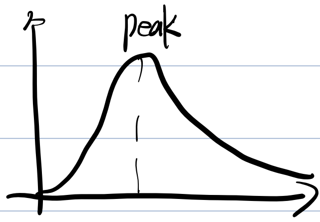
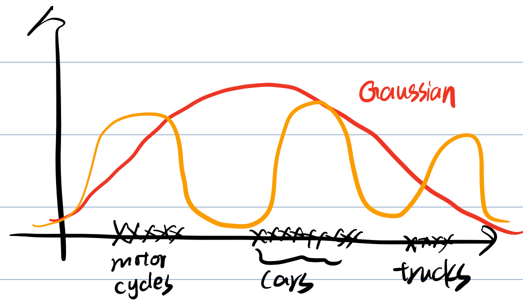


So far, we only have models has prob. dist. w/ one mode $\Rightarrow$ has ⓐ peak

What if it's more complicated?



weight sensor on bridge

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

Gaussian Mixture Model (GMM)

two r.v.

1) z = hidden state (vehicle type in example) ← with k states.

eg. $z \in \{ \underset{1}{\text{scooter}}, \underset{2}{\text{car}}, \underset{3}{\text{truck}} \}$

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

↑ prior probability of type of vehicle occurrency

2) x = observation

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

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

each vehicle type has its own distribution of weight.

Generative Process

1) sample z (vehicle type)

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

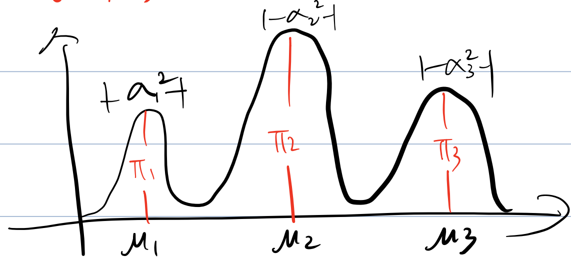
Note: we never see z! only $\underline{x}$ (observation)

Distribution of x:

$$p(x) = \sum_j p(x, z=j) \quad \text{(marginalize all } z\text{)}$$

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

$$p(x) = \sum_j \underbrace{\pi_j}_{\text{component weight (prior)}} \underbrace{\mathcal{N}(x \mid \mu_j, \sigma_j^2)}_{\text{mixture components}}$$ ← weighted sum of Gaussian distributions



Clustering ($x_i \in \mathbb{R}$)

Given data $D = \{x_1, \dots, x_n\}$, estimate a GMM w/ K components.

⇒ 1) Gaussian component μ_j, σ_j^2 ← spread of cluster
μ_j ↑ location 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 $D = \{x_1, x_2, \dots, x_n\}$

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

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

$$\hat{\theta}, \hat{z} = \underset{\theta, z}{\operatorname{argmax}} \sum_i \log \underbrace{p(x_i, z_i)}_{\text{joint pdf of } (x_i, z_i)}$$

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

Indicator variable trick

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

$$p(z_i) = \prod_{j=1}^{k} \pi_j^{z_{ij}} \quad (\text{like bernoulli distribution})$$

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

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

θ, z depend on each other.
So try an alternative maximization scheme.

(1) Given $\theta = \{\pi_j, \mu_j, \sigma_j^2\}_j$, find z_i

– each z_i is independent of others.

$$\underset{\{z_{ij}\}_j}{\operatorname{argmax}} \sum_j z_{ij} \log \pi_j N(x_i|\mu_j, \sigma_j^2)$$

$\Rightarrow$ select j w/ largest $\pi_j N(x_i|\mu_j, \sigma_j^2)$ since only one of $z_{ij} = 1$

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

(2) Given z_i, find $\theta = \{\pi_j, \mu_j, \sigma_j^2\}_j$

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

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

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

$$\Rightarrow \hat{\mu}_j = \frac{1}{\underbrace{\sum_i z_{ij}}_{\text{\# of points assigned to cluster } j}} \cdot \underbrace{\sum_i z_{ij} x_i}_{\text{sum of points assigned to } j}$$

the whole thing $\Rightarrow$ mean of points assigned to j

Similarly, $\hat{\sigma}_j^2 = \dfrac{\sum_i z_{ij}(x_i - \mu_j)^2}{\sum_i z_{ij}}$ $\leftarrow$ Variance of points assigned to j

$\hat{\pi}_j = \dfrac{\sum_i z_{ij}}{N}$ $\leftarrow$ fraction of points assigned to j

23) • repeat (1) and (2) until it converges

$\Rightarrow$ converge to a local maximum

- need an initial value $\{z_i\}_i$ or $\{\pi_j, \mu_j, \sigma_j^2\}_j$
- if we set $\pi_j = \frac{1}{K}$ & $\sigma_j^2 = \text{constant}$

$\Rightarrow$ K-means

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

- problem: not maximizing the actual $\log p(D)$!

maximizing some surrogate $p(X, Z)$

Z is r.v., but we treat it as a parameter.

Expectation - Maximization (EM) algorithm
(real MLE of above case)

(Dempster, Laird, Rubin) 1977 ⇒ 66000 citations

Maximum likelihood estimation for models
with hidden variables

X = observation r.v.

Z = hidden r.v.

$p(X, Z) = p(X|Z) \cdot p(Z)$, $p(X) = \sum_z p(x|z) p(z)$

Goal: $\theta = \arg\max_\theta \log p(x) = \arg\max_\theta \log \sum_z p(x|z) p(z)$

tricky ⇒ Σ inside log.

Key observation.

- if we know (X, Z), then the problem is easy
  ⇒ step 2 of Antoni's hack
- guess the value of Z probabilistically,
  1) select Expected value of Z given the model ⇒ $\hat{z}$
  2) maximize $p(X, \hat{z})$ to get the new model
  3) repeat 1-2

Formally

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

1) E-step: $Q(\theta; \hat{\theta}^{(old)})$

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