Machine Learning - Winter 2015/16

Bernstein Focus: Neurotechnology Frankfurt

LECTURE : EM ALGORITHM

PREPARED BY
PROF. DR. VISVANATHAN RAMESH

(VERSION 1.0)

Outline

- Expectation Maximation Algorithm (Background)
- Convexity
- Jensen's Inequality
- EM Algorithm Formulation
- Short outline of Proofs
- Summary

Convex Functions

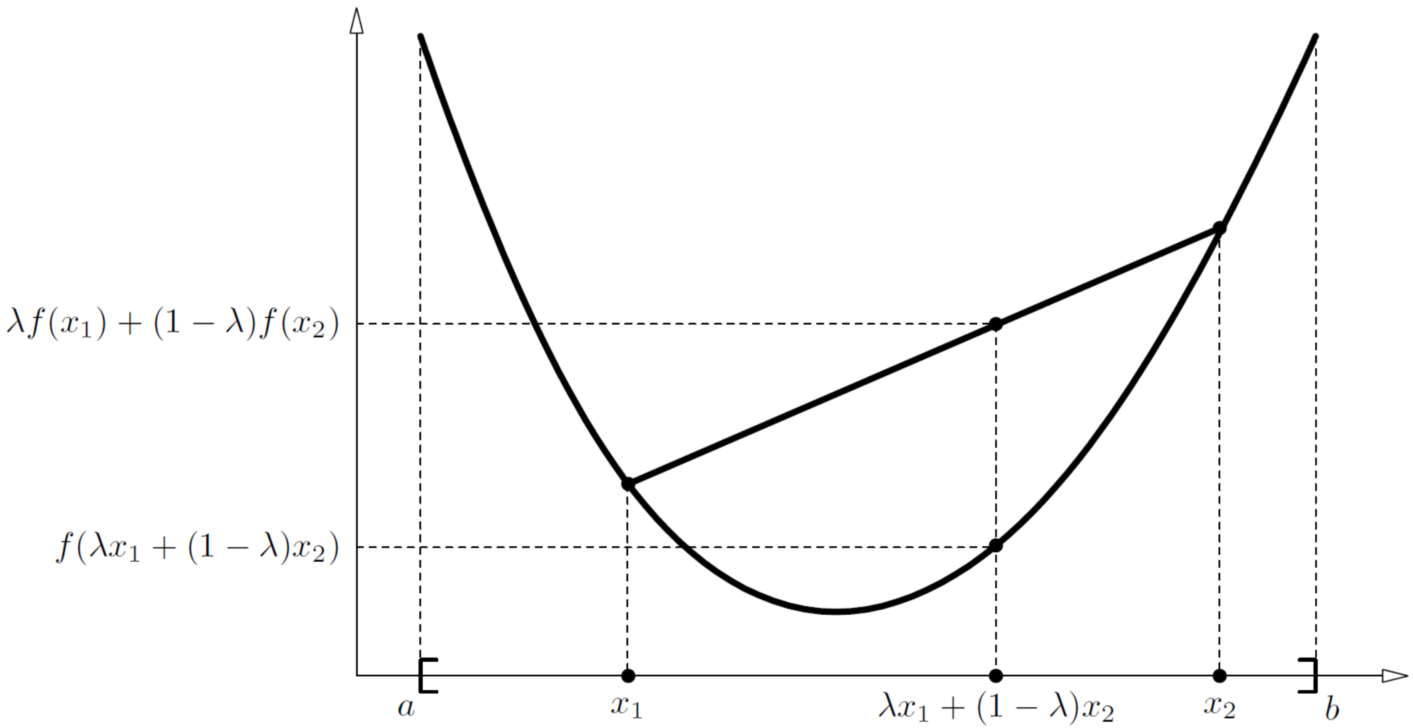

Figure 1: f is *convex* on $[a,b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$ $\forall x_1, x_2 \in [a,b], \quad \lambda \in [0,1]$.

Definitions

Definition 1 *Let f be a real valued function defined on an interval $I = [a, b]$. f is said to be* convex *on I if $\forall x_1, x_2 \in I, \lambda \in [0, 1]$,*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

f is said to be strictly convex *if the inequality is strict. Intuitively, this definition states that the function falls below (strictly convex) or is never above (convex) the straight line (the secant) from points $(x_1, f(x_1))$ to $(x_2, f(x_2))$. See Figure (1).*

Definition 2 *f is concave (strictly concave) if $-f$ is convex (strictly convex).*

Theorem 1 *If $f(x)$ is twice differentiable on $[a, b]$ and $f''(x) \geq 0$ on $[a, b]$ then $f(x)$ is convex on $[a, b]$.*

Jensen's Inequality

Theorem 2 (Jensen's inequality) *Let f be a convex function defined on an interval I. If $x_1, x_2, \ldots, x_n \in I$ and $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$,*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

- Proof follows by Induction (Trivial for n = 1, n=2 → follows from convexity, demonstrate for n+1 assuming theorem true for n).

Since $\ln(x)$ is concave, we may apply Jensen's inequality to obtain the useful result,

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i). \tag{6}$$

This allows us to lower-bound a logarithm of a sum, a result that is used in the derivation of the EM algorithm.

EM Algorithm Overview

Let $\mathbf{X}$ be random vector which results from a parameterized family. We wish to find θ such that $\mathcal{P}(\mathbf{X}|\theta)$ is a maximum. This is known as the Maximum Likelihood (ML) estimate for θ. In order to estimate θ, it is typical to introduce the *log likelihood function* defined as,

$$L(\theta) = \ln \mathcal{P}(\mathbf{X}|\theta). \tag{7}$$

The likelihood function is considered to be a function of the parameter θ given the data $\mathbf{X}$. Since $\ln(x)$ is a strictly increasing function, the value of θ which maximizes $\mathcal{P}(\mathbf{X}|\theta)$ also maximizes $L(\theta)$.

The EM algorithm is an iterative procedure for maximizing $L(\theta)$. Assume that after the n^{th} iteration the current estimate for θ is given by θ_n. Since the objective is to maximize $L(\theta)$, we wish to compute an updated estimate θ such that,

$$L(\theta) > L(\theta_n) \tag{8}$$

Equivalently we want to maximize the difference,

$$L(\theta) - L(\theta_n) = \ln \mathcal{P}(\mathbf{X}|\theta) - \ln \mathcal{P}(\mathbf{X}|\theta_n). \tag{9}$$

EM Algorithm (Derivation)

$$L(\theta) - L(\theta_n) = \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n). \tag{11}$$

$$\begin{aligned}
L(\theta) - L(\theta_n) &= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta) \cdot \frac{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \frac{\mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&\geq \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \ln \left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) && (12) \\
&= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \ln \left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z}, \theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \right) && (13) \\
&\triangleq \Delta(\theta|\theta_n). && (14)
\end{aligned}$$

$$l(\theta|\theta_n) \triangleq L(\theta_n) + \Delta(\theta|\theta_n)$$

EMAlgorithm (Continued)

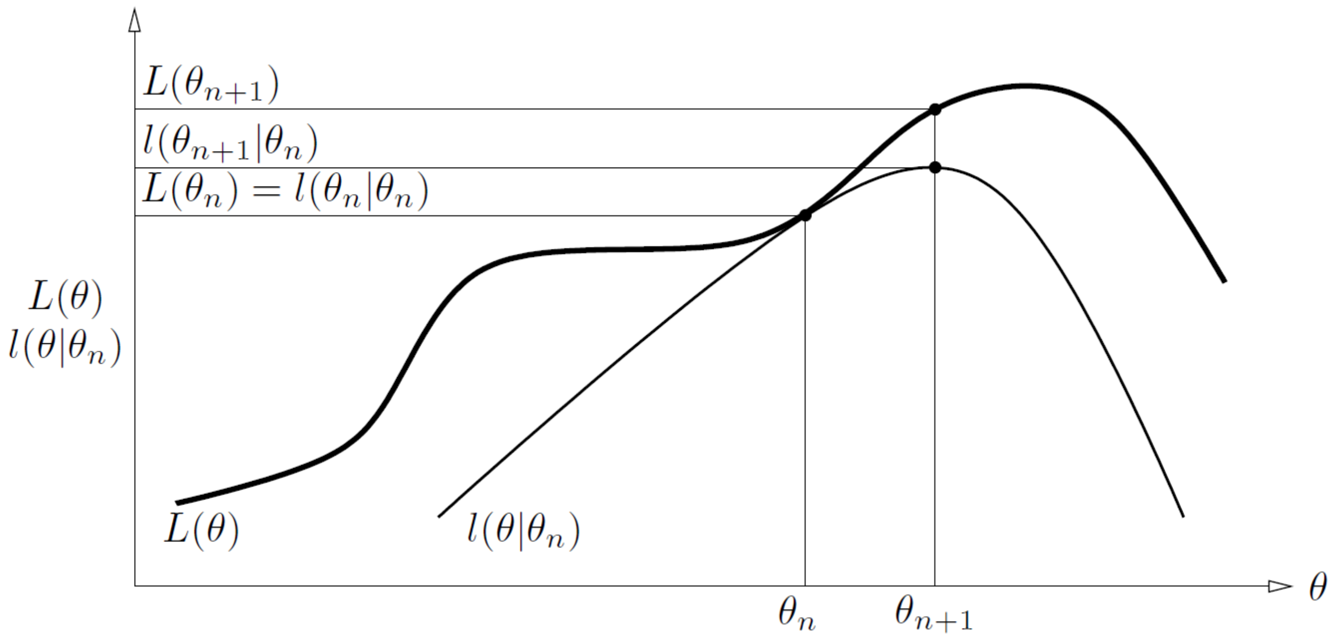

Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

EM Algorithm (Derivation)

$$
\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta} \{ l(\theta|\theta_n) \} \\
&= \arg\max_{\theta} \left\{ L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right\} \\
& \quad \text{Now drop terms which are constant w.r.t. } \theta \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X},\mathbf{z},\theta)}{\mathcal{P}(\mathbf{z},\theta)} \frac{\mathcal{P}(\mathbf{z},\theta)}{\mathcal{P}(\theta)} \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n} \{ \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \} \right\} \qquad (17)
\end{aligned}
$$

EM Algorithm - Summary

1. *E-step*: Determine the conditional expectation $\mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n}\{\ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta)\}$
2. *M-step*: Maximize this expression with respect to θ.

Key Points:

- Iteratively converges to a local maximum
- Detailed Proof done later demonstrates convergence may not be only to Maxima (e.g. saddle points)
- Method is a unified principle for a number of estimation problems with Hidden variables and/or missing data.
- Several methods followed addressing computational speedups of algorithm