

Maximum Likelihood Estimate and Expectation Maximization

Ming-Chi Tsai

Overview

- Maximum Likelihood Estimate (MLE)
 - Classic example, binomial distribution
 - Multinomial distribution
 - Bayesian Network
- Expectation Maximization (EM)
 - Likelihood function
 - E-step
 - M-step

Maximum Likelihood Estimate

Data: $\{X_1, X_2, \dots, X_n\}$
 $X_i = [x_1 \dots x_p]$

Objective: Estimate the underlying 'true' probability distribution over random variables X given a set of observations drawn from the distribution.

How: Find Θ , such that $P(D | \Theta)$ is maximized, by taking derivative of $\log P(D | \Theta)$ and set it to zero.

Coin Toss (Binomial Distribution)

- Suppose you have a biased coin and wants to know the chances of seeing a head (or tail), how would you do that?
- Throw it n times, and count the number of heads and divide it by n .
- Why?
- It turned out this is the maximum likelihood estimate.
- To find a Θ that maximizes $P(D | \Theta)$ we would find Θ that maximizes $\log P(D | \Theta)$ since find Θ that maximizes $\log P(D | \Theta)$ also maximizes $P(D | \Theta)$.

MLE Derivation of Coin Toss

- Log likelihood function:

$$\begin{aligned}
 \log P(X | \theta) &= \log P(X_1, \dots, X_n | \theta) \\
 &= \log P(X_1 | \theta) \dots P(X_n | \theta) \\
 &= \log P(X_1 | \theta) + \dots + \log P(X_n | \theta) \\
 &= \sum_{i=1}^n \log P(X_i | \theta) \\
 &= \sum_{i=1}^{n_h} \log P(X_i = H | \theta) + \sum_{i=1}^{n_t} \log P(X_i = T | \theta) \\
 &= n_h \log \theta + n_t \log(1 - \theta)
 \end{aligned}$$

- Take derivative and set it to zero:

$$\begin{aligned}
 \frac{d}{d\theta} [n_h \log \theta + n_t \log(1 - \theta)] &= 0 \\
 \frac{n_h}{\theta} - \frac{n_t}{1 - \theta} &= 0 \\
 \frac{n_h}{\theta} &= \frac{n_t}{1 - \theta} \\
 n_h - n_h \theta &= n_t \theta \\
 n_h &= n_h \theta + n_t \theta \\
 \theta &= \frac{n_h}{n_h + n_t}
 \end{aligned}$$

MLE for Multinomial Distribution

- Rather than head or tail, suppose we have a dice. What is the probability of seeing a 1?
- We would do the same thing as in the binomial except now we have k Θ and we have k derivative set to 0 that we need to solve.
- Log likelihood function:

$$\begin{aligned}
 \log P(X | \theta) &= \log P(X_1, \dots, X_n | \theta) \\
 &= \log P(X_1 | \theta) \dots P(X_n | \theta) \\
 &= \sum_{i=1}^n \log P(X_i | \theta) \\
 &= \sum_{i=1}^k n_i \log \theta_i
 \end{aligned}$$

MLE for Multinomial Derivation

- Before taking the derivative, we need to add a Lagrange multiplier. Otherwise, taking derivative would not give you any solution. The Lagrange multiplier ensures sum of all Θ_i is 1.

$$l(\theta; \lambda) = \sum_{i=1}^n \log P(X_i | \theta) + \lambda \left(1 - \sum_{i=1}^k \theta_i \right)$$

- Taking the derivative with respect to each Θ_i , we have the following k equations

$$\frac{d}{d\theta_1} \sum_{i=1}^k n_i \log \theta_i + \lambda \left(1 - \sum_{i=1}^k \theta_i \right) = 0$$

$$\frac{d}{d\theta_2} \sum_{i=1}^k n_i \log \theta_i + \lambda \left(1 - \sum_{i=1}^k \theta_i \right) = 0$$

...

$$\frac{d}{d\theta_k} \sum_{i=1}^k n_i \log \theta_i + \lambda \left(1 - \sum_{i=1}^k \theta_i \right) = 0$$

Solving For Θ

- Taking the derivative and solve for Θ, λ

$$\frac{d}{d\theta_j} \sum_{i=1}^k n_i \log \theta_i + \lambda \left(\sum_{i=1}^k \theta_i - 1 \right) = 0$$

$$\frac{n_i}{\theta_i} = 0$$

$$\frac{d}{d\lambda} \sum_{i=1}^k n_i \log \theta_i + \lambda \left(\sum_{i=1}^k \theta_i - 1 \right) = 0$$

$$\sum_{i=1}^k \theta_i - 1 = 0$$

- Solving k+1 set of equations gives you:

$$\theta_i = \frac{n_i}{\lambda}$$

MLE for Bayesian Networks

- Likelihood function and log likelihood function:

$$\begin{aligned}
 L(X; \theta) &= P(x_1^1, x_2^1, \dots, x_p^1, x_1^2, \dots, x_p^2, \dots, x_1^n, \dots, x_p^n | \theta) \\
 &= \prod_{i=1}^n P(x_1^i, x_2^i, \dots, x_p^i | \theta) \\
 &= \prod_{i=1}^n \prod_{j=1}^p P(x_j^i | Pa(x_j^i), \theta) \\
 &= \prod_{j=1}^p \prod_{s \in x_j, t \in Pa(x_j)} \theta_{x_j=s, Pa(x_j)=t}^{n_{x_j=s, Pa(x_j)=t}} \\
 \log L(X; \theta) &= \sum_{j=1}^p \sum_{s \in x_j, t \in Pa(x_j)} n_{x_j=s, Pa(x_j)=t} \log(x_j = s | Pa(x_j) = t)
 \end{aligned}$$

Derivation of MLE for Bayesian Network

- Like, MLE for multinomial distribution, taking derivative with respect to $\theta_{x_j | pa(x_j)}$ will not give any solution, so we can again add the Lagrange multiplier for each state of parents.

$$L(X; \theta) = \sum_{j=1}^p \sum_{s \in x_j, t \in Pa(x_j)} n_{x_j=s, Pa(x_j)=t} \log(\theta_{x_j=s, Pa(x_j)=t}) + \sum_{x_j} \sum_{t \in Pa(x_j)} \lambda_j \left(1 - \sum_{s \in x_j} \theta_{x_j=s, Pa(x_j)=t} \right)$$

- Taking derivative and set it to zero (like the multinomial) will give us:

$$\begin{aligned}
 \theta_{x_j=s, Pa(x_j)=t} &= \frac{n_{x_j=s, Pa(x_j)=t}}{\sum_{u \in x_j} n_{x_j=u, Pa(x_j)=t}} \\
 &= \frac{n_{x_j=s, Pa(x_j)=t}}{n_{Pa(x_j)=t}}
 \end{aligned}$$

Expectation Maximization

- Goal: Maximize the likelihood function the best you can
- Problem: You don't know value of hidden variables
- Likelihood function:

$$\begin{aligned}\log L(X; \theta) &= \log \prod_{i=1}^n P(x^i | \theta) \\ &= \sum_{i=1}^n \log P(x^i | \theta) \\ &= \sum_{i=1}^n \log \sum_z P(x^i, z | \theta)\end{aligned}$$

Jensen's Inequality

$$\log \sum_z P(z) f(z) \geq \sum_z P(z) \log f(z)$$



$$\begin{aligned}\log L(X; \theta) &= \sum_{i=1}^n \log \sum_z P(x^i, z | \theta) \\ &= \sum_{i=1}^n \log \sum_z Q(z | x^i) \frac{P(x^i, z | \theta)}{Q(z | x^i)} \\ &\geq \sum_{i=1}^n \sum_z Q(z | x^i) \log \frac{P(x^i, z | \theta)}{Q(z | x^i)} \\ &= \sum_{i=1}^n \sum_z (Q(z | x^i) \log P(x^i, z | \theta)) - \sum_{i=1}^n \sum_z Q(z | x^i) \log Q(z | x^i)\end{aligned}$$

EM algorithm

- Guess some values for all the parameters.
- Iterate until convergence
 - E-Step
 - Compute probability of missing value given current choice of Θ
 - M-Step
 - Find the MLE of given current expectation of z

E-Step

- Define $Q(z | x, \Theta)$ as

$$Q(z | x^i) = P(z | x^i, \theta)$$

- Why?
- It turned out setting Q as the condition probability of z given x makes the lower bound the same as the original log likelihood.

M-Step

- Maximize the lower bound by finding the MLE given the current estimate of z

$$\theta^{(t+1)} \leftarrow \arg \max_{\theta} \sum_{i=1}^n \sum_z Q(z | x^i) \log P(z, x^i | \theta)$$