MA4270 Lec 17

Mixture Models ⇒ clusters of Data

E.g. Sopherical Gaussian MM

$$P(x; \underline{\theta}, m) = \sum_{j=1}^{m} P(j)\, N(x; \mu_j, \sigma_j^2 I)$$

$$\underline{\theta} = \{P(j), \mu_j, \sigma_j^2\}_{j=1}^{m}$$

Find where the clusters are, RATHER THAN $\underline{\theta}$.

→ To keep the model simpler, $P(j) = \frac{1}{m}$, $\forall j = 1, 2, \ldots, m$

$$P(x; \underline{\theta}, m) = \frac{1}{m} \sum_{j=1}^{m} N(x; \mu_j, \sigma_j^2 I)$$

→ To simplify the model <u>even more</u>, let $\sigma_j^2 = \sigma^2$ $\forall j = 1, 2, \ldots, m$

$$P(x; \underline{\theta}, m) = \frac{1}{m} \sum_{j=1}^{m} N(x; \mu_j, \sigma^2 I)$$

$$\underline{\theta} = \{\{\mu_j\}_{j=1}^{m}, \sigma^2\}$$

Q: How are points are assigned to clusters in the EM alg.?

E-step: We compute $P(i \mid x_t, \underline{\theta})$ (→ posteior assignment!) under current paras $\underline{\theta}$

for every x_t & for every $i=1,2,\dots,m$

Consider the points $x \in \mathbb{R}^d$ s.t $P(i|x_t,\theta) = P(j|x_t,\theta)$ for $i \neq j$

↙ calculation

$$\frac{P(x_t|i,\theta)\,P(i)}{P(x_t|\theta)} = \frac{P(x_t|j,\theta)\,P(j)}{P(x_t|\theta)} \qquad P(i)=P(j)=\frac{1}{m}$$

$$\Rightarrow P(x|i,\theta) = P(x|j,\theta)$$

$$\Rightarrow N(x;\mu_i,\sigma^2 I) = N(x;\mu_j,\sigma^2 I)$$ 常数项相等!

$$\Rightarrow \exp\left(-\frac{1}{2\sigma^2}\|x-\mu_i\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}\|x-\mu_j\|^2\right)$$

The set of points x s.t $P(i|x,\theta) = P(j|x,\theta)$ is precisely those x s.t

$$\|x-\mu_i\| = \|x-\mu_j\|$$ 二次项消失

$$\Leftrightarrow 2x^T(\mu_i-\mu_j) = \|\mu_i\|^2 - \|\mu_j\|^2$$

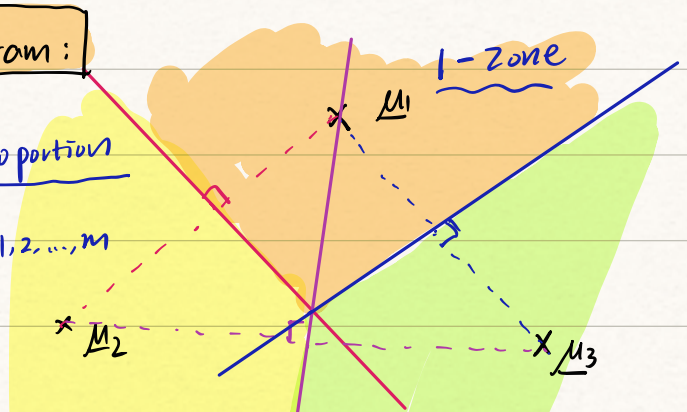
⇓

define a line!

⇓

Boundary is linear with x!

Diagram:



equal mixture proportion

$P(j) = \frac{1}{m} \quad \forall j=1,2,\dots,m$

x: $P(x|1,\theta) = P(x|2,\theta)$

x: $P(x|1,\theta) = P(x|3,\theta)$

x: $P(x|2,\theta) = P(x|3,\theta)$

2-zone | 3-zone

(k-means)

Consider the following simpler (hard-assignment) version of EM.

E-step: Assign each point x_t to the Voronoi region (different (which is hard assignment)) with hard prob.

$$x_t \longrightarrow j_t = \underset{1 \le j \le m}{\operatorname{argmin}} \|x_t - \mu_j\|^2$$

M-step: Recompute μ_j as the mean of assigned points

$$C_j = \{ x_t : j_t = j \}$$

$$\Downarrow$$

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x \qquad \forall j = 1, 2, \dots, m$$

Rmk: The M-step is equivalent to optimizing (maximizing) the Complete Log-likelihood w.r.t the assignment $j_1, \dots, j_t$ (obtained in E-step) over the unknown means $\mu_1, \dots, \mu_j$

K-means: Algorithm

Input: $\mathcal{D} = \{x_t\}_{t=1}^n \subset \mathbb{R}^d$

1) step 0: Initialize centroids $\mu_j^{(0)}$ $j = 1, 2, \dots, m$ SMARTLY.

2) step $\ell \in \mathbb{N}$ (E-step): Assign each pt to CLOSEST Mean

$$j_t^{(\ell)} = \underset{}{\operatorname{argmin}} \|x_t - \mu_j\|^2 \qquad t = 1, 2, \dots, n$$

$j = 1, \dots, m$

3) Step $\ell \in \mathbb{N}$ (M-step): Recompute mean:

$$\mu_j^{(\ell+1)} = \frac{1}{|C_j^{(\ell)}|} \sum_{t : j_t^{(\ell)} = j} x_t \qquad C_j^{(\ell)} = \{ x_t : j_t^{(\ell)} = j \}$$

4) Terminate when cluster assignment $j_1^{(\ell)}, j_2^{(\ell)}, \dots, j_n^{(\ell)}$ do not change!

⇓

Natural terminate criterion

The cost Function for a Given partition $\mathcal{D} = \bigcup_{j=1}^{m} C_j$

$$\rightarrow C_j = \{ x \in D : j = \underset{j'=1,2,\dots,m}{\operatorname{argmin}} \|x - \mu_{j'}\|^2 \}$$

Consider the cost function / objective function:

$$\Rightarrow J(C_1, \dots, C_m ; \mu_1, \dots, \mu_m) = \sum_{j=1}^{m} \sum_{x \in C_j} \|x - \mu_j\|^2$$

this part actually is fixed by certain partition

⇓

$\sum$ the distance^2 for each x to their own centroid μ

C_j: set of points closer to μ_j than any other μ_i $(i \neq j)$

Last Lecture teaches us!

[Proposition] : GMM ⇒ log-likelihood is non-decreasing ! (EM)

Let $J^{(\ell)} = J(C_1^{(\ell)}, \dots, C_m^{(\ell)}; \mu_1^{(\ell)}, \dots, \mu_m^{(\ell)})$

$C_1^{(\ell)}, \dots, C_m^{(\ell)}$: ℓ^{th} iteration of k-means

① The k-means algorithm results in :

$J^{(\ell+1)} \leq J^{(\ell)}$ for all $\ell \in \mathbb{N}$

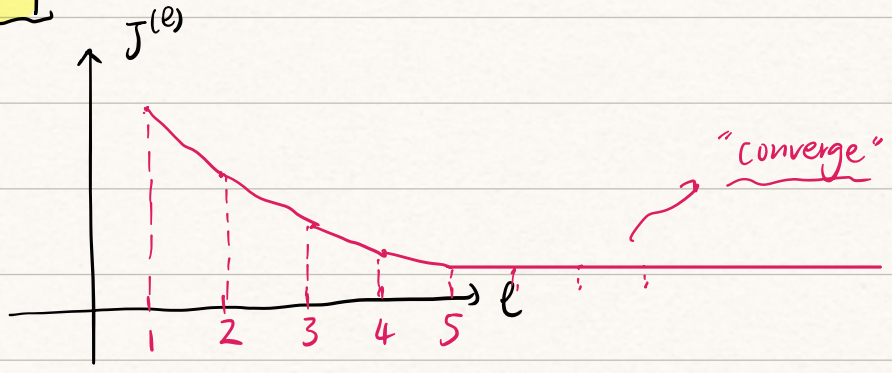
this is not enough for termination !

$J^{(\ell)}$ cannot increase !

② The alg. terminates in a finite # of steps ($\leq m^n$)

stop criterion is effective

diagram



---

① Pf: Once $\mathcal{D}$ are assigned to their initial clusters, we compute

$\| x_t - \mu_{j_t} \|^2$

Summing this up over all $t = 1, 2, \dots, n$, yields :

$J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(1)}, \dots, \mu_m^{(1)})$

$C_j^{(1)} = \{ x_t \in \mathcal{D} : j = j_t^{(1)} = \underset{j'=1,2,\dots m}{\operatorname{argmin}} \| x_t - \mu_{j'}^{(1)} \|^2 \}$.

For the M-step, we min the total dist. of each clusters samples from the resp. centroid so the cost $f^{\wedge}$ is

$$J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(2)}, \dots, \mu_m^{(2)})$$

The function $f(\mu) = \sum_{x \in C_j} \|x - \mu\|^2$ convex with μ.

$$\Rightarrow \nabla_\mu f(\mu) = 0 \Rightarrow \sum_{x \in C_j} x = \mu |C_j|$$

$$\Rightarrow \mu_j^{(2)} = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

the M-step actually minimizes! (for fixed assignment)

$\Updownarrow$

the choice of update paras yields $\mu_j^{(t)} \to \mu_j^{(t+1)}$ minimize for a fixed assignment! (M-step)

$$J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(2)}, \dots, \mu_m^{(2)})$$
$$\leq J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(1)}, \dots, \mu_m^{(1)})$$

partition

E-step: ① if a sample is assigned to a different centroid, it must be that the sample is closer to its new centroid than the previously assigned one!

② If a sample is not reassigned, then nothing happen!

From ① & ②

$$J(C_1^{(2)}, \dots, C_m^{(2)}; \mu_1^{(2)}, \dots, \mu_m^{(2)})$$
$$\leq J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(1)}, \dots, \mu_m^{(1)})$$

Thus, $J(C_1^{(2)}, \dots, C_m^{(2)}; \mu_1^{(2)}, \dots, \mu_m^{(2)})$

$\leq J(C_1^{(1)}, \dots, C_m^{(1)}; \mu_1^{(1)}, \dots, \mu_m^{(1)})$

i.e., $J^{(2)} \leq J^{(1)}$ → decrease!

Pf ②: Procedure terminates in a finite # of steps

There is a finite # of samples n

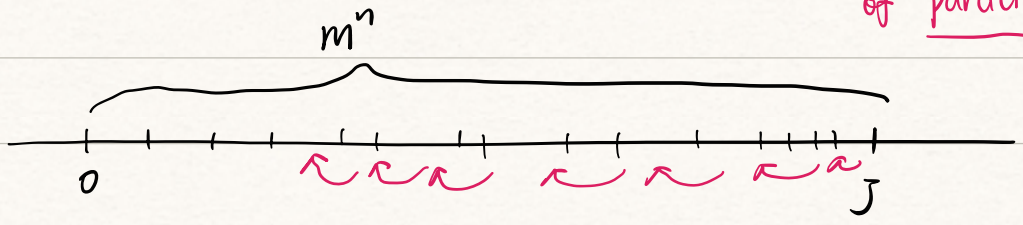
⇒ There is a finite # of labellings → m^n

Note that $J(C_1, \dots, C_m; \hat{\mu}_1, \dots, \hat{\mu}_m)$

where $\hat{\mu}_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$

60次幂

can only take on m^n values! ⇒ Actually J is a function of partition (m^n 个)



⇒ Alg. will terminate at iteration $\ell^* \in \mathbb{N}$ with

$J^{(\ell^*)} = J^{(\ell^*-1)}$ ⇒ then $J^{(\ell)} = J^{(\ell^*)}$ for all $\ell > \ell^*$

$0 \leq J^* \leq J^{(\ell^*)}$ → the 'convergence' Given By K-means

the true minimum of cost function

NP-Hard Problem to find J^*

⇓ Reason

since $J^* = \min_{C_i} J(C_1, \dots, C_m; \hat{\mu}_1, \dots \hat{\mu}_m)$

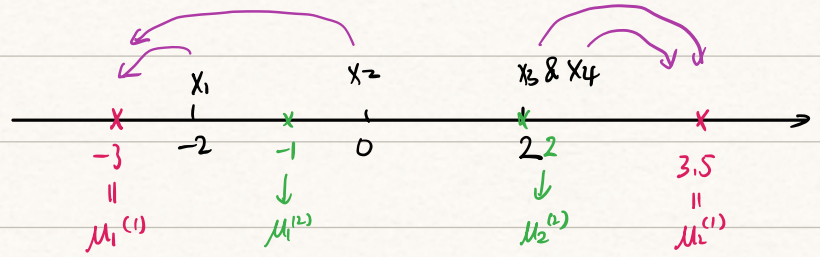
⇓

find the optimal over all possible m

partition (m^n ↑) ⇒ NP-Hard

Ex: $x_1 = -2,\ x_2 = 0,\ x_3 = x_4 = 2$

a) $\mu_1^{(1)} = -3 \quad \mu_2^{(1)} = 3.5$



Good initialization $J^{(1)} = 1^2 + 3^2 + 2 \times 1.5^2$

E-step: $j_1 = 1 \quad j_2 = 1 \quad j_3 = j_4 = 2$

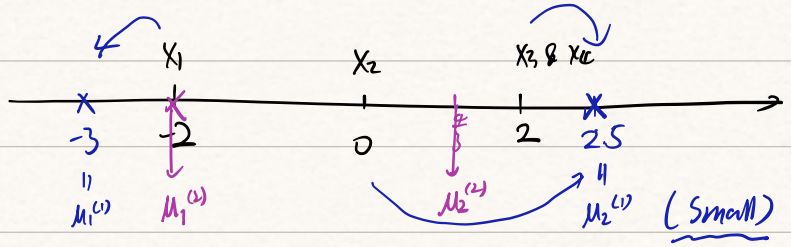
M-step: $\mu_1^{(2)} = -1 \quad \mu_2^{(2)} = 2$

⇒ $J^{(2)} = 1^2 + 1^2 + 0^2 + 0^2 = 2$ shrink!

→ Good

b) BAD Initialization

$\mu_1^{(1)} = -3, \quad \mu_2^{(1)} = 2.5$



E-step: $j_1 = 1 \quad j_2 = j_3 = j_4 = 2$

M-step: $\mu_1^{(2)} = -2 \quad \mu_2^{(2)} = \frac{4}{3}$

$J^{(2)} = 0^2 + (\frac{4}{3})^2 + 2 \cdot (\frac{2}{3})^2 = \frac{16}{9} + \frac{8}{9} = \frac{8}{3} > 2$

Stuck in $J^{(2)} = \frac{8}{3}$ ⇝ not global minimum of J

a BIG Cost! ⇒ 被困住了

Option Part:

K-means ++ Algorithm (2017)

(1) Take an center μ_1 uniformly chosen from $\mathcal{D} = \{x_t\}_{t=1}^{n}$

(2) Take μ_i $(i>1)$ choosing $x \in \mathcal{D}$ with prob. given by $\frac{D(x)^2}{\sum D(x')^2}$

$D(x)$: distance of $x \in \mathcal{D}$ to the closest center we have chosen.

chosen from data points!

(3) Repeat Step (2) with K centroids are found.

Intuition: Select centroids using the idea that centroids should be as far apart as possible!

⇓

with BAD Initialization, we will have BAD $J^{(t^*)}$

Given a set of K centers $C = \{\mu_1, \dots, \mu_K\}$

$$J = \sum_{x \in D} \min_{\mu \in C} \|x - \mu\|^2$$

$$J_{opt} = \min_{C} J \longrightarrow \text{NP-Hard Problem}$$

Thm: [AV. 2007]

If C is constructed based on a K-means ++

let its cost function be J a random variable.

↳ r.v since C is random

Then $\mathbb{E}[J] \leq 8(\ln K + 2) J_{opt} = O(\log K) J_{opt}$

J_{opt} ↓ Global minimum

$\mathbb{E}[J]$ ↓ control by SMART Initialization!

⇒ K-means ++ is $O(\log K)$-competitive w.r.t the optimal clustering (which is NP-HARD)

ICML (2020)

Informal

Using εK ($\varepsilon > 0$) local search steps, then with K-means +++, we can attain $O(1)$-competitive approx. the J_{opt}.

Some important point of K-means:

1. The decrease of cost function $J^{(t)}$

$J^{(t)} = J(D_1^{(t)}, \ldots, D_m^{(t)}; \mu_1^{(t)}, \ldots, \mu_m^{(t)})$.

partition according to $\mu_1^{(t)} \ldots \mu_m^{(t)}$

→ decrease

decrease

$J_1^{(t)} = J(D_1^{(t)}, \dots, D_m^{(t)}; \mu_1^{(t+1)}, \dots, \mu_m^{(t+1)})$.

↘ reset of centroid

$J^{(t+1)} = J(D_1^{(t+1)}, \dots, D_m^{(t+1)}; \mu_1^{(t+1)}, \dots, \mu_m^{(t+1)})$. ⇒ decrease

↘ reassign (repartition)

① if $J^{(t)} = J_1^{(t)}$, ⇒ no reset of centroid ⇒ $J^{(t)} = J^{(t+1)} = \dots$

⇓

no need of reassignment

② if $J_1^{(t)} = J^{(t+1)}$ ⇒ no reassignment

⇓

no need of reset ⇒ $J^{(t+1)} = J^{(t+2)} = \dots$

③ if $J^{(t)} = J^{(t+1)}$, then $J^{(t)} \overset{=}{\leq} J_1^{(t)} \overset{=}{\leq} J^{(t+1)}$ ⇒ $J^{(t)} = J^{(t+1)} = \dots$

⇓ no reset ⇓ no reassignment

2. local minimum of K-means

$\binom{m+n}{n}$

$J(D_1, \dots, D_m; \mu_1, \dots, \mu_m) = \sum_{i=1}^{m} \sum_{x \in D_i} \|x - \mu_i\|_2^2$

3. What kind of minimum (local) we can find?

⇓