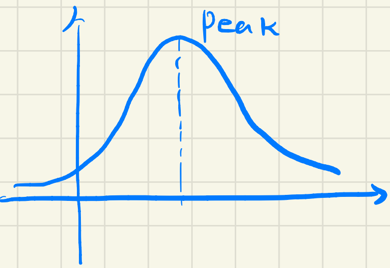
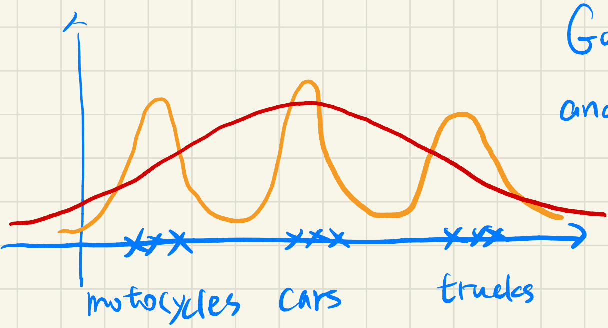
Lecture 4 Mixture Models and Clustering



so far, we only have looked at probability distribution with respect to one mode (peak)

what if it is more complicated?



Gaussian doesn't fit the det well and doesn't tell the whole story.

Gaussian Mixture Model

two r.v.

1) Z = hidden state (vehicle type) ← with K states

e.g. $Z \in \{$scooter, car, truck$\}$
1 , 2 , 3

$P(Z=j) = \pi_j$, $\sum_j \pi_j = 1$

prior probability of a type of vehicle occurring

2) X = observation

observation model conditioned on $Z=j$ (weight)

$$p(x \mid z=j) = N(x \mid \mu_j, \sigma_j^2)$$

each vehicle type has its own distribution of weight

Generative process

1) sample z (vehicle type)

2) sample $x \mid z$ (weight given type)

Note: We never see z! only see x!

Distribution of x:

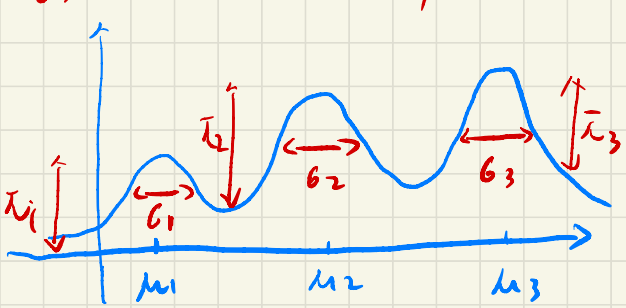
$$p(x) = \sum_j p(x, z=j)$$

$$= \sum_j p(x \mid z=j)\, p(z=j)$$

$$p(x) = \sum_j \pi_j\, p(x \mid z=j) \leftarrow \text{"weight sum of Gausian dist"}$$

π_j: component weight (prior)

$p(x \mid z=j)$: mixture components



Clustering

Given data $D = \{x_1, \ldots x_n\}$, $x_i \in \mathbb{R}$, estimate a GMM with K components (# of clusters, Gaussians)

1) Gaussian component μ_j, σ_j^2

μ_j : location of cluster

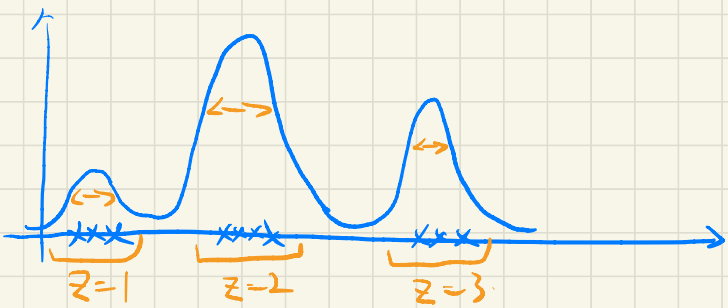
σ_j^2 : spread of cluster.

2) Component weight π_j

π_j : probability / size of cluster.

3) Cluster assignments z_i for each x_i

z_i : cluster membership



Antoni's hack

Data $= \{x_1, \cdots x_n\}$

Assignment variable $z_i \in \{1, \cdots, k\}$ = cluster assignment for x_i

Objective: treat z_i's as a parameter and optimize them

maximize the joint likelihood $p(x, z)$

$$(\hat{\theta}, \hat{z}) = \arg\max_{(\theta, z)} \sum_i \log p(x_i, z_i)$$

$$= \arg\max_{\theta, z} \sum_i \log p(x_i | z_i) p(z_i)$$

Indicator variable trick:

Let $z_{ij} = \begin{cases} 1 & (z_i = j) \quad (x_i \text{ is assigned to } j) \\ 0, & \text{otherwise.} \end{cases}$

$$\Rightarrow p(z_i) = \prod_{j=1}^{k} \pi_j^{z_{ij}}, \quad p(x_i | z_i) = \prod_{j=1}^{k} N(x | \mu_j, \sigma_j^2)^{z_{ij}}$$

$$\Rightarrow \hat{\theta}, \hat{z} = \arg\max \sum_i \left(\log \prod_{j=1}^{k} \pi_j^{z_{ij}} \cdot \prod_{j=1}^{k} N(x | \mu_j, \sigma_j^2)^{z_{ij}} \right)$$

$$= \arg\max \sum_i \sum_{j=1}^{k} z_{ij} \log \pi_j + z_{ij} \log N(x | \mu_j, \sigma^2)$$

variables depend on each other, so try an alternative maximization scheme.

1) Given $\theta = \{\pi_j, \mu_j, \sigma_j^2\}$, find the z_i's

- each z_i is independent of other z_i in the objective

$$\arg\max_{\{z_{ij}\}_j} \sum_j z_{ij} \log \pi_j N(x|\mu_j, \sigma_j^2)$$

only one term can be selected ($z_{ij}=1$)

$\Rightarrow$ select j with largest $\pi_j N(x|\mu_j, \sigma_j^2)$

$$z_i = \arg\max_j \pi_j N(x_i|\mu_j, \sigma_j^2)$$

2) Given z_i, find $\{\pi_j, \mu_j, \sigma_j^2\}$

$$(\pi_j, \mu_j, \sigma_j^2) = \arg\max \sum_i z_{ij} \log \pi_j + z_{ij} \log N(x_i|\mu_j, \sigma_j^2)$$

mean $\hat{\mu_j} = \arg\max_{\mu_j} \sum_i z_{ij} \left[-\frac{1}{2\sigma_j^2}(x_i - \mu_j)^2\right]$

$$\frac{\partial}{\partial \mu_j} = \sum_i z_{ij} \left[\frac{1}{\sigma_j^2}(x_i - \mu_j)\right] = 0$$

$$= \sum_i z_{ij}(x_i - \mu_j) = 0$$

$$\Rightarrow \mu_j = \frac{1}{\sum_i z_{ij}} \left(\sum_i z_{ij} x_i\right)$$ ← mean of points assigned to j

$\sum_i z_{ij}$ = # of points assigned to cluster j

$\sum_i z_{ij} x_i$ = sum of points assigned to j

variance $\hat{\sigma}_j^2 = \arg\max_{\sigma_j^2} \sum_i z_{ij} \left[\log \frac{1}{\sqrt{2\pi\sigma_j^2}} - \frac{1}{2\sigma_j^2}(x_i - \mu_j)^2 \right]$

$$\frac{\partial}{\partial \sigma_j^2} = \sum_i z_{ij} \left[\sqrt{2\pi}\sigma_j^{-1} \cdot -\frac{1}{2}(2\pi\sigma_j^2)^{-\frac{3}{2}} \cdot 2\pi + \frac{1}{2(\sigma_j^2)^2}(x_i - \mu_j)^2 \right]$$

$$= \sum_i z_{ij} \left[\sqrt{2\pi}\sigma_j^{-1} - \frac{1}{2}\cdot(2\pi)^{-\frac{1}{2}}(\sigma_j^2)^{-\frac{3}{2}} + \frac{1}{2(\sigma_j^2)^2}(x_i - \mu_j)^2 \right]$$

$$= \sum_i z_{ij} \left[-\frac{1}{2}(\sigma_j^2)^{-1} + \frac{1}{2(\sigma_j^2)^2}(x_i - \mu_j)^2 \right] = 0$$

$$\Rightarrow \sum_i z_{ij} \left[-\frac{1}{2}\sigma_j^2 + \frac{1}{2}(x_i - \mu_j)^2 \right] = 0$$

$$\Rightarrow \hat{\sigma}_j^2 = \frac{\sum_i z_{ij}(x_i - \mu_j)^2}{\sum_i z_{ij}}$$

prior: $\hat{\pi}_j = \arg\max_{\pi_j} \sum_i z_{ij} \log \pi_j + \lambda\left(\sum_{j=1}^K \pi_j - 1\right)$

$$\frac{\partial}{\partial \pi_j} = \frac{\sum_i z_{ij}}{\pi_j} + \lambda = 0$$

$$\sum_i z_{ij} = -\lambda \pi_j \qquad \Rightarrow \hat{\pi}_j = \frac{\sum_i z_{ij}}{-\lambda} = \frac{\sum_i z_{ij}}{N}$$

$$\sum_{j=1}^K \sum_i z_{ij} = -\lambda = N$$

3) Repeat (1) and (2) until converge

Notes:
- this 2 step procedure always maximizes the objective $\Rightarrow$ converges to a local maximum
- need an initial value $\{z_i\}$ or $\{\pi_j, \mu_j, \sigma_j^2\}$
- If we set $\pi_j = \frac{1}{k}$ and $\sigma_j^2 =$ constant
  $\Rightarrow$ k-means algorithm (Lloyd's algorithm)

$$z_i = \underset{j}{\operatorname{argmin}} (x_i - \mu_j)^2$$

μ_j = mean of points assigned to j

$$= \frac{1}{\sum_i z_{ij}} \sum_i z_{ij} x_i$$

- problem: not maximizing the actual $\log p(D)$!
  maximizing some surrogate $p(X, Z)$

Expectation-Maximization (EM) algorithm

Maximum likelihood estimation for models with hidden variables

X = observation r.v.

Z = hidden r.v. r.v.

$p(X, Z) = p(X|Z)\,p(Z)$, $p(X) = \sum_Z p(X|Z)\,p(Z)$

Goal: MLE

$$\theta = \arg\max_{\theta} \log p(x) = \arg\max_{\theta} \log \sum_{z} p(x|z)\,p(z)$$

key observation

- if we knew (x, z), then problem is easy

⇒ step 2 of Antoni's hack.

- guess the value of z probabilitically:

1) select Expected value of z given the model ⇒ $\hat{z}$

2) maximize $p(x, \hat{z})$ to get the new model

3) repeat 1) and 2)

Formally: EM algorithm

0) select initial model $\hat{\theta}^{(old)}$

1) E-step: $Q(\theta; \theta^{(old)}) = E_{z|x,\hat{\theta}^{(old)}}[\log p(x, z|\theta)]$

conditional expectation using the current $\hat{\theta}^{old}$

new parameter; old parameter (fixed); joint LL using θ

2) M-step: $\hat{\theta}^{new} = \arg\max_{\theta} Q(\theta; \theta^{(old)})$

3) $\hat{\theta}^{(old)} \leftarrow \hat{\theta}^{new}$, repeat 1 and 2 until convergence

$$\log p(D|\theta) = \sum_{i=1}^{n} \log p(x_i|\theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} p(x_i, z_j|\theta)$$

assume $z_j \sim Q(z_j)$

$$= \sum_{i=1}^{n} \log \sum_{j=1}^{k} Q(z_j) \frac{p(x_i, z_j|\theta)}{Q(z_j)}$$

$$\geq \sum_{i=1}^{n} \sum_{j=1}^{k} Q(z_j) \log \frac{p(x_i, z_j|\theta)}{Q(z_j)} \quad \text{(Jenson)}$$

IFF: $\frac{p(x_i, z_j|\theta)}{Q(z_j)} = C = \sum_{j=1}^{k} p(x_i, z_j|\theta)$

$$\Rightarrow Q(z_j) = \frac{p(x_i, z_j|\theta)}{\sum_{j=1}^{k} (x_i, z_j|\theta)} = p(z_j|x_i, \theta)$$

0) select initial model $\hat{\theta}^{old}$.

1) $$Q(z_j) = \frac{p(x_i, z_j|\hat{\theta}^{old})}{\sum_{j=1}^{k} (x_i, z_j|\hat{\theta}^{old})}$$

$$M(\theta^{old}) = \sum_{i=1}^{n} \sum_{j=1}^{k} Q(z_j) \log \frac{p(x_i, z_j|\theta^{old})}{Q(z_j)} = E(\theta^{old})$$

2) $\hat{\theta}^{new} = \arg\max M(\theta^{old}) \geq E(\theta^{old})$

3) repeat 1) and 2) until converages.

EM for GMMs

Joint LL: $\log p(x, z) = \sum_i \sum_j z_{ij} \log \pi_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)$

1) E-step

$$Q(\theta; \theta^{old}) = E_{z|x,\hat{\theta}^{old}}[\log p(x, z)]$$

$$= \sum_i \sum_j E_{z|x,\hat{\theta}^{old}} z_{ij} \underbrace{\log \pi_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)}_{\text{const.}}$$

$E_{z|x,\hat{\theta}^{old}} z_{ij}$ — Expectation of an Indicator. p.s1-5

$$= p(z_{ij} \mid x, \hat{\theta}^{old}) = p(z_i = j \mid x, \hat{\theta}^{old})$$

Bayes' Rule.

$$= \frac{p(x \mid z_i = j)\, p(z_i = j)}{p(x)}$$

$$= \frac{p(x_{i,v})\, p(x_{-i} \mid z_i = j)\, p(z_i = j)}{p(x_{i,v})\, p(x_{-i})}$$

independent of x_i w.r.t. oth x

$$= \frac{p(x_i \mid z_i = j)\, p(z_i = j)}{p(x_i)}$$

$$= \frac{\hat{\pi}_j \mathcal{N}(x_i \mid \hat{\mu}_j, \hat{\sigma}_j^2)}{\sum_j \hat{\pi}_j \mathcal{N}(x_i \mid \hat{\mu}_j, \hat{\sigma}_j^2)}$$

"soft assignment to cluster j using $\hat{\theta}^{old}$

$= p(z_i = j \mid x_i, \hat{\theta}^{old})$ posterior prob of $z_i | x_i$ (using $\hat{\theta}^{old}$)

2) M-step: Same as before, replace z_{ij} with $\hat{z}_{ij}$

Summary EM-GMMs

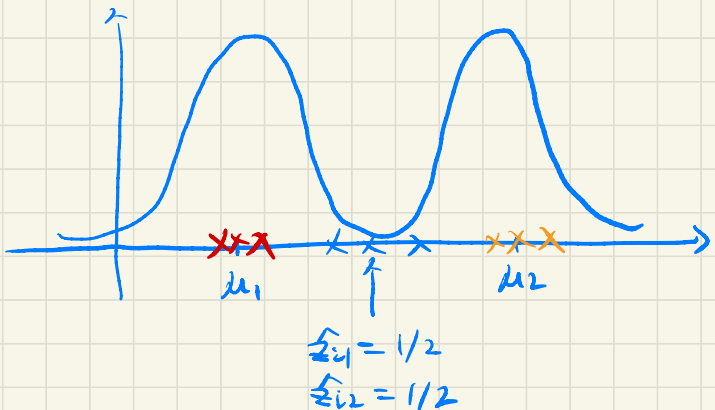
E-step: $\hat{z}_{ij} = p(z_{ij} \mid x_i, \hat{\theta}^{old}) = \frac{\pi_j N(x_i \mid \mu_j, \sigma_j^2)}{\sum_k \pi_k N(x_i \mid \mu_k, \sigma_k^2)}$ (using $\hat{\theta}^{old}$)

M-step: $\hat{\mu}_j = \frac{1}{N_j} \sum_i \hat{z}_{ij} x_i$ ← sample mean with points weighted by soft assignment z_{ij}

$N_j = \sum_i \hat{z}_{ij}$ ← weight of points assigned to j

$\hat{\sigma}_j^2 = \frac{1}{N_j} \sum_i \hat{z}_{ij} (x_i - \hat{\mu}_j)^2$

$\hat{\pi}_j = N_j / N$



Notes on EM:

1) converages: after each iteration of EM, the data LL increases → converages to local max and could be slow

2) depends on initialization

different init → different $\hat{\theta}$

pick $\hat{\theta}$ with largest LL $p(x|\hat{\theta})$

3) General framework for MLE on any model with hidden variables: Linear dynamical system

Hidden Markov model

Prob. graphical models.