Mathematical Tools: Expectation Maximization Algorithm

Lecture 5

Instructor: Su-In Lee

Maximum likelihood estimation review

- Find parameter estimates which make observed data most likely
- General approach, as long as tractable likelihood function exists
- Can use all available information

Not all data are perfect

- Most MLE problems are simple to solve with complete data.
- Available data are “incomplete” in some way.

Expectation maximization algorithm

- Consider a set of starting parameters
- Use these to “estimate” the complete data
- Use estimates of complete data to update parameters
- Repeat as necessary

[Concalo Abecasis, biostat666]

The *Diabetes* example

- A mixture of M individuals
  - Some individuals have type II diabetes (we do not know who they are).
  - Denote by z_i the disease state.
- For each individual i, we measure various phenotypes (blood glucose level, insulin level,…), denoted by $\mathbf{x}_i$.
- Learn the parameters that encode the distribution of **x** give the disease state z.

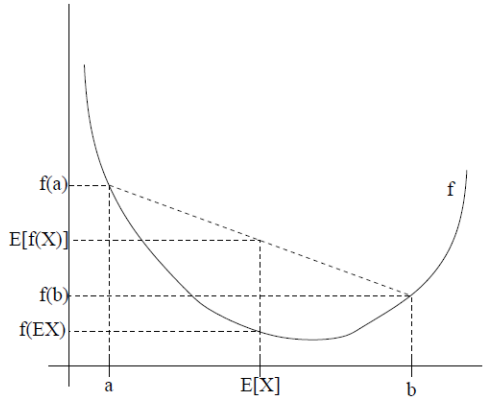
EM approach for the *Diabetes* problem

- Consider a set of starting parameters Θ.
- Infer the latent variables **z** given observed data **x** and the current parameters Θ.
- Use estimates of **z** to update parameters Θ.
- Repeat until the convergence.

Jensen's inequality

- **Theorem:** Let f be a convex function, and let X be a random variable. Then:

$$E[f(X)] \geq f(EX)$$



[Andrew Ng, cs229]

The EM algorithm

- The *Diabetes* problem
  - Training data: measurements from M individuals {x[1],…,x[M]}
  - Goal: Fit the parameters Θ of a model P(x,z) to the data
- The likelihood function:

The EM algorithm

The EM algorithm

- Procedure:

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

}

[Andrew Ng, cs229]

Convergence of the EM algorithm

Acknowledgement

- This set of slides is based on the following materials:
  - Andrew Ng's lecture note
  - Goncalo Abecasis' lecture note