MLE + EM

The Expectation-Maximization Algorithm

Maximum Likelihood Parameter Estimation

One (of many) approaches to param. est.
Likelihood of (indp) observations $x_1, x_2, ..., x_n$

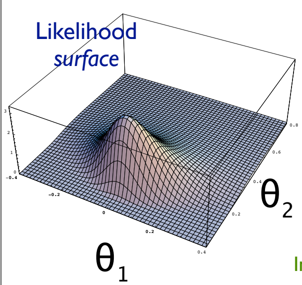
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

As a function of θ, what θ maximizes the likelihood of the data actually observed
Typical approach: $\frac{\partial}{\partial\theta} L(\vec{x} \mid \theta) = 0$ or $\frac{\partial}{\partial\theta} \log L(\vec{x} \mid \theta) = 0$

Ex 3: $x_i \sim N(\mu, \sigma^2),\ \mu, \sigma^2$ both unknown

$$\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_1} \ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} \frac{(x_i - \theta_1)}{\theta_2} = 0$$



$$\boxed{\widehat{\theta}_1 = \left(\sum_{i=1}^{n} x_i\right) / n = \overline{x}}$$

Sample mean is MLE of population mean, again

In general, a problem like this results in 2 equations in 2 unknowns.
Easy in this case, since θ_2 drops out of the $\partial/\partial\theta_1 = 0$ equation

Bias

Maximum likelihood estimation tells us how to take a bunch of i.i.d. samples $X_1, X_2, ..., X_n$
from a distribution with density $f(\cdot|\theta)$
and compute the most likely value $\hat{\Theta}$ of θ

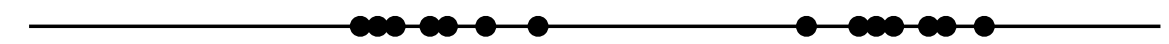
The MLE $\hat{\Theta}$ is unbiased if $E(\hat{\Theta}) = \theta$

Ex. 3, (cont.)

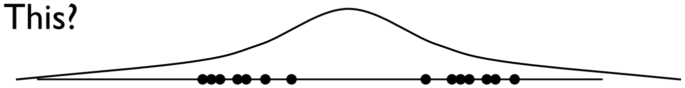
$$\begin{aligned}
\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i-\theta_1)^2}{2\theta_2} \\
\frac{\partial}{\partial\theta_2}\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} -\frac{1}{2}\frac{2\pi}{2\pi\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} = 0 \\
\boxed{\widehat{\theta}_2 = \left(\textstyle\sum_{i=1}^{n}(x_i-\widehat{\theta}_1)^2\right)/n = \bar{s}^2}
\end{aligned}$$

Sample variance is MLE of *population* variance

More Complex Example

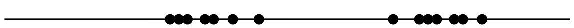


This?

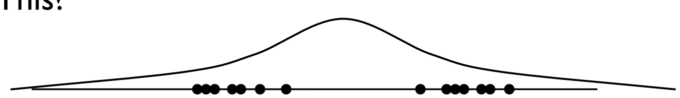


A modeling problem, not a math problem...

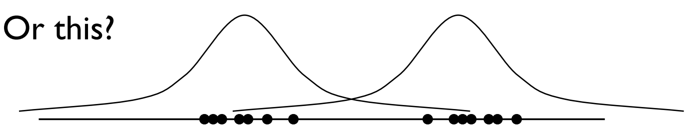
More Complex Example



This?

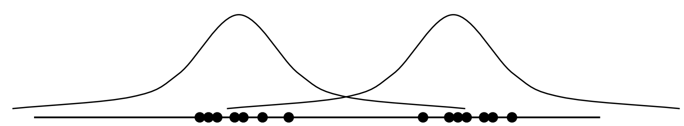


Or this?



(A modeling decision, not a math problem...,
but if the latter, what math?)

Gaussian Mixture Models / Model-based Clustering



Parameters θ

means	μ_1	μ_2
variances	σ_1^2	σ_2^2
mixing parameters	τ_1	$\tau_2 = 1 - \tau_1$

P.D.F. $\xrightarrow{\text{separately}}$ $f(x|\mu_1, \sigma_1^2)$ $\quad f(x|\mu_2, \sigma_2^2)$

Likelihood $\xrightarrow{\text{together}}$ $\boxed{\tau_1 f(x|\mu_1, \sigma_1^2) + \tau_2 f(x|\mu_2, \sigma_2^2)}$

$$L(x_1, x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$$

No closed-form max

EM

The Expectation-Maximization Algorithm

A What-If Puzzle I

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

But *what if* we knew the *hidden data?*

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

A What-If Puzzle II

Likelihood

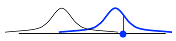
$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

What if we knew the θ, how would we estimate P[zij=1 | xi]?

Assume θ = known & fixed



$Pr(z_{i1} = 1 | x_i) = ?$

think of $Pr(x_i | ...)$ as probability of seeing a value within $\pm\delta/2$ of x_i

$$Pr(z_{i1} = 1 | x_i) = \frac{Pr(x_i | z_{i1} = 1) Pr(z_{i1} = 1)}{Pr(x_i)}$$

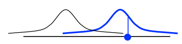
$$Pr(x_i) = Pr(x_i | z_{i1} = 1) Pr(z_{i1} = 1) + Pr(x_i | z_{i2} = 1) Pr(z_{i2} = 1)$$

$$Pr(z_{i1} = 1 | x_i) = \frac{\delta f_1(x_i | \theta) \tau_1}{\delta f_1(x_i | \theta) \tau_1 + \delta f_2(x_i | \theta) \tau_2} = \frac{f_1(x_i | \theta) \tau_1}{f_1(x_i | \theta) \tau_1 + f_2(x_i | \theta) \tau_2}$$

EM as Egg vs Chicken

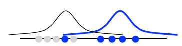
IF parameters θ known, could estimate z_{ij}

P[z_{i1}=1] vs P[z_{i2}=1]



IF z_{ij} known, could estimate parameters θ

E.g., only points in cluster 2 influence μ_2, σ_2



But we know neither; (optimistically) iterate:

E-step: calculate expected z_{ij}, given parameters

M-step: calculate "MLE" of parameters, given E(z_{ij})

Overall, a clever "hill-climbing" strategy

The EM Algorithm

Samples $x_1, \ldots, x_n$ Missing data $z_1, \ldots, z_m$

Desired parameters: $\vec{\theta} : \theta_1, \ldots, \theta_k$

Initialize: $\vec{\theta}^0$

Repeat until convergence: $t = 0, 1, \ldots$

Expectation:

Given $\vec{\theta}^t$ compute $E(z_i | x_1, \ldots, x_n) \quad \forall i$

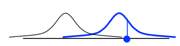
Maximization:

Set $\vec{\theta}^{t+1}$ to maximize $E(\text{LogLikelihood}(\vec{x}, \vec{z} | \vec{\theta})$

the expectation is with respect to hidden parameters $\vec{z}$

The E-step:

Find E(z_{ij}), i.e., P(z_{ij}=1)



think of $Pr(x_i|...)$ as probability of seeing a value within $\pm\delta/2$ of x_i

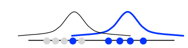
Assume θ = known & fixed

$$Pr(z_{i1} = 1 | x_i) = \frac{Pr(x_i | z_{i1} = 1) Pr(z_{i1} = 1)}{Pr(x_i)}$$

$$Pr(x_i) = Pr(x_i | z_{i1} = 1) Pr(z_{i1} = 1) + Pr(x_i | z_{i2} = 1) Pr(z_{i2} = 1)$$

Repeat for each x_i

$$Pr(z_{i1} = 1 | x_i) = \frac{f_1(x_i|\theta)\tau_1}{f_1(x_i|\theta)\tau_1 + f_2(x_i|\theta)\tau_1}$$



The E-step:

Find E(z_{ij}), i.e., P(z_{ij}=1) for each i knowing θ

$$Pr(z_{i1} = 1 | x_i) = \frac{f_1(x_i|\theta)\tau_1}{f_1(x_i|\theta)\tau_1 + f_2(x_i|\theta)\tau_2}$$

M-step:

Set $\vec{\theta}^{t+1}$ to maximize $E(\text{LogLikelihood}(\vec{x}, \vec{z} | \vec{\theta})$

Complete Data Likelihood

Recall:

$$z_{1j} \;=\; \begin{cases} 1 & \text{if } x_1 \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly,

$$L(x_1, z_{1j} \mid \theta) \;=\; \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

equal, if z_{ij} are 0/1

Formulas with "if's" are messy; can we blend more smoothly?
Yes, many possibilities. Idea 1:

$$L(x_1, z_{1j} \mid \theta) \;=\; z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2 (Better):

$$L(x_1, z_{1j} \mid \theta) \;=\; (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$