}$ → $\frac{\partial ll}{\partial S_\ell}$ → ☆ need to check

$$\mathcal{LL} = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i \mid \mu_k, \Sigma_k) \right\}$$

$$\frac{\partial \mathcal{LL}}{\partial \Sigma_\ell} = \sum_{i=1}^{n} \frac{\partial}{\partial \Sigma_\ell} \left[\log \left\{ \sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i \mid \mu_k, \Sigma_k) \right\} \right]$$

$$= \sum_{i=1}^{n} \frac{\pi_\ell}{\sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i \mid \mu_k, \Sigma_k)} \cdot \frac{\partial}{\partial \Sigma_\ell} P_{Gauss}(x_i \mid \mu_\ell, \Sigma_\ell)$$

$P_{Gauss}(x_i \mid \mu_\ell, \Sigma_\ell)$

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

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

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

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

$$= \sum_{i=1}^{n} \frac{\pi_\ell}{\sum_{k=1}^{K} \pi_k \cdot P_{Gauss}(x_i \mid \mu_k, \Sigma_k)} \left[-\frac{1}{2} P_{Gauss}(x_i \mid \mu_\ell, \Sigma_\ell) \cdot \Sigma_\ell^{-2} \left(\Sigma_\ell - (x_i - \mu_\ell)(x_i - \mu_\ell)^T \right) \right]$$

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

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

→ Not closed form

③ → π_ℓ → Lagrangian Multiplier

$$\mathcal{L} = \mathcal{LL} + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_\ell} = \sum_{i=1}^{N} \frac{1}{\pi_\ell} \frac{\pi_\ell P_{Gaus}(x_i; \mu_\ell, \Sigma_\ell)}{\sum \dots \dots} - \lambda = 0$$

$$\Rightarrow \sum_{i=1}^{N} \gamma_{i\ell} - \pi_\ell \cdot \lambda = 0$$

$$\begin{cases} \sum_{\ell=1}^{K} \pi_\ell = 1 \\ \sum_{\ell=1}^{K} \sum_{i=1}^{N} \gamma_{i\ell} = \sum_{i=1}^{N} \cdot 1 = N \end{cases}$$

$$\Rightarrow N - \lambda = 0 \Rightarrow \boxed{\lambda = N}$$

$$\Rightarrow \pi_\ell = \frac{\sum_{i=1}^{N} \gamma_{i\ell}}{N} \rightarrow \text{Not closed form}$$

→ not closed form
→ only solve iteratively

To conclude: after solving MLE, we achieve:

① $\mu_\ell = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}} \cdot \sum_{i=1}^{n} \gamma_{i\ell} x_i$

② $\Sigma_\ell = \frac{1}{\sum_{i=1}^{n} \gamma_{i\ell}} \cdot \sum_{i=1}^{n} \gamma_{i\ell} (x_i - \mu_\ell)(x_i - \mu_\ell)^T$

③ $\pi_\ell = \frac{\sum_{i=1}^{n} \gamma_{i\ell}}{N}$

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

Self-Recall

EM → GMM

Notation

z_{ik} → i-th X ⇒ k-th class

Primal formulation:

$$LL_1 = \log L_n(X|\theta) = \log\left\{\prod_{i=1}^{n}\cdot\sum_{k=1}^{K} P(z_{ik})\cdot P(x_i|z_{ik})\right\}$$

$$= \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K} P(z_{ik})\cdot P(x_i|z_{ik})\right\}$$

EM Approximation ← Effectiveness of EM Algo can be proved by MM

$$LL_2 = \mathbb{E}_{z\sim P(\cdot|X,\theta^e)}[\log L_n(X,Z|\theta)]$$ Framework

$$\theta^{t+1} = \arg\max_{\theta\in\Theta} LL_2 = \arg\max_{\theta\in\Theta} \mathbb{E}_{z\sim P(\cdot|X,\theta^e)}[l_n(X,Z|\theta)]$$

$$L_n(X,Z|\theta) = \prod_{i=1}^{n} P(x_i|z_{i,k(i)})$$

$$= \prod_{i=1}^{n}\cdot\prod_{k=1}^{K} P(x_i|z_{i,k})^{\delta_{i,k}}$$

It can be proved: $l_n(\theta^{t+1}) \geq l_n(\theta^e)$

$$\Rightarrow \log l_n(X,Z|\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{i,k}\log P(x_i|z_{i,k})$$

$$= \sum_{k=1}^{K}\left\{\sum_{i=1}^{n}\delta_{i,k}\log P_{Gauss}(x_i|\mu_k,\Sigma_k)\right\}$$

$$\Rightarrow \mathbb{E}_{z\sim P(Z|X,\theta^e)}[l_n(X,Z|\theta)] = \sum_{k=1}^{K}\left\{\sum_{i=1}^{n}\mathbb{E}[\delta_{i,k}|X,\theta^e]\log P_{Gauss}(x_i,\mu_k,\Sigma_k)\right\}$$

$$\mathbb{E}[\delta_{ik}|X,\theta^e] = P(x_i\to \text{class } k|X,\theta^e)$$

$$= \frac{\pi_k^e P_G(x_i|\mu_k^e,\Sigma_k^e)}{\sum_{k=1}^{K}\pi_k^e\cdot P_G(x_i|\mu_k^e,\Sigma_k^e)}$$