ECEN 758 Data Mining & Analysis: Likelihood, MLE & EM for Gaussian Mixture Clustering

Nick Duffield

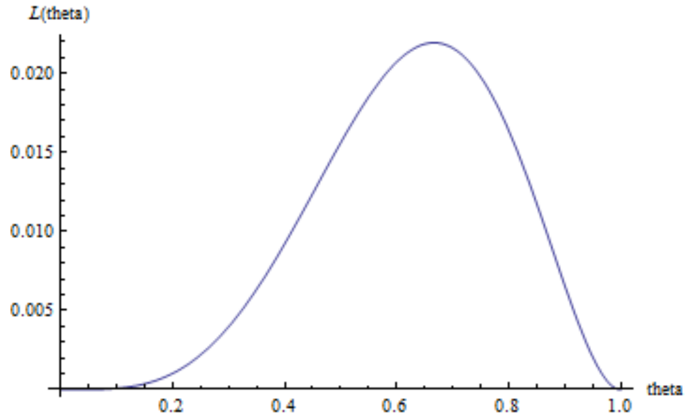
Texas A&M University

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:**
  - θ is probability of "heads" (parameter)
  - x = HHHTTH is outcome form 6 flips

Likelihood for Coin-flip Example

- **Probability of outcome given parameter:**
  - $p(x = HHHTTH \mid \theta = 0.5) = 0.5^6 = 0.016$
- **Likelihood of parameter given outcome:**
  - $L(\theta = 0.5 \mid x = HHHTTH) = p(x \mid \theta) = 0.016$



General Θ:

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

- **Likelihood *maximal* when $\theta = 0.6666...$**
- **Likelihood function not a probability density**

Coin Flip MLE details

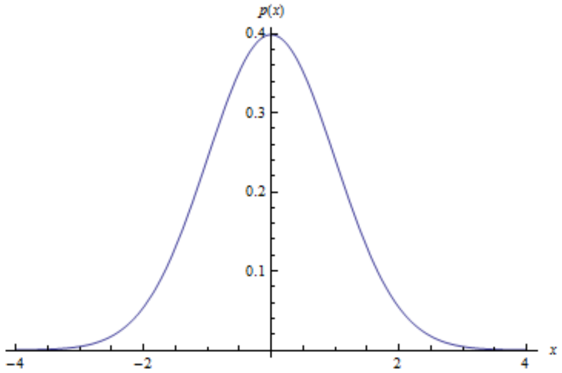
- **$L(\Theta|HHHTTH) = \Theta^4(1-\Theta)^2$**
- **$\log L(\Theta) = 4 \log \Theta + 2 \log (1-\Theta)$:**

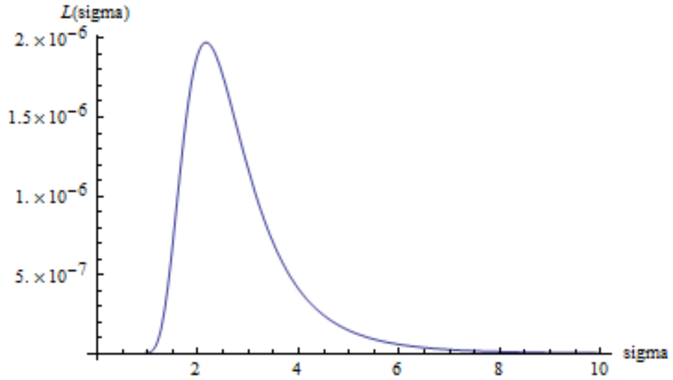
  $(d/d\Theta) \log L(\Theta) = 4/\Theta - 2/(1-\Theta)$

  Stationary point: derivative = 0 when $\Theta = 2/3$
- **Stationary point is maximizer**
  - Because logarithm is a concave function
    - Second derivative is negative
- **Intuitive result:**
  - MLE of H probability Θ = fraction of H in sample

Likelihood for Continuous Distributions

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





- **Likelihood of σ:**

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

- **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, σ^2)
- **Likelihood of σ:**

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

- **Maximum likelihood:**

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

- **Intuitive: MLE σ^2 = sample variance**

Maximum Likelihood Estimate

- **Parameterized family of distributions of some r.v. X**
- **P[X|θ] for θ in some paramter set**
- **Likelihood L(θ,X) = P[X|θ]**
- **MLE = argmax_θ L(θ,X)**
- **Clustering with normal distribution:**
  - Single point $f(x_j) = \Sigma^k_{i=1} f(x_j \mid \mu_i, \Sigma_i)P(C_i)$
  - $P[X|\theta]=\text{Prod}_j\, f(x_j)$
  - Log-LLHD
  - $\log P(X|\theta)=\Sigma^n_{j=1} \log f(x_j) = \Sigma^n_{j=1} \log \Sigma^k_{i=1} f(x_j \mid \mu_i, \Sigma_i)P(C_i)$
- **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 Normal distrbutions
    Latent variables $Y = \{y_1, y_2, \ldots, yn\}$ 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:
      - μ_i , $\Sigma^2{}_i$ parametrize Normal 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 $L(\theta|X,Y) = P_\theta(X,Y)$**

Expectation-Maximization

- **Let E_θ denote the expectation w/ parameter θ**
- **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 θ, 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 $|| \theta(n+1) - \theta(n) || < \varepsilon$**
- **Declare victory**
  - hope θ is MLE of the original problem: argmax $L(\theta, X)$

3 questions

- **What is the unobserved data Y?**
  - How does it relate to any given problem
  - How do I know its distribution?
- **How is this related to the MLE problem?**
  - Function $Q(\theta', \theta) = E_{\theta}[\log L(\theta' | X,Y) \mid X]$
  - New parameters $f(\theta) = \arg\max_{\theta'} Q(\theta', \theta)$
    - Find parameters θ’ that maximize E log likelihood
    - Seems to have something to do with MLE
- **How to find maximizer to compute iterates f(θ)**

Why Expectation Maximization?

- **$P_{\theta'}(X,Y) = P_{\theta'}(X)\, P_{\theta'}(Y|X)$**
- **$\log P_{\theta'}(X) = \log P_{\theta'}(X,Y) - \log P_{\theta'}(Y|X)$**
- **$\log L(\theta' |\, X). = \log L(\theta' |\, X,Y) - \log P_{\theta'}(Y|X)$**
  - Taking logs and rearrainging

- **Take expectation E_θ conditional on X**
  - $\log L(\theta' |\, X) = Q(\theta', \theta) + H(\theta', \theta)$
    - Definition of Q
    - Where $H(\theta', \theta) = - E_\theta [\, \log P_{\theta'} (Y|X) \mid X \,]$

Why Expectation Maximization?

- **Recap:**

  $\log L(\theta' \mid X) = Q(\theta', \theta) + H(\theta', \theta)$

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

- **$\log L(\theta^* \mid X) - \log L(\theta \mid X)$**

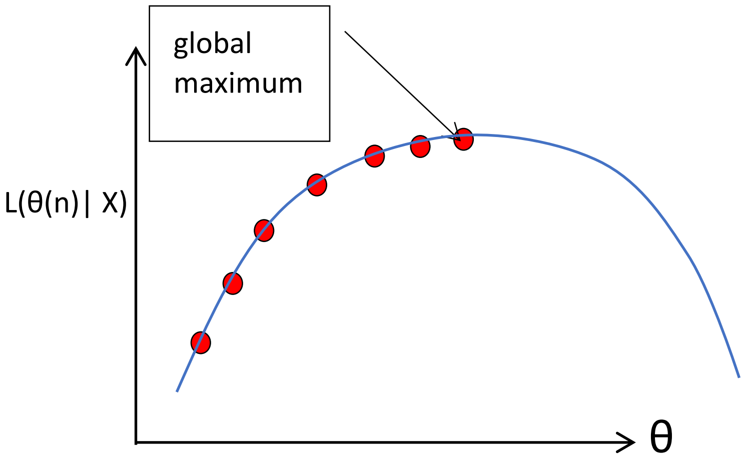
  $$= Q(\theta^*, \theta) - Q(\theta, \theta) + (H(\theta^*, \theta) - H(\theta, \theta))$$

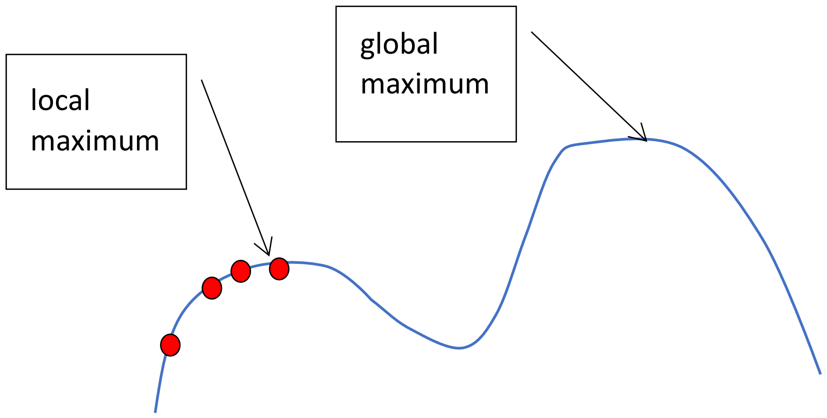
  $$\geq H(\theta^*, \theta) - H(\theta, \theta)$$

- **By Gibbs inequality (see later): $H(\theta^*, \theta) - H(\theta, \theta) \geq 0$**
- **Then conclude that $\log L(\theta^* \mid X) \geq L(\theta \mid X)$**
- **Observed data likelihood is higher for θ^* than for θ**

Expectation-Maximization: Monotonicity

- **Map $\theta \rightarrow f(\theta)=\theta^*$**
- **Sequence $\theta(n+1) = f(\theta(n))$.**
- **$L(\theta(n)|\ X)$ is non-decreasing function of n**
- **Best case:**
  - $L(\theta(n)|\ X)$ increases monotonically to a limit which is the ML
  - $\theta(n)$ converges to MLE $\text{argmax}_{\theta} L(\theta|\ X)$
- **Beware:**
  - $L(\theta(n)|\ X)$ could converge to a *local* maximum of the likelihood





Gibbs' Inequality

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

  $D(p,q) = \Sigma_i\, p_i \log (p_i / q_i) \geq 0$ with equality iff $p = q$
- **$H(\theta', \theta) = - E_\theta [\log P_{\theta'} (Y|X) \mid X] = \Sigma_y P_\theta (Y=y \mid X) \log P_{\theta'} (Y=y \mid X)$**

- **$H(\theta^*, \theta) - H(\theta, \theta) = \Sigma_y\, P_\theta (Y=y \mid X) \log P_\theta (Y=y \mid X) - P_\theta (Y=y \mid X) \log P_{\theta*} (Y=y \mid X)$**

- **$= \Sigma_y\, P_\theta (Y=y \mid X) \{\log P_\theta (Y=y \mid X)/P_{\theta*} (Y=y \mid X)\}$**

- **$= D(P_\theta\ (Y|X), P_{\theta*}\ (Y \mid X) \geq 0$ due to Gibbs Inequality**

Proof of Gibbs Inequality

- **Want to show $D(p,q) = \Sigma_i\, p_i \log (p_i / q_i) \geq 0$**
- **$g(x) = \log(x)$ 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 $\log x > 1 - 1/x$**
- **$\log p/q > 1 - q/p$**
- **$\Sigma_i\, p_i \log (p_i / q_i) \geq \Sigma_i\, p_i (1 - q_i / p_i) = \Sigma_i\, p_i - q_i = 1 - 1 = 0$ with equality only if $p_i = q_i$ for all i**

When does EM converge to MLE?

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

Gaussian Mixture models

- **Observed Data LLHD:**
  - $\log P(X|\theta)=\Sigma^n_{j=1} \log f(x_j) = \Sigma^n_{j=1} \log \Sigma^k_{i=1} f(x_j \mid \mu_i , \Sigma_i)P(C_i)$
- **Complete data:**
  - Each point x_j comes with vector $c_j = (c_{j1},\ldots,c_{jk})$
  - c_j indicates which component j lies in:
    - c_{ji} = 1 if j in component i and zero otherwise.
  - $Y = \{c_j\}$ is latent or unobserved data.
  - $E[c_{ji}] = P[C_i \mid x_j] = w_{ij}$

Full Data Likelihood

- **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
    - $f(x_{j,}\ c_j) = \Pi^k_{i=1} (f(x_j | \ \mu_i , \Sigma_i)P(C_i))^{**}c_{ji}$
    - c_{ji} = 1 for exactly 1 of the i:
    - Terms in product for all other i are 1
- $\mathbf{P(X,Y|\theta)=\Pi^n_{j=1}f(x_{j,}\ c_j)=\Pi^n_{j=1}\Pi^k_{i=1} (f(x_j| \ \mu_i , \Sigma_i)P(C_i))^{**}c_{ji}}$
- $\mathbf{\log P(X,Y|\theta) = \Sigma^n_{j=1} \Sigma^n_{j=1} c_{ji} \log \{f(x_j| \ \mu_i , \Sigma_i)P(C_i)\}}$

Computation of Q

- **Much nicer than for observed data likelihood**
  - no sum inside log
- $Q(\theta', \theta) = E_\theta [\log L(\theta' | X,Y) \mid X]$
- $= E_\theta[\ \Sigma^n_{j=1} \Sigma^k_{i=1} c_{ji}\ \{\log f(x_j | \mu'_i , \Sigma'_i) + \log P'(C_i)\} \mid X]$
  - Conditioned on X, so the x_j are just constant
- $E_\theta [c_{ji}] = P[C_i \mid x_j] = w_{ij}$

$$Q(\theta', \theta) = \Sigma^n_{j=1} \Sigma^n_{j=1}\ w_{ij}\ \{\log f(x_j | \mu'_i , \Sigma'_i) + \log P'(C_i) \}$$

Computation of maximizer (1-dim)

- **Recap:**
  - $Q(\theta', \theta) = \Sigma^{k}_{i=1} \Sigma^{n}_{j=1} w_{ij} \{\log f(x_j | \mu'_i, \Sigma'_i) + \log P'(C_i) \}$
  - $\log f(x_j | \mu'_i, \Sigma'_i) = -(x_j-\mu'_i)^2 / 2(\sigma'_i)^2 - \log \sigma'_i + \text{const.}$
- **Differentiate w.r.t. μ'_i:**
  - $\Sigma^{n}_{j=1} w_{ij} (x_j-\mu'_i) = 0$
  - Occurs when $\mu'_i = \mu_i^* = \Sigma^{n}_{j=1} w_{ij} x_j / \Sigma^{n}_{j=1} w_{ij}$
- **Differentiate w.r.t. σ'_i**
  - $-\Sigma^{n}_{j=1} \{ w_{ij} (x_j-\mu'_i)^2 / (\sigma'_i)^3 - 1/ \sigma'_i \} = 0$
  - Occurs when $(\sigma'_i)^2 = (\sigma^*_i)^2 = \Sigma^{n}_{j=1} w_{ij} (x_j-\mu'_i)^2 / \Sigma^{n}_{j=1} w_{ij}$

Computation of maximizer (1-dim)

- **Recap:**
  - $Q(\theta', \theta) = \Sigma^{k}_{i=1} \Sigma^{n}_{j=1} w_{ij} \{\log f(x_j | \mu'_i, \Sigma'_i) + \log P'(C_i)\}$
- **Differentiate w.r.t. $P'(C_i)$**

  Subject to constraint $\Sigma^{k}_{i=1} P'(C_i) = 1$
  - $\Sigma^{n}_{j=1} w_{ij} / P'(C_i)$ = constant independent of i
  - Occurs when $P'(C_i) = \Sigma^{n}_{j=1} w_{ij} / n$
  - Maximizer
- **Have recovered stated iteration of parameters**
  - $\mu_i \rightarrow \mu_i^{*}$, $\Sigma_i \rightarrow \Sigma_i^{*}$, $P(C_i) \rightarrow P^{*}(C_i)$