Lecture 6.7: Maximum Likelihood, Model Selection, Bayesian Learning

General EM algorithm:

1. Initialize θ^{old}
2. E step: $P(z|x,\theta) \Rightarrow$ guess the value of z
3. M step: $\theta \leftarrow \arg\max_{\theta} Q(\theta, \theta^{old})$

$$Q(\theta,\theta^{old}) \leftarrow \sum_{z} P(z|x,\theta^{old}) \cdot \ln P(x,z|\theta) \Rightarrow \text{discrete}$$

or $$Q(\theta,\theta^{old}) \leftarrow \int_{z} P(z|x,\theta^{old}) \cdot \ln P(x,z|\theta) \cdot dz \Rightarrow \text{continuous}$$

$\Rightarrow$ MLE to optimize parameters.

4. Repeat 2, 3 until converged.

MLE and MAP:

Maximum Likelihood Estimator: means model params.

$$\theta \leftarrow \arg\max_{\theta} P(X|\theta)$$

Maximum A Posterior:

$$\theta \leftarrow \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} \frac{P(X|\theta)\cdot P(\theta)}{P(X)}$$

$$\leftarrow \propto \arg\max_{\theta} P(X|\theta)\cdot P(\theta)$$

正比于

Model Selection:

Step 1: For each model k, compute the MLE:

$$\hat{\theta}_{ML}(k) = \arg\max_{\theta} \log P(X|\theta, k), \quad k = 1, 2, \cdots m$$

Step 2:

$$k^* = \arg\max_{k} \underbrace{J\left(\hat{\theta}_{ML}(k)\right)}_{\text{Information criterion}}$$

$$J\left(\hat{\theta}_{ML}(k)\right) = \begin{cases} \text{AIC:} & \underbrace{\ln P(X|\hat{\theta}(k))}_{\text{Maximum Likelihood}} + \underbrace{d_k}_{\#\text{params}} \\ \text{BIC:} & \ln P(X|\hat{\theta}(k)) + \frac{1}{2} \cdot d_k \cdot \underbrace{\ln N}_{\#\text{samples}} \end{cases}$$

Bayesian Model Selection: VBEM

$$\text{MAP form} \Leftarrow \boxed{P(m|y)} = \boxed{\frac{P(y|m) \cdot P(m)}{P(y)}} \Rightarrow \text{Bayesian form}$$

$P(m|y)$: model class m, data set y

$$\underbrace{P(y|m)}_{\text{Bayesian Evidence}} = \int_{\theta_m} P(y|\theta_m, m) \cdot P(\theta_m|m) \cdot d\theta_m$$

$P(y|\theta_m, m)$ and $P(\theta_m|m)$: variational form

E-step: compute $P(z|x,\theta_m) \Rightarrow$ as general EM

$$P(z|x,\theta_m) \propto e^{-E[\log p(x,y,\theta)]}$$

M-step: $\theta_m \propto$

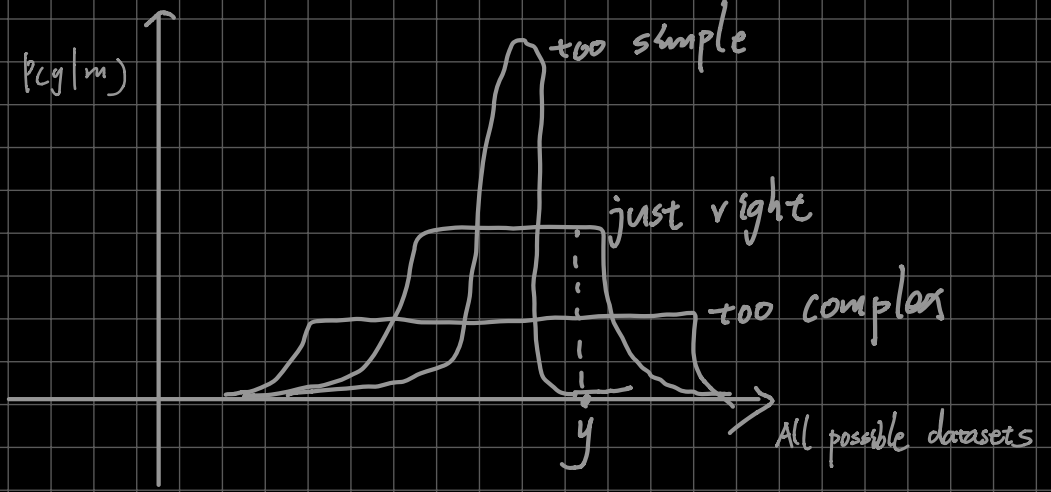
$$e^{\left[\int_z P(z|x,\theta_m)\cdot \ln p(x,z|\theta)\, dz\right]}$$

general EM: $\theta_m \leftarrow \arg\max_\theta Q(\theta,\theta^{old})$

Change $\arg\max_\theta\{\}$ to $e^{\{\}}$

△ Use EM algorithm to optimize $P(y|\theta)$.

Occam's Razor in Bayesian Model Selection:



⇒ [m is too simple: can't generate y

m is too complex: can generate y, but P(y|m) is too low.

△ Bayesian inference automatically implements Occam's Razor principle.

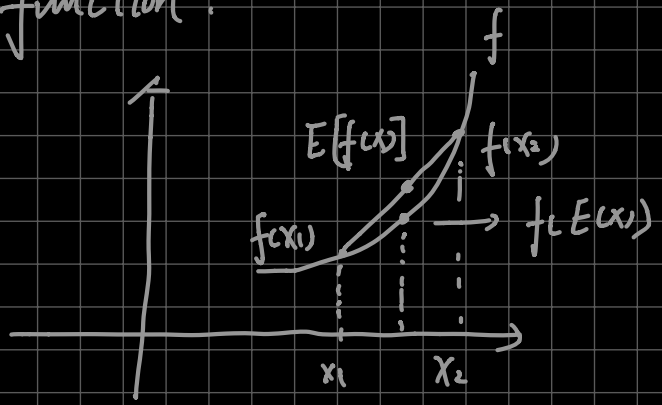
EM algorithm derivation

—— Standford Ng

sometimes called bootstrap.

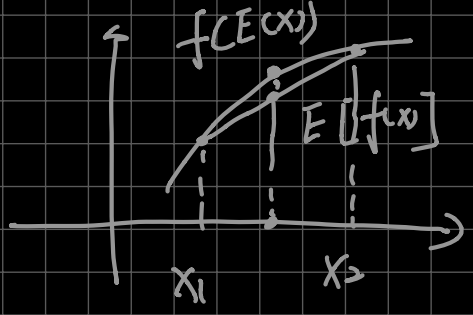
Jensen's inequality:

Convex function:



$$f(E(x)) \le E[f(x)]$$

Concave function:



$$E[f(x)] \le f(E(x))$$

EM process:

Have model for $P(x, z|\theta)$, only observe $x = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$.

$$l(\theta) = \sum_{i=1}^{m} \log P(x^{(i)}|\theta)$$

$$= \sum_{i=1}^{m} \cdot \log \sum_{k=1}^{k} P(x^{(i)}, z_k^{(i)}|\theta)$$

$$= \sum_{i=1}^{m} \cdot \log \sum_{k=1}^{k} Q(z_k^{(i)}) \left[\frac{P(x^{(i)}, z_k^{(i)}|\theta)}{Q(z_k^{(i)})}\right], \text{ where } \sum_{k=1}^{k} Q(z_k^{(i)}) = 1$$

$$\max_\theta \{l(\theta)\} \rightarrow \sum_{i=1}^{m} \cdot \log \sum_{k=1}^{k} Q(z_k^{(i)}) \left[\frac{P(x^{(i)}, z_k^{(i)} | \theta)}{Q(z_k^{(i)})} \right]$$

$$= \sum_{i=1}^{m} \log E \left[\frac{P(x^{(i)}, z_k^{(i)} | \theta)}{Q(z_k^{(i)})} \right]$$

Jenson's inequality in concave form $\Leftarrow (\geq)$

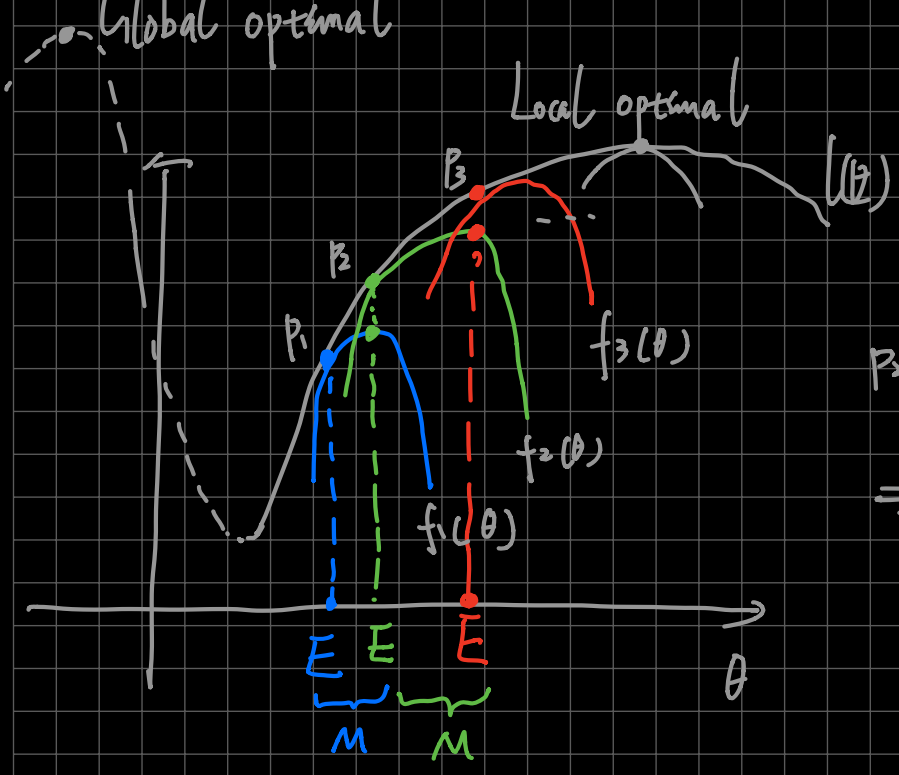
$$\sum_{i=1}^{m} E_{z^{(i)} \sim Q} \left[\log \left(\frac{P(x^{(i)}, z_k^{(i)} | \theta)}{Q(z_k^{(i)})} \right) \right]$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{k} \cdot Q(z_k^{(i)}) \cdot \log \frac{P(x^{(i)}, z_k^{(i)} | \theta)}{Q(z_k^{(i)})}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{k} Q(z_k^{(i)}) \left(\log \frac{P(z_k^{(i)} | x^{(i)}, \theta) \cdot P(x^{(i)} | \theta)}{P(z_k^{(i)} | x^{(i)}, \theta)} \right) = \sum_{i=1}^{m} \sum_{k=1}^{k} Q(z_k^{(i)}) \log P(x^{(i)} | \theta)$$

Lower bound, z, x is fixed so it's a function of $\theta \Rightarrow f(\theta)$

Graphical Explanation:



E-step: Draw curve f_1 that in point P_1 Jenson inequality satisfies.

M-step: Find the optimal point P_2 of curve f_1 drew by E-step.

$\Rightarrow$ Repeat EM until it converged

(Point find by M-step is the same as E-step)

△ E-step 即让 Jensen's inequality 相等. 因此这也是为什么 E-step 直观的解释为缩小 lower bound function $f(\theta)$ 与 $l(\theta)$ 之间的 gap.

M-step 即 optimize θ 使得上述不等式左边变大，这也是为什么 M-step 直观的解释为拔高 lower bound function $f(\theta)$ 的值。

Classification Model (分类问题)

- Unsupervised (clustering) 聚类问题: GMM, K-means, Hierarchical...
- Supervised
  - Discriminative: LR, NN, DL...
    ⇒ try to find decision boundary
  - Generative: GDA, ...
    ⇒ capture distribution of each class