Likelihood, MLE & EM for Gaussian Mixture Clustering

Nick Duffield
Texas A&M University
Probability vs. Likelihood

• Probability: predict unknown outcomes based on known parameters:
  – $P(x \mid \theta)$

• Likelihood: estimate unknown parameters based on known outcomes:
  – $L(\theta \mid x) = p(x \mid \theta)$

• Coin-flip example:
  – $\theta$ is probability of “heads” (parameter)
  – $x = \text{HHHTTH}$ is outcome form 6 flips
Likelihood for Coin-flip Example

• Probability of outcome given parameter:
  \[ p(x = \text{HHHTTH} \mid \theta = 0.5) = 0.5^6 = 0.016 \]

• Likelihood of parameter given outcome:
  \[ L(\theta = 0.5 \mid x = \text{HHHTTH}) = p(x \mid \theta) = 0.016 \]

• Likelihood maximal when \( \theta = 0.6666... \)
• Likelihood function not a probability density
Coin Flip MLE details

• $L(\Theta|HHHTTH) = \Theta^4(1-\Theta)^2$

• Differentiate log $L$ w.r.t. $\Theta$:
  
  $4/\Theta - 2/(1-\Theta) = 0 \rightarrow \Theta = 2/3$

• Maximizer since logarithm is concave

• Intuitive result:
  
  – MLE of $H$ probability $\Theta = \text{fraction of } H \text{ in sample}$
Likelihood for Cont. Distributions

- Six samples \{-3, -2, -1, 1, 2, 3\} believed to be drawn from some Gaussian \(N(0, \sigma^2)\)

- Likelihood of \(\sigma\):
  
  \[
  L(\sigma \mid \{-3,-2,-1,1,2,3\}) = p(x = -3 \mid \sigma) \cdot p(x = -2 \mid \sigma) \cdots p(x = 3 \mid \sigma)
  \]

- Maximum likelihood:
  
  \[
  \sigma = \sqrt{\frac{(-3)^2 + (-2)^2 + (-1)^2 + 1^2 + 2^2 + 3^2}{6}} = 2.16
  \]
Likelihood for Cont. Distributions

• Six samples {-3, -2, -1, 1, 2, 3} believed to be drawn from some Gaussian N(0, $\sigma^2$)

• Likelihood of $\sigma$:

$L(\sigma | \{-3, -2, -1, 1, 2, 3\}) = p(x = -3 | \sigma) \cdot p(x = -2 | \sigma) \cdots p(x = 3 | \sigma)$

• Maximum likelihood:

$\sigma = \sqrt{\frac{(-3)^2 + (-2)^2 + (-1)^2 + 1^2 + 2^2 + 3^2}{6}} = 2.16$

• Intuitive: MLE $\sigma^2$ = sample variance
Maximum Likelihood Estimate

- Parameterized family of distributions of some r.v. $X$
- $P[X|\theta]$ for $\theta$ in some parameter set
- Likelihood $L(\theta,X) = P[X|\theta]$
- MLE = $\arg\max_\theta L(\theta,X)$
- Generally can’t compute explicitly
  - Single point $f(x_j) = \sum_{i=1}^{k} f(x_j | \mu_i, \Sigma_i)P(C_i)$
  - $P[X|\theta] = \prod_j f(x_j)$
  - Log-LLHD
    - $\log P(X|\theta) = \sum_{j=1}^{n} \log f(x_j) = \sum_{j=1}^{n} \log \sum_{i=1}^{k} f(x_j | \mu_i, \Sigma_i)P(C_i)$
- Find max by differentiation?
  - Difficult due to sum inside logarithms
Latent data

• Observations \( X = \{x_1, x_2, \ldots, x_n\} \)

• Suppose “latent data” \( Y \), unobserved, that explains the observations
  
  – E.g. if clustering with mixture of \( k \) Gaussians
    
    Latent variables \( Y = \{y_1, y_2, \ldots, y_n\} \) where each \( y_i \) in \( \{1, \ldots, k\} \) tells which mixture component \( i \) actually followed.
  
  – Parameters \( \theta = (P(C_i)) \) describe joint distribution of \( P_\theta(X,Y) \) of \( X,Y \)
    
    • \( Y \) part: \( P(C_i) \) = probability to be in component \( i \)
    
    • Distribution of \( X \) given \( Y \):
      
      – \( \mu_i, \Sigma_i \) parametrize Gaussian distribution of \( x \) in component \( i \)
  
• Problem: we don’t know the \( y_i \)

• Call \( (X,Y) \) complete data
  
  – Contrast with observed data \( X \)

• Complete data likelihood \( L(\theta \mid X,Y) = P_\theta(X,Y) \)
Expectation-Maximization

• Let \( E_\theta \) denote the expectation w/ parameter \( \theta \)
• Expectation Step: Compute Expected value of the complete data log likelihood, conditioned on observed data \( X \)
  – \( Q(\theta', \theta) = E_\theta [\log L(\theta'|X,Y) \mid X] \)
• Maximization step: given \( \theta \), find the parameters \( f(\theta) = \arg \max_{\theta'} Q(\theta', \theta) \)
EM Procedure:

- Initialize $\theta(0)$
- Iterate E and M steps sequence of iterates $\theta(n+1) = f(\theta(n))$
- Iterate until some stopping criterion is met, e.g., small successive differences
  - until $||\theta(n+1) - \theta(n)|| < \epsilon$
- Declare victory and hope $\theta$ is MLE of the original problem: $\arg\max L(\theta,X)$
3 questions

• What is the unobserved data $Y$ and how does it related to any given problem, and how do I know its distribution?

• How is this related to the MLE problem?
  – Function $Q(\theta’, \theta) = E_{\theta}[\log L(\theta’|X,Y) | X]$
  – parameters $f(\theta) = \arg\max_\theta Q(\theta’, \theta)$
  – Find parameters $\theta’$ that maximize $E$ log likelihood, so seems to have something to do with MLE

• How to find the maximizer to compute iterates $f(\theta)$
Why Expectation Maximization?

- \( P_{\theta'}(X,Y) = P_{\theta'}(X) P_{\theta'}(Y|X) \)
- \( \log P_{\theta'}(X) = \log P_{\theta'}(X,Y) - \log P_{\theta'}(Y|X) \)
- \( \log L(\theta'|X) = \log L(\theta'|X,Y) - \log P_{\theta'}(Y|X) \)
- Take expectation \( E_\theta \) conditional on \( X \)
- \( \log L(\theta'|X) = Q(\theta', \theta) + H(\theta', \theta) \)
- Where \( H(\theta', \theta) = - E_\theta [ \log P_{\theta'}(Y|X) | X ] \)
Why Expectation Maximization?

Suppose $\theta^* = \text{argmax}_\theta Q(\theta', \theta)$

$$\log L(\theta^* | X) - \log L(\theta | X)$$
$$= Q(\theta^*, \theta) - Q(\theta, \theta) + (H(\theta^*, \theta) - H(\theta, \theta))$$
$$\geq H(\theta^*, \theta) - H(\theta, \theta) \quad \text{(Gibbs inequality: this is } \geq 0)$$

If we can show this is $\geq 0$, then we can conclude that $\log L(\theta^* | X) \geq L(\theta | X)$, i.e. observed data llhd of $\theta^*$ is higher than that of $\theta$. 
Expectation-Maximization: Monotonicity

- Map $\theta \rightarrow f(\theta) = \theta^*$
- Sequence $\theta(n+1) = f(\theta(n))$.
- $N \rightarrow L(\theta(n) \mid X)$ is non-decreasing.
- Best case:
  - $L(\theta(n) \mid X)$ increases monotonically to a limit which is the ML
  - $\theta(n)$ converges to MLE $\arg\max_\theta L(\theta \mid X)$
- Caveat:
  - $L(\theta(n) \mid X)$ could converge to a *local* maximum of the likelihood
Gibbs’ Inequality

• Theorem:
  • If $p$ and $q$ are two probability distributions with $q_i = 0 \rightarrow p_i = 0$

  \[ D(p,q) = \sum_i p_i \log \left( \frac{p_i}{q_i} \right) \geq 0 \text{ with equality iff } p = q \]

• $H(\theta', \theta) = - E_\theta [ \log P_{\theta'}(Y|X) | X ]$
  \[ = \sum_y P_\theta(Y=y | X) \log P_{\theta'}(Y=y | X) \]

• $H(\theta^*, \theta) - H(\theta, \theta) =
  \sum_y P_\theta(Y=y | X) \log P_\theta(Y=y | X) - P_\theta(Y=y | X) \log P_{\theta^*}(Y=y | X)$
  \[ = \sum_y P_\theta(Y=y | X) \{ \log P_\theta(Y=y | X)/P_{\theta^*}(Y=y | X) \} \]

• = $D(P_\theta(Y|X), P_{\theta^*}(Y | X)) \geq 0$ by Gibbs
Proof of Gibbs

• Want to show $D(p,q) = \sum_i p_i \log \left( \frac{p_i}{q_i} \right) \geq 0$
• $g(x) = \log(x)$ a concave function,
  – derivative $1/x$ is decreasing
• $\log x - \log 1 \leq g'(1)(x-1) = x-1$
  – with equality only of $x = 1$.
• $-\log x = \log(1/x) < 1/x-1$ so $\log x > 1-1/x$
• $-\log p/q > 1 - q/p$
• $\sum_i p_i \log \left( \frac{p_i}{q_i} \right) \geq \sum_i p_i \left( 1 - \frac{q_i}{p_i} \right) = \sum_i p_i - q_i = 1 - 1 = 0$
  with equality only if $p_i = q_i$ for all $i$
When does EM converge to MLE?

- Certain abstract conditions sometimes difficult to check.
- Theorem:
  - Assume $L(\theta,X)$ has unique maximum and $(d/d\theta)Q(\theta, \theta')$ is continuous in $\theta$ and $\theta'$
  - Then EM sequence converges to MLE
Gaussian Mixture models

- Observed Data LLHD:
  \[ \log P(X|\theta) = \sum_{j=1}^{n} \log f(x_j) = \sum_{j=1}^{n} \log \sum_{i=1}^{k} f(x_j | \mu_i, \Sigma_i) P(C_i) \]

- Complete data:
  - Each point \( x_j \) comes with vector \( c_j = (c_{j1},...,c_{jk}) \)
  - \( c_j \) indicates which component \( j \) lies in:
    - \( c_{ji} = 1 \) if \( i \) in component \( i \) and zero otherwise.
  - \( Y = \{c_j\} \) is latent or unobserved data.
  - \( E[c_{ji}] = P[C_i | x_j] = w_{ij} \)
LLHD

- **Full data likelihood:**
  - Suppose we know the $c_{ji}$ as data, i.e. which component $i$ each point $j$ belongs to
  - For each point $x_j$
    - $f(x_j, c_j) = \prod_{i=1}^{k} (f(x_j | \mu_i, \Sigma_i)P(C_i))^{c_{ji}}$
    - $c_{ji} = 1$ for exactly 1 of the $i$:
    - Terms in product for all other $i$ are 1
  - $P(X,Y|\theta) = \prod_{j=1}^{n} f(x_j, c_j) = \prod_{j=1}^{n} \prod_{i=1}^{k} (f(x_j | \mu_i, \Sigma_i)P(C_i))^{c_{ji}}$
  - $\log P(X,Y|\theta) = \sum_{j=1}^{n} \sum_{j=1}^{n} c_{ji} \log \{f(x_j | \mu_i, \Sigma_i)P(C_i)\}$
Computation of Q

- Much nicer, no sum inside log
- \( Q(\theta', \theta) = E_\theta [\log L(\theta' | X,Y) \mid X] \)
- \( = E_\theta [ \sum_{j=1}^{n} \sum_{j=1}^{n} c_{ji} \{\log f(x_j \mid \mu'_i, \Sigma'_i) + \log P'(C_i)\} \mid X] \)
- Conditioned on \( X \), so \( x_j \) are just spectators
- \( E_\theta [c_{ji}] = P[C_i \mid x_j] = w_{ij} \)
- \( Q(\theta', \theta) = \sum_{j=1}^{n} \sum_{j=1}^{n} w_{ij} \{\log f(x_j \mid \mu'_i, \Sigma'_i) + \log P'(C_i) \} \)
Computation of maximizer (1-dim)

- $Q(\theta', \theta) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} \{ \log f(x_j | \mu'_i, \Sigma'_i) + \log P'(C_i) \}$
- $\log f(x_j | \mu'_i, \Sigma'_i) = -(x_j - \mu'_i)^2 / 2(\sigma'_i)^2 - \log \sigma'_i + \text{const.}$
- Differentiate w.r.t. $\mu'_i$:
  - $\sum_{j=1}^{n} w_{ij} (x_j - \mu'_i) = 0$
  - Occurs when $\mu'_i = \mu^*_i = \sum_{j=1}^{n} w_{ij} x_j / \sum_{j=1}^{n} w_{ij}$

- Differentiate w.r.t. $\sigma'_i$
  - $-\sum_{j=1}^{n} \{ w_{ij} (x_j - \mu'_i)^2 / (\sigma'_i)^3 - 1 / \sigma'_i \} = 0$
  - Occurs when $(\sigma'_i)^2 = (\sigma^*_i)^2 = \sum_{j=1}^{n} w_{ij} (x_j - \mu'_i)^2 / \sum_{j=1}^{n} w_{ij}$
Computation of maximizer (1-dim)

- \(Q(\theta', \theta) = \sum_{i=1}^k \sum_{j=1}^n w_{ij} \{ \log f(x_j | \mu'_i, \Sigma'_i) + \log P'(C_i) \} \)
- \(\log f(x_j | \mu'_i, \Sigma'_i) = -\frac{(x_j - \mu'_i)^2}{2(\sigma'_i)^2} - \log \sigma'_i + \text{const.} \)
- Differentiate w.r.t. \(P'(C_i)\)
  - Subject to constraint \(\sum_{i=1}^k P'(C_i) = 1\)
  - \(\sum_{j=1}^n w_{ij} / P'(C_i) = \text{constant independent of } i\)
  - Occurs when \(P'(C_i) = \sum_{j=1}^n w_{ij} / n\)
  - Maximizer

- Have recovered stated iteration of parameters
  - \(\mu_i \rightarrow \mu_i^*, \Sigma_i \rightarrow \Sigma_i^*, P(C_i) \rightarrow P^*(C_i)\)