Chapter 7
Point Estimation
Expectation Maximization Algorithm

Outline

Overview

Recall θ_0 is more likely that another parameter value θ_1, then

$$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta_1|X)$$

Let write this out in terms of densities for n independent observations $X_1, \ldots, X_n$ from densities indexed by a parameter $\theta \in \Theta$.

$$\begin{aligned}
f_X(x_1|\theta_0)\cdots f_X(x_n|\theta_0) &> f_X(x_1|\theta_1)\cdots f_X(x_n|\theta_1) \\
\ln(f_X(x_1|\theta_0)\cdots f_X(x_n|\theta_0)) &> \ln(f_X(x_1|\theta_1)\cdots f_X(x_n|\theta_1)) \\
\ln f_X(x_1|\theta_0) + \cdots + \ln f_X(x_n|\theta_0) &> \ln f_X(x_1|\theta_1) + \cdots + \ln f_X(x_n|\theta_1) \\
\ln(f_X(x_1|\theta_0)/f_X(x_1|\theta_1)) + \cdots + \ln(f_X(x_n|\theta_0)/f_X(x_n|\theta_1)) &> 0 \\
\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f_X(x_i|\theta_0)}{f_X(x_i|\theta_1)} &> 0
\end{aligned}$$

Kullback-Leibler Divergence

If θ_0 is the true state of nature, then, by the strong law of large numbers, this sum converges almost surely and in mean to

$$D_{KL}(\theta_0||\theta_1) = E_{\theta_0}\left[\ln \frac{f_X(X|\theta_0)}{f_X(X|\theta_1)}\right] = \int_{\mathcal{X}} \ln \frac{f_X(x|\theta_0)}{f_X(x|\theta_1)} f_X(x|\theta_0)\ \nu(dx)$$

$D_{KL}(\theta_0||\theta_1)$ is called the Kullback-Leibler divergence.

Example. For $X \sim N(\theta, \sigma^2)$,

$$\begin{aligned} \ln \frac{f_X(x|\theta_0)}{f_X(x|\theta_1)} &= \ln \frac{(2\pi\sigma^2)^{-1/2}\exp(x-\theta_0)^2/2\sigma^2}{(2\pi\sigma^2 i)^{-1/2}\exp(x-\theta_1)^2/2\sigma^2} \\ &= \frac{1}{2}((x-\theta_0)^2 - (x-\theta_1)^2) = \frac{1}{2\sigma^2}(\theta_0-\theta_1)(2x-\theta_1-\theta_0) \end{aligned}$$

and

$$D_{KL}(\theta_0||\theta_1) = \frac{1}{2\sigma^2}(\theta_0-\theta_1)E_{\theta_0}[2X-\theta_1-\theta_0] = \frac{1}{2\sigma^2}(\theta_0-\theta_1)^2.$$

Kullback-Leibler Divergence

Example. For $X \sim Ber(\theta)$,

$$\ln \frac{f_X(x|\theta_0)}{f_X(x|\theta_1)} = \ln \frac{\theta_0^x(1-\theta_0)^{1-x}}{\theta_1^x(1-\theta_1)^{1-x}} = x \ln \frac{\theta_0}{\theta_1} + (1-x) \ln \frac{1-\theta_0}{1-\theta_1}$$

and

$$\begin{aligned} D_{KL}(\theta_0||\theta_1) &= E_{\theta_0}\left[X \ln \frac{\theta_0}{\theta_1} + (1-X) \ln \frac{1-\theta_0}{1-\theta_1}\right] \\ &= \theta_0 \ln \frac{\theta_0}{\theta_1} + (1-\theta_0) \ln \frac{1-\theta_0}{1-\theta_1} \end{aligned}$$

This shows that the Kullback-Leibler divergence is *not symmetric*, i.e., $D_{KL}(\theta_0||\theta_1) \neq D_{KL}(\theta_1||\theta_0)$

Kullback-Leibler Divergence

More generally, we can write the Kullback-Leibler divergence

$$D_{KL}(f_0||f_1) = \int_{\mathcal{X}} \ln \frac{f_0(x)}{f_1(x)} f_0(x)\ \nu(dx)$$

for densities f_0 and f_1.

Theorem. $D_{KL}(f_0||f_1) \geq 0$ and equals zero if and only if $f_0 = f_1$.

Proof. Use Jensen's inequality and the concavity of the logarithm

$$\begin{aligned} D_{KL}(f_0||f_1) &= \int_{\mathcal{X}} \ln \frac{f_0(x)}{f_1(x)}\ f_0(x)\ \nu(dx) = \int_{\mathcal{X}} -\ln \frac{f_1(x)}{f_0(x)}\ f_0(x)\ \nu(dx) \\ &\geq -\ln \int_{\mathcal{X}} \frac{f_1(x)}{f_0(x)} f_0(x)\ \nu(dx) = -\ln \int_{\mathcal{X}} f_1(x)\ \nu(dx) \\ &= -\ln 1 = 0 \end{aligned}$$

For equality to hold, we must have $\ln(f_1(x)/f_0(x))$ constant. This constant must be 0 for both f_0 and f_1 to be densities.

Connection to Fisher Information

We have a simple formula to connect the Kullback-Leibler divergence to the Fisher information, provided that we can exchange integration over the state space and differentiation over the parameter space.

$$
\begin{aligned}
D_{KL}(\theta_0||\theta) &= \int_{\mathcal{X}} \ln \frac{f_X(x|\theta_0)}{f_X(x|\theta)} \cdot f_X(x|\theta_0)\ \nu(dx) \\
&= \int_{\mathcal{X}} \ln f_X(x|\theta_0) \cdot f_X(x|\theta_0)\ \nu(dx) - \int_{\mathcal{X}} \ln f_X(x|\theta) \cdot f_X(x|\theta_0)\ \nu(dx) \\
\frac{\partial^2}{\partial\theta_i \partial\theta_j} D_{KL}(\theta_0||\theta) &= -\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f_X(x|\theta) \cdot f_X(x|\theta_0)\ \nu(dx) \\
\frac{\partial^2}{\partial\theta_i \partial\theta_j} D_{KL}(\theta_0||\theta)\big|_{\theta=\theta_0} &= -\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln f_X(x|\theta_0) \cdot f_X(x|\theta_0)\ \nu(dx) = -I_{ij}(\theta_0)
\end{aligned}
$$

the entries in the Fisher information matrix.

Kullback-Leibler Divergence

This concept has many names and uses

1. (information theory, coding theory) relative entropy
2. (machine learning, Bayesian estimation) information gain
3. (large deviations) rate function
4. (classical statistics) Fisher information, expectation-maximization algorithm

Mixtures

If your data come from two (or more) distinct population groups, then the density is a mixture

$$f_X(x|p, \theta_0, \theta_1) = pf_X(x|\theta_0) + (1-p)f_X(x|\theta_1)$$

In this case, simple maximization of the likelihood would be a difficult task. The partial derivatives do not take on a simple form.

The impetus behind the Expectaton-Maximization Algorithm is to add an unobserved or latent variable that would facilitate the likelihood maximization procedure. In the case of mixtures of two groups, we add a sequence of $Ber(p)$ random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$. If these Y_i were observed, then the group for each individual would be known.

Expectation-Maximization Algorithm

Goal Find $\hat{\Theta} = \arg\max_\theta L(\theta|\mathbf{x})$ and $L(\hat{\Theta}|\mathbf{x})$
With "missing" data $\mathbf{y}$ we introduce

$$Q(\theta, \tilde{\theta}) = E_\theta \left[\ln \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}|\tilde{\theta})}{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}|\theta)} | \mathbf{X} = \mathbf{x} \right]$$

For the algorithm, pick an initial parameter value θ_0, then iterate these steps.

- E-step: Given θ_n, compute $Q(\theta_n, \tilde{\theta})$.
- M-step: Find θ_{n+1} so that $Q(\theta_n, \theta_{n+1})$ satisfies

$$Q(\theta_n, \tilde{\theta}) = \max_{\tilde{\theta}} Q(\theta_n, \tilde{\theta}).$$

Expectation-Maximization Algorithm

Next, we show that each iteration of the EM algorithm always increases the likelihood.

$$\begin{aligned} Q(\theta_n, \theta_{n+1}) &= E_{\theta_n}\left[\ln \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\theta_{n+1})}{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\theta_n)}|\mathbf{X}=\mathbf{x}\right] \\ &= E_{\theta_n}\left[\ln \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x},\theta_{n+1}) f_{\mathbf{X}}(\mathbf{x}|\theta_{n+1})}{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x},\theta_n) f_{\mathbf{X}}(\mathbf{x}|\theta_n)}|\mathbf{X}=\mathbf{x}\right] \\ &= \ln \frac{f_{\mathbf{X}}(\mathbf{x}|\theta_{n+1})}{f_{\mathbf{X}}(\mathbf{x}|\theta_n)} + E_{\theta_n}\left[\ln \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x},\theta_{n+1})}{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x},\theta_n)}|\mathbf{X}=\mathbf{x}\right] \end{aligned}$$

Note that the first term on the right side is the logarithm of the likelihood. For the likelihood to increase with each step, we must show that this term is positive.

Expectation-Maximization Algorithm

$$\begin{aligned} \ln \frac{f_{\mathbf{X}}(\mathbf{x}|\theta_{n+1})}{f_{\mathbf{X}}(\mathbf{x}|\theta_n)} &= Q(\theta_n, \theta_{n+1}) - E_{\theta_n}\left[\ln \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x}, \theta_{n+1})}{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x}, \theta_n)} | \mathbf{X} = \mathbf{x}\right] \\ &= Q(\theta_n, \theta_{n+1}) + E_{\theta_n}\left[\ln \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x}, \theta_n)}{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{x}, \theta_{n+1})} | \mathbf{X} = \mathbf{x}\right] \end{aligned}$$

Note that $Q(\theta, \theta) = 0$ for all θ. Thus,

$$Q(\theta_n, \theta_{n+1}) = \max_{\tilde{\theta}} Q(\theta_n, \tilde{\theta}) \geq 0.$$

The second term on the right is a Kullback-Leibler divergence for a conditional density and thus must be non-negative.

Expectation-Maximization Algorithm

Example Gaussian mixture model.

The density for the mixture, $\theta = (p, \mu_0, \sigma_0, \mu_1, \sigma_1)$

$$f_X(x|\theta) = \frac{1-p}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\pi}\left(\frac{x-\mu_0}{\sigma_0}\right)^2\right) + \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\pi}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right).$$

Missing data - group membership. Let's assume that the variances take on a known common value, σ^2

Expectation-Maximization Algorithm

The likelihood

$$L(p, \mu_0, \mu_1 | \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} ((1-p) f_X(x_i|\mu_0, \sigma^2))^{1-y_i} (p f_X(x_i|\mu_1, \sigma^2))^{y_i}$$

The log-likelihood

$$\begin{aligned} \ln L(p, \mu_0, \mu_1 | \mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{n} (1-y_i) \left(\ln(1-p) - \frac{1}{2}(\ln \pi\sigma^2) - \frac{(x-\mu_0)^2}{2\sigma^2} \right) \\ &\quad + y_i \left(\ln p - \frac{1}{2}(\ln \pi\sigma^2) - \frac{(x_i-\mu_1)^2}{2\sigma^2} \right) \\ &= \sum_{i=1}^{n} (1-y_i) \left(\ln(1-p) - \frac{(x-\mu_0)^2}{2\sigma^2} \right) + y_i \left(\ln p - \frac{(x_i-\mu_1)^2}{2\sigma^2} \right) - \frac{1}{2}(\ln \pi\sigma^2) \end{aligned}$$

Expectation-Maximization Algorithm

$$\ln L(p, \mu_0, \mu_1 | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (1 - y_i) \left(\ln(1-p) - \frac{(x-\mu_0)^2}{2\sigma^2} \right) + y_i \left(\ln p - \frac{(x_i - \mu_1)^2}{2\sigma^2} \right) - \frac{1}{2} (\ln \pi\sigma^2)$$

$$\begin{aligned} Q(\theta_n, \theta) &= E_{\theta_n} \left[\ln \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y} | \theta)}{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y} | \theta_n)} | \mathbf{X} = \mathbf{x} \right] \\ &= -\sum_{i=1}^{n} P\{Y_i = 0 | X = x_i, \theta_n\} \frac{(x_i - \mu_0)^2}{2\sigma^2} + P\{Y_i = 1 | X = x_i, \theta_n\} \frac{(x_i - \mu_1)^2}{2\sigma^2} + H(\theta_n) \end{aligned}$$

For the latent variables, Y_i the odds

$$\frac{P\{Y_i = 1 | X_i = x_i, \theta^n\}}{P\{Y_i = 0 | X_i = x_i, \theta^n\}} = \frac{f_X(x_i | \mu_1^n, \sigma^2) p^n}{f_X(x_i | \mu_0^n, \sigma^2 (1 - p^n)} = \frac{p_i^{n+1}}{1 - p_i^{n+1}}$$

Expectation-Maximization Algorithm

$$\begin{aligned} Q(\theta_n, \theta) &= -\sum_{i=1}^{n} P\{Y_i = 0 | X = x_i, \theta_n\} \frac{(x_i - \mu_0)^2}{2\sigma^2} + P\{Y_i = 1 | X = x_i, \theta_n\} \frac{(x_i - \mu_1)^2}{2\sigma^2} + H(\theta_n) \\ &= -\sum_{i=1}^{n} (1 - p_i^{n+1}) \frac{(x_i - \mu_0)^2}{2\sigma^2} + p_i^{n+1} \frac{(x_i - \mu_1)^2}{2\sigma^2} + H(\theta_n) \end{aligned}$$

Differentiate with respect to μ_0 and set to 0.

$$0 = \sum_{i=1}^{n} (1 - p_i^{n+1})(x_i - \mu_0^{n+1}) \quad \mu_0^{n+1} = \frac{\sum_{i=1}^{n} (1 - p_i^{n+1}) x_i}{\sum_{i=1}^{n} (1 - p_i^{n+1})}$$

We have a similar weighted average for μ_1^{n+1},

$$\mu_1^{n+1} = \frac{\sum_{i=1}^{n} p_i^{n+1} x_i}{\sum_{i=1}^{n} p_i^{n+1}}$$

Expectation-Maximization Algorithm

Start with estimates $\theta_0 = (p_0, \mu_0^0, \mu_1^0)$

- E-step: Estimate the probability that each individual is in group 1.

$$p_i^{n+1} \propto \frac{p^n}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}\left(\frac{x - \mu_1^n}{\sigma}\right)^2.$$

- M-step: Maximize the means

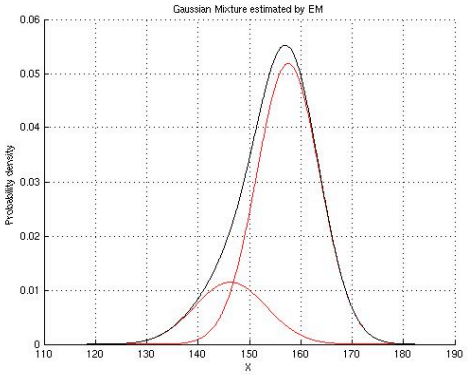
$$\mu_0^{n+1} = \frac{\sum_{i=1}^n (1 - p_i^{n+1}) x_i}{\sum_{i=1}^n (1 - p_i^{n+1})} \quad \mu_1^{n+1} = \frac{\sum_{i=1}^n p_i^{n+1} x_i}{\sum_{i=1}^n p_i^{n+1}} \quad p^{n+1} = \frac{1}{n}\sum_{i=1}^n p_i^n.$$

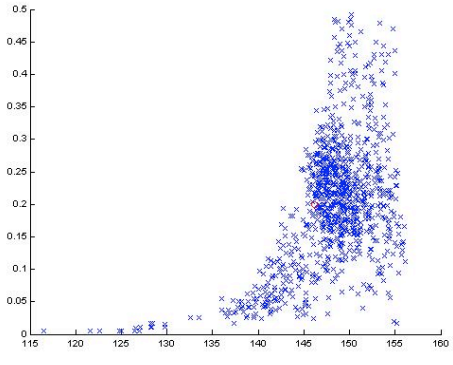
Expectation-Maximization Algorithm

Flores

- Island - 400km east-west, 80km north-south
- Site of *Homo floresiensis*, is a pygmy archaic human
- Pygmies near Rampasasa
- Two villages that were sampled trace their descent matrilineally
- Do they have living descendants?

Expectation-Maximization Algorithm





So the EM algorithm divides the population on Flores into two groups as anticipated. However, the EM algorithm finds a way to divide even homogeneous populations with the same overall mean and standard deviation.