Expectation Maximization:

A general approach for learning with latent variables

CS534

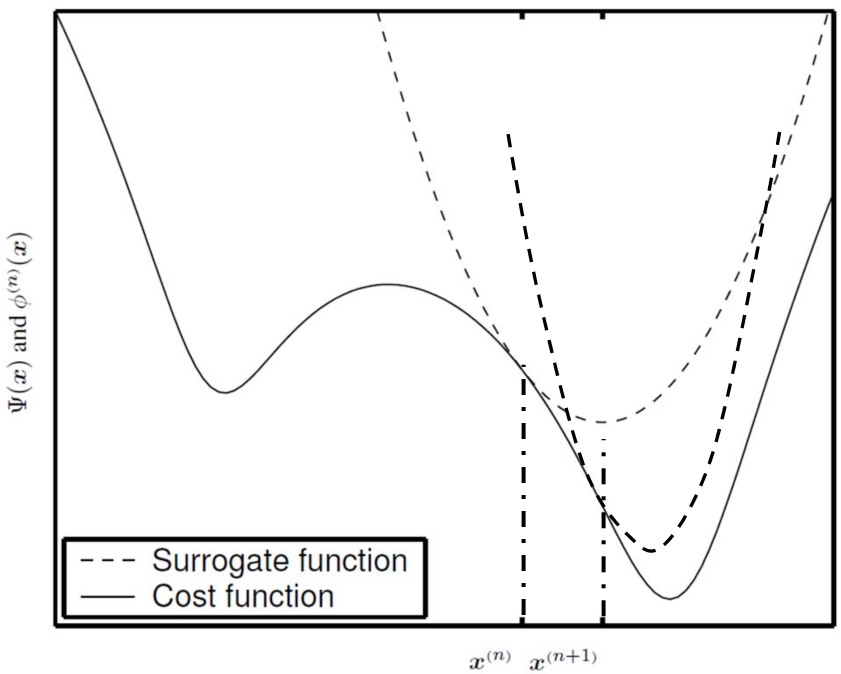
Maximum Likelihood Estimation with Latent Variables

- Suppose we have an estimation problem with a training set $\{x_1, x_2, \dots, x_m\}$
- We wish to fit the parameter of a model $p(x, z)$ to the data
- The log-likelihood function is:

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z_i} p(x_i, z_i; \theta)$$

- Directly maximizing $l(\theta)$ can be hard
- Here the z's are the latent variables
- It is often the case that if z is observed, the maximum likelihood estimation is easy to compute for θ

Optimization Transfer (OT)



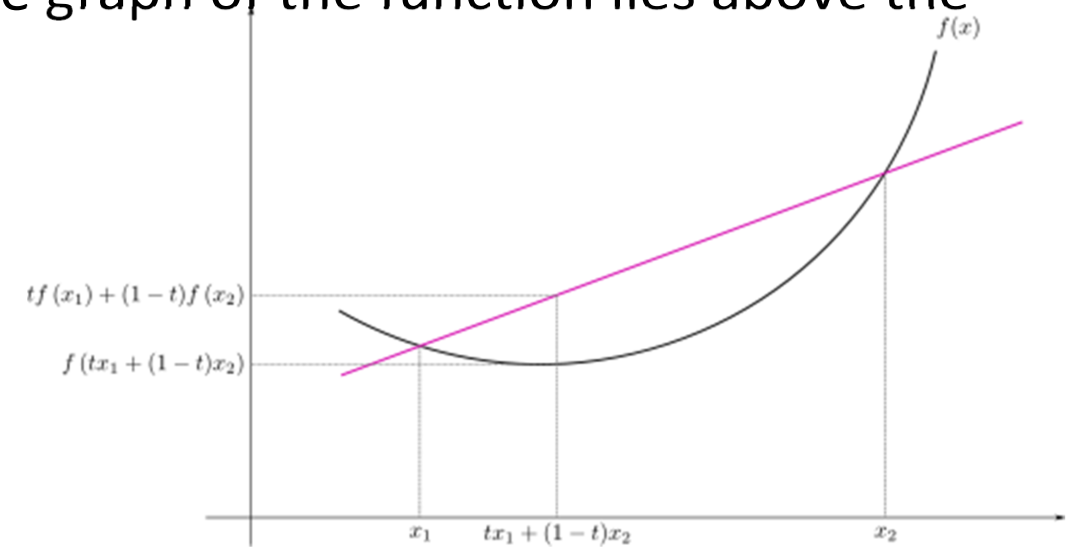
- Given a complex function Ψ to minimize (shown as the solid line)
- OT works iteratively, minimizes a surrogate function ϕ_n at each iteration:
  - $x^{n+1} = \arg\min_{x} \phi_n(x)$
- In any iteration n, if the surrogate function satisfy the following condition:
  - $\phi_n(x^n) = \Psi(x^n)$ (match at current pos)
  - $\phi_n(x) \geq \Psi(x)$ (lies above)
- we are guaranteed to monotonically improve in each iteration
  - $\Psi(x^{n+1}) \leq \Psi(x^n)$

Expectation maximization uses optimization transfer to maximize the log-likelihood

- Each iteration, it finds an lower bound of the log-likelihood using Jensen's inequality
- It then maximizes the lower bound

Jensen's Inequality

- Definition: a function is **convex** if the line segment between any two points on the graph of the function lies above the graph



- **Jensen's inequality:**

If f is convex, and let x be a random variable, then:

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

If f is concave, then: $E[f(x)] \leq f(E[x])$

Log-likelihood and lower bound

- Objective: Log-likelihood function

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z_i} p(x_i, z_i; \theta) = \sum_{i=1}^{m} \log \sum_{z_i} \frac{q_i(z_i) p(x_i, z_i; \theta)}{q_i(z_i)}$$

$$= \sum_{i=1}^{m} \log \sum_{z_i} q_i(z_i) [\frac{p(x_i, z_i; \theta)}{q\_i(z_i)}] = \sum_{i=1}^{m} \log E_{z_i \sim q_i(z_i)} \left[\frac{p(x_i, z_i; \theta)}{q_i(z_i)} \right]$$

$$\geq \sum_{i=1}^{m} E_{z_i \sim q_i(z_i)} [\log \frac{p(x_i, z_i; \theta)}{q_i(z_i)}]$$

Log is a concave function
Using Jensen's inequality

- For any distribution $q_i(z_i)$, this gives a lower bound to the log-likelihood
- To be a valid surrogate for optimization transfer, we also need it to match l at current parameter θ^n:

$$\log \sum_{z_i} q_i(z_i) [\frac{p(x_i, z_i; \theta^n)}{q_i(z_i)}] = \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; \theta^n)}{q_i(z_i)}$$

Further developing the surrogate

- We want to satisfy

$$\log \sum_{z_i} q_i(z_i)\left[\frac{p(x_i, z_i; \theta^n)}{q_i(z_i)}\right] = \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; \theta^n)}{q_i(z_i)}$$

- If the circled part is constant across all possible z_i values then we will have:
  - Left side: $\log \sum_{z_i} C q_i(z_i) = \log C \sum_{z_i} q_i(z_i) = \log C$
  - Right side: $\sum_{z_i} q_i(z_i) \log C = \log C \sum_{z_i} q_i(z_i) = \log C$
- So we have

$$\frac{p(x_i, z_i; \theta^n)}{q_i(z_i)} = C \rightarrow q_i(z_i) = \frac{1}{C} p(x_i, z_i; \theta^n)$$

- Note that q_i must satisfy $\sum_{z_i} q_i(z_i) = 1$

$$\frac{1}{C} \sum_{z_i} p(x_i, z_i; \theta^n) = 1$$

$$C = \sum_{z_i} p(x_i, z_i; \theta^n) = p(x_i; \theta^n)$$

$$q_i(z_i) = \frac{p(x_i, z_i; \theta^n)}{p(x_i; \theta^n)} = p(z_i | x_i; \theta^n)$$

Expectation Maximization

Repeat until convergence {

$//\theta^n$: current parameters in iteration n

(E-step) For each data point i, compute posterior of z_i:

$$q_i(z_i) = p(z_i|x_i;\theta^n)$$

(M-step) Maximize the expected log-likelihood

$$\theta^{n+1} = \arg\max_\theta \sum_i \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{q_i(z_i)}$$

$$= \arg\max_\theta \sum_i \sum_{z_i} q_i(z_i) \log p(x_i, z_i; \theta)$$

$n \leftarrow n + 1$

}

Mixture of Gaussian revisited

- Goal: given $(x_1, \ldots, x_m)$, fitting parameters $\alpha_1, \ldots, \alpha_k; \mu_1, \ldots, \mu_k; \Sigma_1, \ldots, \Sigma_k$
- The cluster labels are the latent variables $z_i' s$
- E-step:
  - Compute $q_i(z_i) = p(z_i | x_i; \theta^n)$ - posterior of cluster label
- M-step:
  - $\arg\max_\theta \sum_i \sum_{z_i} q_i(z_i) \log p(x_i, z_i | \theta)$ - maximize the expected complete loglikelihood

M-step

$$\sum_i \sum_{z_i} q_i(z_i) \log p(x_i, z_i; \theta)$$

$$= \sum_i \sum_{z_i} q_i(z_i) \log p(x_i | z_i; \theta) p(z_i; \theta)$$

$$= \sum_i \sum_{z_i} q_i(z_i) \left(\log N\left(x_i; \mu_{z_i}, \Sigma_{z_i}\right) + \log \alpha_{z_i} \right)$$

This can be viewed as doing maximum likelihood estimation of weighted complete data:

- Each data point is used to create k weighted labeled examples, one for each label
- The weight of the data point is $q_i(z_i)$, i.e., the cluster posterior of the data points

Summary of EM

- Expectation maximization is a general approach based on optimization transfer for maximum likelihood estimation with latent data
- In each iteration,
  - E-step computes posterior prob. of the latent variable given observed data and current parameters
  - M-step maximizes the expected complete data likelihood assuming that the latent variable follows the posterior distribution computed in E-step
- It is guaranteed to improve the log-likelihood objective monotonically