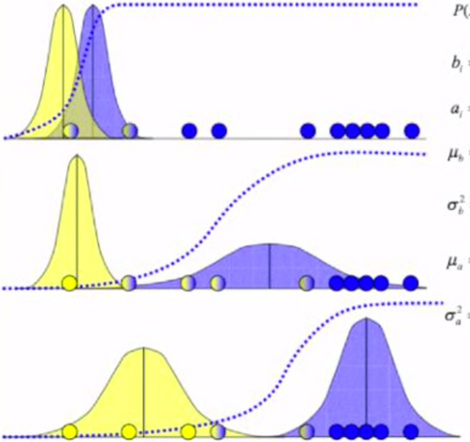
Expectation-Maximization Algorithm

EM: 1-d example



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i-\mu_b)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1x_1 + b_2x_2 + \ldots + b_nx_{n_b}}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1-\mu_1)^2 + \ldots + b_n(x_n-\mu_n)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1x_1 + a_2x_2 + \ldots + a_nx_{n_b}}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1-\mu_1)^2 + \ldots + a_n(x_n-\mu_n)^2}{a_1 + a_2 + \ldots + a_n}$$

could also estimate priors:

$P(b) = (b_1 + b_2 + \ldots b_n) / n$

$P(a) = 1 - P(b)$

Expectation-Maximization Algorithm

A general technique for finding maximum likelihood estimators in latent variable models is the expectation-maximization (EM) algorithm.

- E-Step
  Estimate the missing variables in the dataset.
  Calculate the expectation of complete-data log-likelihood:

  $$Q(\theta|\theta^{(t)}) := E[\log P(y_{obs}, y_{mis}|\theta)|y_{obs}, \theta^{(t)}]$$

- M-Step
  Maximize the parameters of the model in the presence of the data. Find $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$

  $$\theta^{(t+1)} := \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(t)})$$

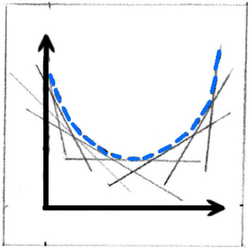
- Iterate the above 2 steps until convergence.

Why EM works?

- Lemma 1: Jensen's inequality
- Proposition 1: Ascent property of EM
- Theorem 1: Convergence property of EM

Convex function

Upper envelop and supporting lines



$$g(x) \geq a_0 x + b_0;\ g(x_0) = a_0 x_0 + b_0.$$

Supporting line at x_0 touches $g(x)$ at x_0, but below $g(x)$ at other places.

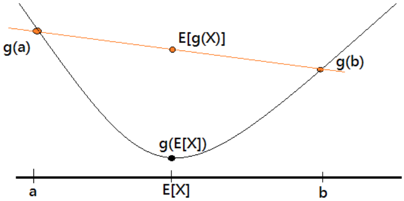
Jensen's inequality

$P(X = a) = P(X = b) = 1/2.$

$\mathrm{E}(X) = (a + b)/2$, $g(\mathrm{E}(X)) = g((a + b/2)$.

$\mathrm{E}(g(X)) = (g(a) + g(b))/2.$

$\mathrm{E}(g(X)) \geq g(\mathrm{E}(X))$. Note: $g(x)$ is convex



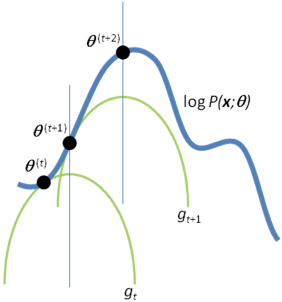
$x_0 = \mathrm{E}(X)$. $g(x_0) = a_0 x_0 + b_0$ (supporting line at x_0)
$g(x) \geq a_0 x + b_0$.
$\mathrm{E}(g(X)) \geq \mathrm{E}(a_0 X + b_0) = a_0\,\mathrm{E}(X) + b_0 = a_0 x_0 + b_0 = g(\mathrm{E}(X)).$

Ascent property of EM

Let $\ell(\theta | Y_{obs}) := \log P(Y_{obs} | \theta)$, which is the observed-data log-likelihood. Then the EM iterations satisfy

$$\ell(\theta^{(t+1)} | Y_{obs}) \geq \ell(\theta^{(t)} | Y_{obs})$$

EM Algorithm: Evidence Lower Bound and Convergence



Theorem

Under some conditions, the sequence $\{\theta^{(t)}\}$ defined by the EM iteratations converges to a stationary point of the observed-data log-likelihood $\log(P(y_{obs}|\theta))$.

Revisit the mixture model

- Observed variables $X = (X_1, X_2, \ldots, X_n)$
  An observation of X is called an incomplete data set
- Unobserved variables $Z = (Z_1, Z_2, \ldots, Z_k)$.
  (An observation (X,Z) is called a complete data set, but we never have a complete dataset)
- Parameters $\theta = (\pi, \mu, \sigma)$

  Cluster probabilities: $\pi = (\pi_1, \ldots, \pi_k)$
  Cluster means: $\mu = (\mu_1, \ldots, \mu_k)$
  Cluster standard deviation: $\sigma = (\sigma_1, \ldots, \sigma_k)$
- Complete data likelihood $P(X = i, Z = j|\theta) = \pi_j N(x_i|\mu_j, \sigma_j^2)$

Revisit the mixture model

1. Choose initial $\theta^{old} = (\pi^0, \mu^0, \sigma^0)$
2. Expectation step:

$$\log(P(X = i, Z = j|\theta)) = \log(\pi_j) + \log(N(x_i|\mu_j, \sigma_j^2))$$

$$p(z = j|x = i, \theta^{old}) = \gamma_i^j = \frac{\pi_j^{old} N(X_i|\mu_j^{old}, \sigma_{j,old}^2)}{\Sigma_{c=1}^k \pi_c^{old} N(X_i|\mu_c^{old}, \sigma_{c,old}^2)}$$

$$Q(\theta, \theta^{old}) = \Sigma_{i=1}^n \Sigma_{j=1}^k \gamma_i^j [\log(\pi_j) + \log N(x_i|\mu_j, \sigma_j^2)]$$

3. Maximization step:

$$\theta^{new} = \operatorname*{argmax}_{\theta} Q(\theta, \theta^{old})$$

4. Let $\theta^{old} = \theta^{new}$, go to step 2, until convergence

Maximization Step

- $\pi_j^{new} = \frac{\Sigma_{i=1}^{n} \gamma_i^j}{n}$
- $(n_j^{new} = n * \pi_j^{new})$
- $\mu_j^{new} = \frac{\Sigma_{i=1}^{n} \gamma_i^j x_i}{n_j^{new}}$
- $\sigma_{j,new}^2 = \frac{1}{n_j^{new}} \Sigma_{i=1}^{n} \gamma_i^j (x_i - \mu_j^{new})^2$

for each $j = 1, \ldots, k$

More examples

- Bivariate binary data
- Multinomial distribution with cell probabilities
- Coin flipping (A paper from Nature Biotechnology)