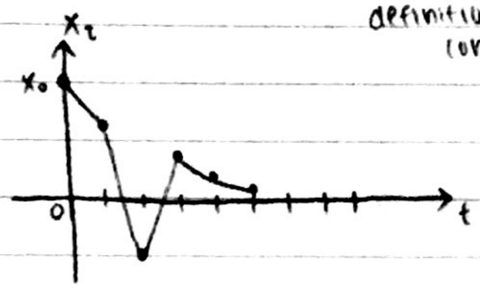
Lecture 18:

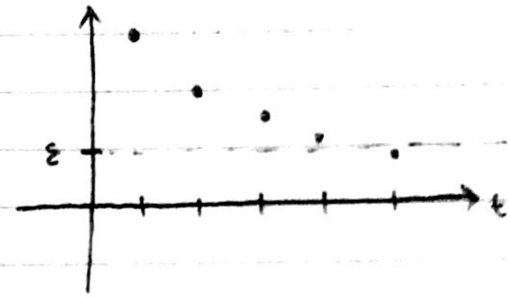
5/4/17

Newton-Raphson

$f(x) = 0$ solve for x. Given ε.

1. Guess solution is x_0
2. Calculate $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ } iteraterate step is an Iteration algorithm
3. Repeat step 2 until $\underbrace{|x_{t+1} - x_t| \le \varepsilon}_{\text{definition of convergence}}$



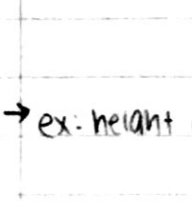


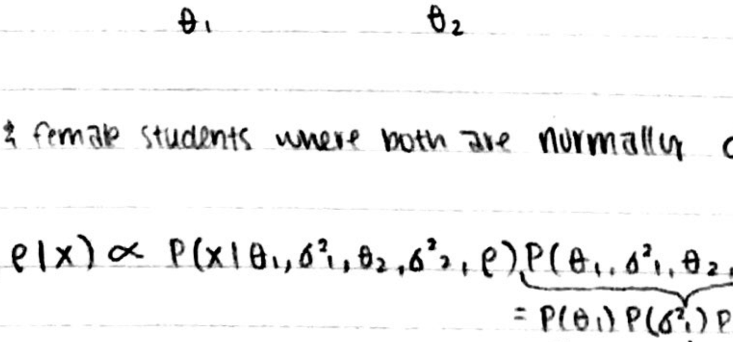
Consider the mixture model

$$X_1, \dots, X_n \mid \vec{\theta}_1, \dots, \vec{\theta}_M, \gamma_1, \dots, \gamma_M \overset{iid}{\sim} \sum_{m=1}^{M} \gamma_m P_m(\vec{\theta}_m) \text{ s.t. } \gamma_1 + \gamma_2 + \dots + \gamma_m = 1$$

eg.

$$X_1, \dots, X_n \mid \underbrace{\theta_1, \sigma_1^2}_{\vec{\theta}_1}, \underbrace{\theta_2, \sigma_2^2}_{\vec{\theta}_2} \underbrace{\rho}_{\gamma_1 \gamma_2} \overset{iid}{\sim} \underset{\gamma_1}{\rho} \underbrace{N(\theta_1, \sigma_1^2)}_{P_1(\vec{\theta}_1)} + \underbrace{(1-\rho)}_{\gamma_2} \underbrace{N(\theta_2, \sigma_2^2)}_{P_2(\vec{\theta}_2)}$$





→ ex: height of male & female students where both are normally distributed.

$$P(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid x) \propto P(x \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho) \underbrace{P(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)}$$

$$= \underset{\propto 1}{P(\theta_1)} \underset{\propto \frac{1}{\sigma_1^2}}{P(\sigma_1^2)} \underset{\propto 1}{P(\theta_2)} \underset{\propto \frac{1}{\sigma_2^2}}{P(\sigma_2^2)} \underset{\propto 1}{P(\rho)}$$

$$= \left(\prod_{i=1}^{n} \rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i - \theta_1)^2} + (1-\rho) \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x_i - \theta_2)^2} \right) \underbrace{\frac{1}{\sigma_1^2} \frac{1}{\sigma_2^2}}_{\text{prior}} = K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid x)$$

How to get inference?

Grid Search

$\hat{G}_{\theta_1} = \langle \cdots \rangle, \hat{G}_{\sigma^2_1} = \langle \cdots \rangle, \ldots$

$\Rightarrow$ inaccurate in large M (ie: high dimension).

↳ Δ is large and the grid will be spread out and lose accuracy.

What if we know which component each X_i belong to?

Define: $I_1 := \mathbb{1}_{X_1 \text{ is in } m=1}$

$I_2 := \mathbb{1}_{X_2} \ldots$

$\vdots$

$I_n := \mathbb{1}_{X_n} \ldots$

"Latent variable / information"

↳ the I_i's are unobserved but important

Let $I = \{I_1, \ldots, I_n\}$

Recall

~~$f(z) = \int f(z,y)dy = \int f(z|y)f(y)dy$~~

$$\underbrace{P(X|\theta) = \int P(X, I|\theta)}_{\text{"data augmentation"}} dI = \int P(X|I,\theta)P(I|\theta)dI$$

"data augmentation" - adding more data & averaging

$$P(\theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho|X) \propto \int P(X|I, \theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho)P(I|\theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho)P(\theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho)dI$$

$$= K(\theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho|X)$$

$$= \int \underbrace{K(\theta_1, \sigma^2_1, \theta_2, \sigma^2_2, \rho|X, I)}_{\text{body of integral}} dI$$

Modest Goal:

Get $\hat{\theta}_{MAP} = \text{argmax}\{K(\theta|X)\}$

Expectation-Maximization Algorithm (1977)

Step 1: Guess $\hat{\theta}_{MAP} = \theta_0$ to start

expectation step → Step 2: Compute $I_0 = E[I|X, \theta = \theta_0]$

maximization step → Step 3: Consider likelihood $\mathcal{L}(\theta; I = I_0, X) = \underbrace{K(\theta|I = I_0, X)}_{\text{body of integral}}$

and find $\hat{\theta}_1 = \text{argmax}\{\mathcal{L}(\theta; I, X)\}$ ie: the MLE procedure

Step 4: Repeat steps 2 & 3 until $\|\theta_{t+1} - \theta_t\| < \varepsilon$, where ε is the predefined tolerance level.

E-M Incrematation for our 2-normal mixture:

Step 1: Initialize

$$\theta_{1,0} = 0$$
$$\sigma^2_{1,0} = 1$$
$$\theta_{2,0} = 0$$
$$\sigma^2_{2,0} = 1$$
$$\rho = 0.5$$

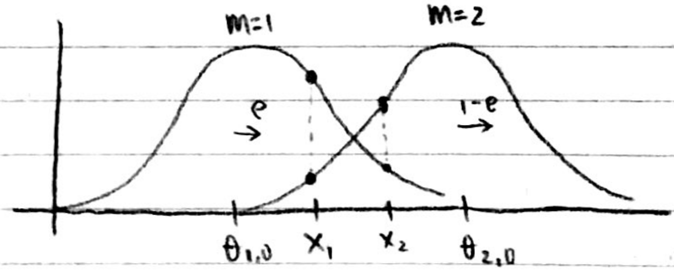
$$Q := \mathbb{1}_A \sim \text{Bern}(P(A))$$
$$E[Q] = P(A)$$

$$\rightarrow I_1 \sim \text{Bern}(P(I_1 = 1 | \cdots))$$

Step 2: $I_{1,0} = E[I_1 | X, \theta_1 = \theta_{1,0}, \sigma^2_1 = \sigma^2_{1,0}, \theta_2 = \theta_{2,0}, \sigma^2_2 = \sigma^2_{2,0}, \rho = \rho_0]$

$$= P(I_1 = 1 | X, \dots) = \frac{P(X | I_1 = 1, \dots) P(I_1 = 1 | \dots)}{\underbrace{P(X | \dots)}_{P(X|I_1=1,\dots)\cdot P(I_1=1|\dots) + P(X|I_1=0,\dots)\cdot P(I_1=0|\dots)}}$$ => Bayes Theorem

$$= \frac{\rho \frac{1}{\sqrt{2\pi\sigma^2_{1,0}}} e^{-\frac{1}{2\sigma^2_{1,0}}(x_1 - \theta_{1,0})^2}}{\rho \frac{1}{\sqrt{2\pi\sigma^2_{1,0}}} e^{-\frac{1}{2\sigma^2_{1,0}}(x_1 - \theta_{1,0})^2} + (1-\rho)\frac{1}{\sqrt{2\pi\sigma^2_{2,0}}} e^{-\frac{1}{2\sigma^2_{2,0}}(x_1 - \theta_{2,0})^2}}$$



$$I_{2,0} = E[I_2 | X_2 \cdots]$$
$$\vdots$$
$$I_{3,0} = E[I_3 | X_3, \dots]$$
$$\vdots$$
$$I_{n,0} = E[I_n | X_n, \dots]$$

Step 3: $\mathcal{L}(\theta_1, \sigma_1^2, \theta_2, \sigma^2_2, \rho; I, X) = P(X | I, \theta) P(I | \theta) P(\theta)$

$$= \left(\prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2_1}} e^{-\frac{1}{2\sigma^2_1}(x_i - \theta_1)^2} \right)^{I_i} \left(\frac{1}{\sqrt{2\pi\sigma^2_2}} e^{-\frac{1}{2\sigma^2_2}(x_i - \theta_2)^2} \right)^{1 - I_i} \right) \cdot$$

$$\left(\prod_{i=1}^{n} \rho^{I_i} (1 - \rho)^{1 - I_i} \right) \cdot \left((\sigma^2_1)^{-1} (\sigma^2_2)^{-1} \right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n (\sigma_1^2)^{-1}(\sigma_2^2)^{-1} (\sigma_1^2)^{-\frac{1}{2}\sum I_i} (\sigma_2^2)^{-\frac{1}{2}\sum 1-I_i} e^{-\frac{1}{2\sigma_1^2}\sum I_i(x_i-\theta_1)^2 - \frac{1}{2\sigma_2^2}\sum(1-I_i)(x_i-\theta)^2}$$

$$\cdot \rho^{\sum I_i} (1-\rho)^{\sum 1-I_i}$$

$$\left(\frac{\sum I_i x_i^2}{2\sigma_1^2} - \frac{\theta_1 \sum x_i I_i}{\sigma_1^2} + \frac{\theta_1^2 \sum I_i}{2\sigma_1^2}\right) \left(\frac{\sum(1-I_i)x_i^2}{2\sigma_2^2} - \frac{\theta_2 \sum x(1-I_i)}{\sigma_2^2} + \frac{\theta_2^2 \sum 1-I_i}{2\sigma_2^2}\right)$$

$$= \ell(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho; I, x)$$

$$= n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - (1+\tfrac{1}{2}\sum I_i)\ln(\sigma_1^2) - (1+\tfrac{1}{2}\sum 1-I_i)\ln(\sigma_2^2) - \overbrace{\frac{1}{2\sigma_1^2}\sum I(x_i-\theta_1)^2} - \overbrace{\frac{1}{2\sigma_2^2}\sum(1-I_i)(x_i-\theta_2)^2}$$

$$+(\sum I_i)\ln\rho + (\sum 1+I_i)\ln(1-\rho)$$

Get $\hat{\theta}_1$ by $\frac{\partial}{\partial\theta_1}[\ell] \overset{set}{=} 0$

$$\frac{\sum x_i I_i}{\cancel{\sigma_1^2}} - \frac{\cancel{2}\theta_1 \sum I_i}{\cancel{2}\cancel{\sigma_1^2}} = 0$$

$$\hat{\theta}_1 = \frac{\sum x_i I_i}{\sum I_i} \quad \bar{x}_{mix\,1}$$

$$\hat{\theta}_2 = \frac{\sum x_i (1-I_i)}{\sum(1-I_i)} \quad \bar{x}_{mix\,2}$$

Get $\hat{\sigma}_1^2$,

$$\frac{\partial}{\partial\sigma_1^2}[\ell] = -\frac{1+\frac{1}{2}\sum I_i}{\cancel{\sigma_1^2}} + \frac{1}{2(\sigma_1^2)^{\cancel{2}}}\sum I_i(x_i-\theta_1)^2 = 0$$

$$= 1 + \tfrac{1}{2}\sum I_i = \frac{1}{2\sigma_1^2}\sum I_i(x_i-\theta_1)^2$$

$$= 2 + \sum I_i = \frac{1}{\sigma_1^2}\sum I_i(x_i-\theta_1)^2 \Rightarrow \hat{\sigma}_1^2 = \frac{\sum I_i(x_i-\theta)^2}{2+\sum I_i}$$

likewise $\hat{\sigma}_2^2 = \frac{\sum(1-I_i)(x_i-\theta)^2}{2+\sum(1-I_i)}$

$\hat{\rho}$ get $\frac{\partial}{\partial\rho}[\ell] \overset{set}{=} 0$

$$= \frac{\sum I_i}{\rho} - \frac{\sum 1-I_i}{1-\rho} = 0$$

$$= \frac{\sum I_i}{\rho} = \frac{\sum 1-I_i}{1-\rho}$$

$$= \sum I_i - \rho\cancel{\sum I_i} = \rho^n - \rho^{\cancel{\sum I_i}}$$

$$= \hat{\rho} = \frac{\sum I_i}{n}$$