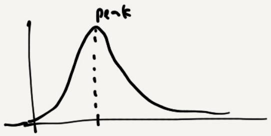
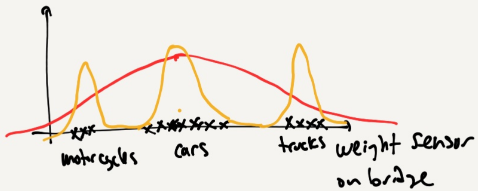
Lecture 4 - Mixture Models & Clustering



So far, we only have looked at prob. dist. w/ one mode (peak).

What if it's more complicated?



Gaussian doesn't fit the data 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. $z \in \{\underset{1}{\text{scooter}}, \underset{2}{\text{car}}, \underset{3}{\text{truck}}\}$

$p(z=j) = \pi_j$, $\sum_j \pi_j = 1$

prior probability of a type of vehicle occuring.

(2) x = observation

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

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

each vehicle type has its own distribution of weight

CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Generative process

1) sample z (vehicle type)

2) sample $x|z$ (weight given type)

Note: we never see z! only see x!

Distribution of x:

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

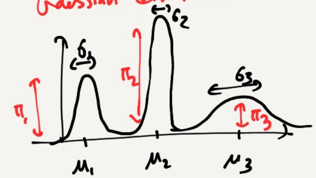
$$= \sum_j p(x|z=j)p(z=j)$$

$$\boxed{p(x) = \sum_j \pi_j \underbrace{N(x \mid \mu_j, \sigma_j^2)}}$$

"weighted sum of Gaussian distributions"

component weight (prior)

mixture components



Clustering

Given data $D = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}$, estimate a GMM w/ K components

\# of clusters (Gaussians)

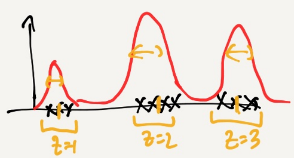
⇒ 1) Gaussian component μ_j, σ_j^2

location of cluster; spread of cluster

2) component weight π_j ← probability/size of cluster

3) cluster assignments z_i for each x_i

(cluster membership)



Antoni's hack

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

Assignment variable $z_i \in \{1, \dots, K\}$ = cluster assignment for x_i.

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

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

$$(\hat{\theta}, \hat{z}) = \underset{(\theta, z)}{\operatorname{argmax}} \sum_i \log p(x_i, z_i)$$

$$= \underset{\theta, z}{\operatorname{argmax}} \sum_i \log p(x_i | z_i) p(z_i)$$

Indicator variable trick

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

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

(z_{ij} selects the π_j & (μ_j, σ_j^2) for the x_i)

$$\Rightarrow \hat{\theta}, \hat{z} = \operatorname{argmax} \sum_i \log \prod_j \pi_j^{z_{ij}} \prod_j N(x_i | \mu_j, \sigma_j^2)^{z_{ij}}$$

$$\hat{\theta}, \hat{z} = \underset{\theta, z}{\operatorname{argmax}} \sum_i \sum_j z_{ij} \log \pi_j + z_{ij} \log N(x_i | \mu_j, \sigma_j^2)$$

Variables depend on each other, so try an alternating maximization scheme.

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

- each z_i is independent of other z_is in the objective.

$$\underset{\{z_{ij}\}_j}{\operatorname{argmax}} \underbrace{\sum_j z_{ij} \log \pi_j N(x_i | \mu_j, \sigma_j^2)}$$ ⇐ only 1 term can be selected ($z_{ij} = 1$)

⇒ select j w/ largest $\pi_j N(x_i | \mu_j, \sigma_j^2)$

$$\boxed{z_i = \underset{j}{\operatorname{argmax}} \pi_j N(x_i | \mu_j, \sigma_j^2)}$$

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

$$\begin{pmatrix} \pi_j, \mu_j \\ \sigma_j^2 \end{pmatrix} = \operatorname{argmax} \sum_i z_{ij} \log \pi_j + z_{ij} \log N(x_i | \mu_j, \sigma_j^2)$$

Mean: $\hat{\mu}_j = \underset{\mu_j}{\operatorname{argmax}} \sum_i z_{ij} \left(-\frac{1}{2\sigma_j^2} \underline{(x_i - \mu_j)^2} \right)$

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

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

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

$$\Rightarrow \boxed{\mu_j = \frac{1}{\sum_i z_{ij}} \sum_i z_{ij} x_i}$$

sum of points assigned to j.

# of points assigned to cluster j.

mean of points assigned to j.

Similarly,

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

variance of points assigned to j.

$$\boxed{\hat{\pi}_j = \frac{\sum_i z_{ij}}{N}}$$

fraction of points assigned to j

3) Repeat (1) & (2) until convergence

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 $\pi_i = \frac{1}{K}$ & $\sigma_j^2 = \text{constant}$,

$\Rightarrow$ K-means algorithm (Lloyd's algorithm)

$$\begin{cases} z_i = \underset{j}{\operatorname{argmin}} (x_i - \mu_j)^2 \\ \\ \mu_j = \text{mean of points assigned to } j \\ \quad = \frac{1}{\sum_i z_{ij}} \sum_i z_{ij} x_i \end{cases}$$

• problem: not maximizing the actual $\log p(D)$!
maximizing some surrogate $p(x, z)$.

Expectation - Maximization (EM) algorithm

(Dempster, Laird, Rubin) 1977 $\Rightarrow$ 66,000 citations on Google Scholar.

Maximum likelihood estimation for models w/ hidden variables

X = observation r.v.

Z = hidden r.v.

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

Goal: MLE
$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log p(X) = \underset{\Theta}{\operatorname{argmax}} \log \sum_Z p(X|Z)\,p(Z)$$

Key observation

- if we knew (X, Z), then the problem is easy
$\Rightarrow$ Step 2 of Antoni's hack

- guess the value of Z probabilistically:
given the model $\Rightarrow \hat{z}$
1) select Expected value of Z given the model
2) maximize $p(X, \hat{Z})$ to get the new model
3) repeat 1 & 2.

Formally: EM algorithm

0) Select initial model $\hat{\Theta}^{(old)}$

1) E-step: $Q(\Theta; \hat{\Theta}^{(old)}) = \underbrace{E_{Z|X, \hat{\Theta}^{(old)}}}_{\text{conditional expectation using the current } \hat{\Theta}^{(old)}} \Big[\overbrace{\log p(X, Z | \Theta)}^{\text{joint LL using } \Theta} \Big]$

new param

old params (fixed)

2) M-step: $\hat{\Theta}^{new} = \underset{\Theta}{\operatorname{argmax}} \; Q(\Theta; \hat{\Theta}^{(old)})$

3) $\hat{\Theta}^{(old)} \leftarrow \hat{\Theta}^{new}$, repeat 1 & 2 until convergence.

EM for GMMs

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

1) E-step

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

$$= \sum_i \sum_j \underbrace{E_{Z|X}[z_{ij}]}_{\hat{z}_{ij}} \log \pi_j N(x_i \mid \mu_j, \sigma_j^2)$$

⇐ Same form as in Antoni's hack

$$\hat{z}_{ij} = E_{Z|X,\hat{\theta}^{(old)}}[z_{ij}]$$

Expectation of an indicator P1-5

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

Bayes' Rule

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

independence of x_i w.r.t. other X

$$= \frac{\cancel{p(X_{\setminus i})}\, p(x_i \mid z_i = j)\, p(z_i = j)}{\cancel{p(X_{\setminus i})}\, p(x_i)}$$

$$p(X) = p(x_i)\, p(X_{\underbrace{\setminus i}_{\text{not } i \text{ (the other } x\text{'s)}}})$$

$$\boxed{\hat{z}_{ij} = \frac{\hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\sigma}_j^2)}{\sum_k \hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\sigma}_k^2)}}$$

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

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

← posterior prob of $z_i \mid x_i$ (using $\hat{\theta}^{old}$)

2) M-step: same as before, replace z_{ij} w/ $\hat{z}_{ij}$

Summary EM-GMM

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

M-step:

$$\hat{\mu}_j = \frac{1}{N_j} \sum_i \hat{z}_{ij} x_i$$

← sample mean w/ points weighted by soft assignment $\hat{z}_{ij}$.

$$N_j = \sum_i \hat{z}_{ij}$$

← weight of points assigned to j

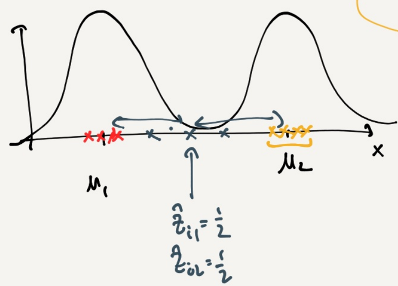
$$\hat{\sigma}_j^2 = \frac{1}{N_j} \sum_i \hat{z}_{ij} (x_i - \hat{\mu}_j)^2 \quad \leftarrow$$

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

w/ = with

w/o = without

wrt = with respect to



Notes on EM:

1) converges - after each iteration of EM, the data LL increases → converges to a local max.
(could be slow)

2) depends on initialization
different init → different $\hat{\theta}$
pick $\hat{\theta}$ w/ largest LL $p(X|\hat{\theta})$

3) general framework for MLE on any model w/ hidden variables: Linear dynamical system
Hidden Markov model
prob. graphical models ...