MA4270 Lec 15 Mixture Model

Last time: Boosting

Ensemble: $h_m(x) = \sum_{j=1}^{m} \hat{\alpha}_j \, h(x; \hat{\theta}_j)$ — $\hat{\alpha}_j$: vote; $h(x;\hat{\theta}_j)$: base learner

Parameters $\theta = \{(\hat{\alpha}_j, \hat{\theta}_j)\}_{j=1}^{m} \Rightarrow \hat{\theta}_j = \{s, \theta_0, \underset{k}{\blacksquare}\}$.

This is a mixture model and we want to find more general ones.

Today: • Gaussian Mixture Model

• Inference / Learning of Parameters in GMM

given { incomplete (EM algorithm)
complete

Basic Mixture model: $\mathcal{D} = \{\underline{x}_t\}_{t=1}^{n} \Rightarrow$ unsupervised

also denote as X

Capture and resolve observable ambiguities in data

↳ m-component GMM is:

$$P(\underline{x}; \underline{\theta}) = \sum_{j=1}^{m} P(j) \, N(\underline{x}; \underline{\mu}_j, \Sigma_j)$$

① $\{P(j)\}_{j=1}^{m}$: mixing parameters $P(j) \geq 0$ & $\sum P(j) = 1$

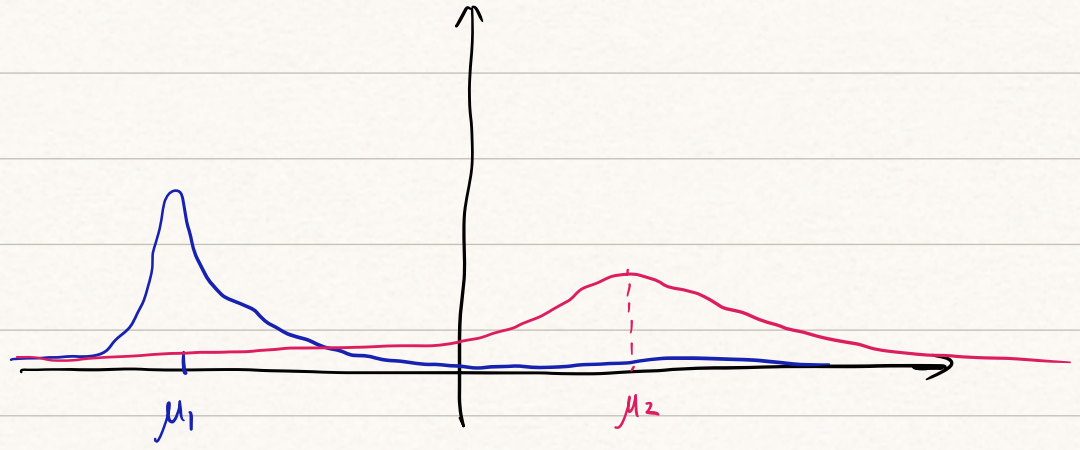
② $N(\underline{x}; \underline{\mu}_j, \Sigma_j) = \frac{1}{(\sqrt{2\pi})^d |\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)^T \Sigma_j^{-1} (\underline{x}-\underline{\mu}_j)\}$

$\{\mu_j\}_{j=1}^m$: component mean

$\{\Sigma_j\}_{j=1}^m$: component covariance matrix $\Rightarrow$ PD matrix

$\underline{\theta} = \{ \{P(j)\}_{j=1}^m ; \{\mu_j\}_{j=1}^m, \{\Sigma_j\}_{j=1}^m \} \Rightarrow$ total # is $\underline{3m}$

d=1



Example: $m = 3$ different neighbourhoods { poor, middle-class, rich

$t = 1, \ldots, n$, n: # of houses

$\underline{x_t} = \begin{bmatrix} x_{t1} \\ x_{t2} \end{bmatrix}$ x_{t1} : price of t-th house

x_{t2} : size of t-th house

Model the dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ as being drawn from GMM

Question:

- How many types of neighbourhoods? Model Selection
- Given a particular nbh. X, what is the { mean, covariance } $\{(\mu_j, \Sigma_j)\}_{j=1}^m$
- In $\mathcal{D}$, what fraction of x_t comes from nbh. of a certain type? $\{P(j)\}_{j=1}^m$

General Mixture Model: (can not be GMM)

$$P(x;\Theta) = \sum_{j=1}^{m} P(j)\, P(x \mid \theta_j)$$

$P(j)$: pick parameter

Likelihood of observing $\mathcal{D}$ given parameters $\Theta = \{\{P(j)\}, \{\theta_j\}\}$ ⇒ 术语: Generating Models

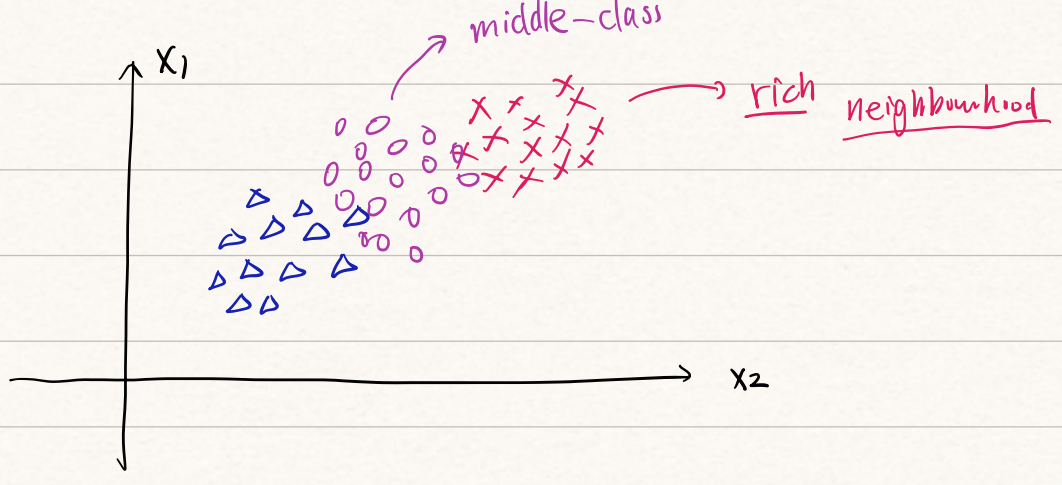
$\mathcal{D} = \{x_1, \dots, x_n\}$

$\theta_j = \{\mu_j, \Sigma_j\}$ for GMM

$$\mathcal{L}(\mathcal{D};\theta) = \prod_{t=1}^{n} P(x_t;\Theta)$$

$$= \prod_{t=1}^{n} \left(\sum_{j=1}^{m} P(j)\, P(x_t \mid \theta_j) \right)$$

Diagram:



Exercise: $\mathcal{L}(\mathcal{D};\theta) = \sum_{j=1}^{m} P(j) \left(\prod_{t=1}^{n} P(x_t \mid \theta_j) \right)$

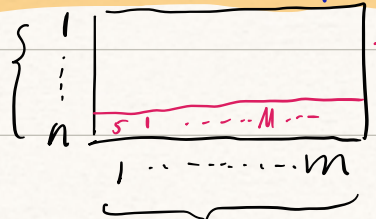
$\prod_{t=1}^{n} P(x_t \mid \theta_j)$: all the houses in certain type j

Story: All the houses in $\mathcal{D}$ are of the same type, but we DO NOT know which type of house represents $\mathcal{D}$.

Collaborative Filtering ⇒ Another Model Example

{ n users
  m movies

user



sparse

$r_{ij} \in \{1, \dots, 5\} \cup \{\text{Missing}\}$

movie

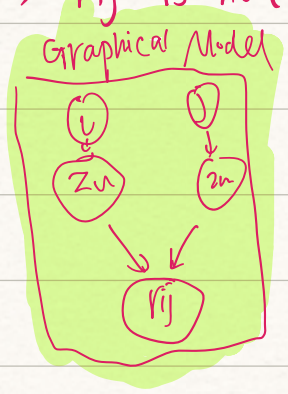
Each of the n users rates a subset of the m movies.

Provide $r_{ij} \in \{1, \dots, 5\}$ or MISSING

☆ defn of I_D

We say that $(i,j) \in I_D$ if r_{ij} is observed, i.e., r_{ij} is not missing. ($\mathcal{D}$ ⇒ set of observed ratings)

$\mathcal{D} = \{ r_{ij} : \text{user } i \text{ rated } j \text{ movie} \}$



{ Types of movies indexed by $z_m \in \{1, \dots, K_m\}$
Types of users indexed by $z_u \in \{1, \dots, K_u\}$ } ⇒ to make problem simpler!

↓ degenerate!

How is a rating generated?

$$P(r_{ij} \mid i, j, \theta) = \sum_{z_u=1}^{K_u} \sum_{z_m=1}^{K_m} P(z_u \mid i)\, P(z_m \mid j)\, P(r_{ij} \mid z_u, z_m)$$

$P(z_u \mid i)$: prob that user i is of type z_u

$P(z_m \mid j)$: prob that movie j is of type z_m

$P(r_{ij} \mid z_u, z_m)$: the rating given that { user is of type z_u, movie is of type z_m }

$$\theta = \{ \{P(r \mid z_u, z_m)\}, \{P(z_u \mid i)\}, \{P(z_m \mid j)\} \}$$

Likelihood of a dataset $\mathcal{D} = \{ r_{ij} : \text{user } i \text{ rate movie } j \}$

$$\mathcal{L}(\mathcal{D}; \theta) = \prod_{(i,j) \in I_D} P(r_{ij} \mid i, j, \theta)$$

$$= \prod_{(i,j) \in I_D} \sum_{z_u} \sum_{z_m} P(z_u \mid i)\, P(z_m \mid j)\, P(r_{ij} \mid z_u, z_m)$$

How many Parameters $\underline{\theta}$?

① $\{P(z_u|i)\}$ ⇒ $i=1,\dots,n$ ⇒ user number; $z_u=1,\dots,K_u$ ⇒ user type number

↓ mixture proportion

⇒ total # if $n\cdot(K_u-1)$

② $\{P(z_m|j)\}$ → $m\cdot(K_m-1)$

③ $\{P(r|z_u,z_m)\}$ → $(5-1)K_uK_m$

→ 自由参数

What is our aim?

fill-in missing value instead!

With a new data ⊻ come in, make prediction!

⇓

→ what kind of films he or she will like?

Once I have learnt $\hat{\theta}$ from $\mathcal{D}=\{r_{ij}: i \text{ rated movie } j\}$

Take user $i^*=1,\dots,n$ & $r_{i^*j^*}$ is not available (missing)

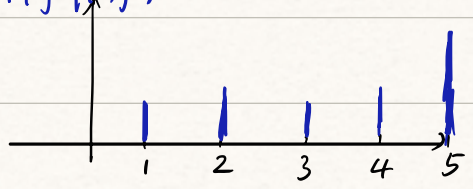
$P(z_u|i^*)$ ★ 这仅以前的投票倾向

$P(z_m|j^*)$

$P(r_{ij}|z_u,z_m)$

to predict $P(r_{i^*j^*}|i^*,j^*)$

$p(r_{i^*j^*}|i^*,j^*)$



Cold-start problem: New user who has not rated any movies comes in.

↳ using social media resources to attain more information

Parameter Estimation: 核心

GMM:

$$P(x;\theta) = \sum_{j=1}^{m} P(j)\, N(x; \mu_j, \Sigma_j)$$

Likelihood

注记：这两个 MLE 是有明显不同的

⇒ Complete Data 退化成了最简单的情形 (separable)

$$\mathcal{D} = \{x_t\}_{t=1}^{n} = X$$

Complete Dataset

① Pretend we know x_j belongs to ○ sub-mixture

Pretend that each observation $x_1, \ldots, x_n$ has additional information about component responsible for generating it

$j_1, \ldots, j_n$ ↳ Generating Information

$j_t \in \{1, \ldots, m\}$: component index of sample t

Def: $\delta(j|t) = \begin{cases} 1, & \text{if } j = j_t \\ 0, & \text{if } j \neq j_t \end{cases}$

$$\mathcal{L}(\mathcal{D};\theta) = \prod_{t=1}^{n} \sum_{j=1}^{m} P(j)\, N(x_t; \mu_j, \Sigma_j)$$ ⇒ difficult to deal with

Complete log-likelihood (Log-likelihood for complete data)

Complete data → $(X, Z) = \{x_t, j_t\}_{t=1}^{n}$

↳ which component generates x_t!

Very Easy to Deal with

$$\ell(\underbrace{x_1,\dots,x_n}_{X},\underbrace{j_1,\dots,j_n}_{Z};\theta)=\sum_{t=1}^{n}\log\left[P(j_t)\,N(x_t;\mu_{j_t},\Sigma_{j_t})\right]$$

✲ we assume that we have known x_t ← the j_t-th mixture

$$=\sum_{t=1}^{n}\sum_{j=1}^{m}\underbrace{\delta(j|t)}_{\text{筛选}}\log\left[P(j)\,N(x_j;\mu_j,\Sigma_j)\right]$$

$$=\sum_{j=1}^{m}\sum_{t=1}^{n}\delta(j|t)\log P(j)+\sum_{j=1}^{m}\sum_{t=1}^{n}\delta(j|t)\log N(x_j;\mu_j,\Sigma_j)$$

derivative → good property of separation

Claim: Maximize $\ell(X,Z;\theta)$ w.r.t $\underline{\theta}=\{\{P(j)\},\{\mu_j\},\{\Sigma_j\}\}$

yields that:

$\hat{n}(j)=\sum_{t=1}^{n}\delta(j|t)$

① $\hat{P}(j)=\dfrac{\hat{n}(j)}{n}$ ② $\hat{\mu}_j=\dfrac{1}{\hat{n}(j)}\sum_{t=1}^{n}\delta(j|t)\,x_t$

③ $\hat{\Sigma}_j=\dfrac{1}{\hat{n}(j)}\sum_{t=1}^{n}\delta(j|t)(x_t-\hat{\mu}_j)(x_t-\hat{\mu}_j)^T$

Pf: (① and ②)

① Note that $\sum_{j=1}^{m}P(j)=1$.

$$L(\{P(j)\},\lambda)=\sum_{j=1}^{m}\underbrace{\sum_{t=1}^{n}\delta(j|t)}_{\hat{n}(j)}\log P(j)+\lambda\left(1-\sum_{j=1}^{m}P(j)\right)$$

($x_1,\dots,x_n$ 属于 j-submodel 的数量.)

$$=\sum_{j=1}^{m}\hat{n}(j)\log P(j)+\lambda\left(1-\sum_{j=1}^{m}P(j)\right)$$

$$\frac{\partial L}{\partial P(j)}=0\Rightarrow\frac{\hat{n}(j)}{P(j)}=\lambda\Rightarrow P(j)=\frac{\hat{n}(j)}{\lambda}\Rightarrow\text{solve for }\lambda=n$$

$\Rightarrow$ optimal $P(j)$ must be proportional to $\hat{n}(j)$

$$\Rightarrow \quad \hat{P}(j) = \frac{\hat{n}(j)}{n}$$

② $$\ell(X, Z; \underline{\theta}) \overset{c}{=} \sum_{j=1}^{m} \sum_{t=1}^{n} \delta(j|t) \log\left[N(\underline{x}_t ; \mu_j, \Sigma_j)\right]$$

$$\overset{c}{=} \sum_{j=1}^{m} \sum_{t=1}^{n} \delta(j|t) \left[-\frac{1}{2} (\underline{x}_t - \mu_j)^T \Sigma_j^{-1} (\underline{x}_t - \mu_j) \right]$$

Then $\frac{\partial \ell}{\partial \mu_j} = 0 \Rightarrow \sum_{t=1}^{n} \delta(j|t) \left[\Sigma_j^{-1} (\underline{x}_t - \mu_j) \right] = 0$

$$\Rightarrow \sum_{t=1}^{n} \delta(j|t) (\underline{x}_t - \mu_j) = 0$$

$$\Rightarrow \boxed{\hat{\mu}_j = \frac{1}{\hat{n}(j)} \sum_{t=1}^{n} \delta(j|t)\, \underline{x}_t}$$

$\Downarrow$ just calculate in the neighbourhood !

while doing ③

↳ we must notice that $\Sigma_j \geq 0$

What if we cannot have the Ground-truth label ?

↳ we only have $X = \{\underline{x}_t\}_{t=1}^{n}$ $\longleftrightarrow$ we don't have $\delta(j|t)$

Method:

Replace them with estimation ! $(\underline{\theta}^{(t)})$

$\Vert$

use $\underline{\theta}^{(\ell)}$ to estimate $\delta(j|t)$

$$P(j|x_t, \underline{\theta}^{(\ell)}) = \frac{P(j, x_t|\underline{\theta}^{(\ell)})}{P(x_t|\underline{\theta}^{(\ell)})}$$

$$\Updownarrow$$

$$\delta(j|t) = \frac{p^{(\ell)}(j)\, P(x_t|j, \underline{\theta}^{(\ell)})}{\sum_{j'=1}^{m} p^{(\ell)}(j)\, P(x_t|j, \underline{\theta}^{(\ell)})}$$

use $(\mu_j^{(\ell)}, \Sigma_j^{(\ell)})$

★ → Highlight

Instead of the HARD Assignments $\{\delta(j|t)\}$, we use the soft Assignments $\{p^{(\ell)}(j|t)\}$ in ℓ-th iteration

→ EM Alg.

Step 0: Initialize $\underline{\theta}^{(0)}$

e.g. $P(j) = \frac{1}{m}$ for $\forall j$

μ_j: randomly chosen

Σ_j: randomly chosen PD.

Step ℓ. (E-step)

事实上这部分就是ELBO

- Evaluate the Posterior assignment probability $p^{(\ell)}(j|t)$ → substitute for $\delta(j|t)$

Step ℓ (M-step)

$$\hat{n}(j) = \sum_{t=1}^{n} p^{(\ell)}(j|t)$$

$\delta(j|t)$ ↓ hidden variable

- Update the parameters $\underline{\theta}^{(j+1)}$.

① $\hat{p}^{(\ell+1)}(j) = \frac{\hat{n}(j)}{n}$ ② $\hat{\mu}_j^{(\ell+1)} = \frac{1}{\hat{n}(j)} \sum_{t=1}^{n} p^{(\ell)}(j|t)\, x_t$

③ $\hat{\Sigma}_j^{(\ell+1)} = \frac{1}{\hat{n}(j)} \sum_{t=1}^{n} p^{(\ell)}(j|t)\,(x_t - \hat{\mu}_j^{(\ell+1)})(x_t - \hat{\mu}_j^{(\ell+1)})^T$

i.e., Replaced $\delta(j|t)$ with soft assignment $p^{(\ell)}(j|t)$

Thm: (Main Thm)

Let $\ell(X; \theta^{(\ell)}) = \sum_{t=1}^{n} \log P(x_t; \theta^{(\ell)})$, the incomplete likelihood.

⇓

$$\ell(X; \theta^{(\ell+1)}) \geq \ell(X; \theta^{(\ell)}) \qquad \forall \ell \in \mathbb{N}_0$$

⇓ Non-decreasing