CS598JHM: **Advanced NLP** (Spring ’10)

More on EM and variational inference

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

Today

- The connection between EM and variational inference
- Exponential families

Bishop (2006) sections 2.4, 9.3, 9.4, 10.1, 10.4

Exponential Families

Exponential families

- Most parametric distributions that we’ve seen so far belong to the exponential family of distributions

Distributions in the exponential family are nice because:

- They have **conjugate priors** (other distributions generally don’t)
- The likelihood and posterior can be expressed in terms of **sufficient statistics**

Definition

The **exponential family of distributions** over x
(x can be scalar or vector; discrete or continuous)
given the "natural" parameters $\boldsymbol{\eta}$
is defined as the set of distributions

$$p(x|\boldsymbol{\eta}) = h(x)g(\boldsymbol{\eta})exp(\boldsymbol{\eta}^T u(x))$$

-$g(\boldsymbol{\eta})$: normalization coefficient:
$g(\boldsymbol{\eta}) = (\int x\ exp(\boldsymbol{\eta}^T u(x)))^{-1}$
-$u(x)$: some function of x

Likelihood

Given a sequence of i.i.d observations $Y=(y_1,.., y_n)$, the likelihood $P(Y|\boldsymbol{\eta})$ is:

$$P(Y|\boldsymbol{\eta}) = \left[\prod_{i=1}^{n} h(y_i)\right] g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^T \sum_{i=1}^{n} u(y_i)\right)$$

Define a function $t(Y)$, called **sufficient statistics:**

$$t(Y) = \sum_{i=1}^{n} u(y_i)$$

Thus:
$$\begin{aligned} P(Y|\boldsymbol{\eta}) &= \left[\prod_{i=1}^{n} h(y_i)\right] g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^T t(Y)\right) \\ &\propto g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^T t(Y)\right) \end{aligned}$$

Conjugate priors

It is straightforward to define a conjugate prior for members of the exponential family:

$$\begin{aligned} \textbf{Likelihood} \quad P(Y|\boldsymbol{\eta}) &\propto g(\boldsymbol{\eta})^n e^{\boldsymbol{\eta}^T t(Y)} \\ \textbf{Prior} \quad P(\boldsymbol{\eta}) &\propto g(\boldsymbol{\eta})^\mu e^{\boldsymbol{\eta}^T \nu} \\ \textbf{Posterior} \quad P(\boldsymbol{\eta}|Y) &\propto P(\boldsymbol{\eta})P(Y|\boldsymbol{\eta}) \\ &= g(\boldsymbol{\eta})^{\mu+n} e^{\boldsymbol{\eta}^T (\nu + t(Y))} \end{aligned}$$

Expectation Maximization revisited

The EM algorithm

The goal of EM: Find the maximum likelihood solution for a model consisting of parameters θ, given observed (incomplete) data **X** and latent variables **Z**

$$\ln p(\mathbf{X}|\theta) \quad = \quad \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\right\}$$

Note: Even if complete likelihood $p(X,Z \mid \theta)$ is in exponential family, incomplete likelihood $p(X \mid \theta)$ may not be.

We just have **incomplete data X**, so don't know $p(X,Z \mid \theta)$.
We can only infer **Z** from posterior $p(Z \mid X,\theta)$.
We will compute the **expectation** of $p(X,Z \mid \theta)$ wrt. $p(Z \mid X, \theta)$

The EM algorithm

1. **Initialization:** Choose initial θ^{old}

2. **Expectation step:**
Compute **posterior of the latent variables** $p(Z \mid X, \theta^{old})$

3. **Maximization step:**
Find θ^{new} which **maximize the expected log-likelihood of the joint** $p(Z,X \mid \theta^{new})$ under $p(Z \mid X, \theta^{old})$:

$$\theta^{new} \quad = \quad \arg\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta)$$

4. **Check for convergence.**
Stop, or set $\theta^{old} := \theta^{new}$ and go to 2.

Another view of EM

We want to maximize $\quad \ln p(\mathbf{X}|\theta) \quad = \quad \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\right\}$

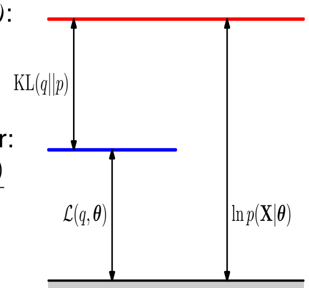
By the product rule: $\quad \ln p(\mathbf{X},\mathbf{Z}|\theta) \quad = \quad \ln p(\mathbf{Z}|\mathbf{X}\theta) + \ln p(\mathbf{X}|\theta)$

Define a functional of distribution $q(Z)$:

$$\mathcal{L}(q,\theta) \quad = \quad \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

KL-divergence btw. $q(Z)$ and posterior:

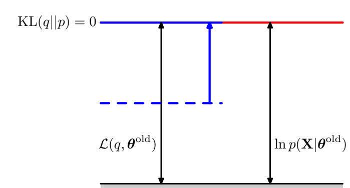
$$KL(q||p) \quad = \quad -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}$$



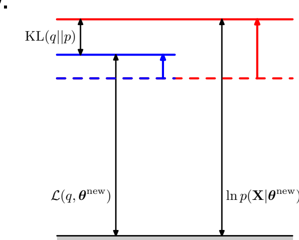
Thus $\ln p(\mathbf{X}|\theta) = \mathcal{L}(q,\theta) + KL(q||p)$

EM again...

$\mathcal{L}(q, \theta)$ is a lower bound on log-likelihood $\ln p(X \mid \theta)$



E-step:
Maximize $\mathcal{L}(q, \theta^{old})$ wrt. $q(Z)$,
keep θ^{old} fixed.
This happens when $KL(q||p) = 0$.



M-step:
Maximize $\mathcal{L}(q, \theta^{old})$ wrt. θ,
keep $q(Z)$ fixed.
$\mathcal{L}(q, \theta)$ will increase.
Thus $\ln p(X \mid \theta)$ will increase.
Hence, now: $KL(q||p) > 0$

Variational inference for Bayesian models

Bayesian model

- In a fully Bayesian model, all parameters θ are stochastic variables with priors.
- Now **Z** consists of latent variables and priors.
- We still want to maximize (incomplete) log-likelihood:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + KL(q||p)$$

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ KL(q||p) &= -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}\end{aligned}$$

Factorized distributions

Assume q factorizes:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

We still want to maximize $\mathcal{L}(q)$.

We can do this by optimizing with respect to each factor q_i in turn

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_i q_i \Big\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \Big\} d\mathbf{Z} \\ &= \int q_j \langle \ln p(\mathbf{X},\mathbf{Z}) \rangle_{i \neq j} + c' \; d\mathbf{Z_j} - \int q_i \ln q_j d_{Z_j} + c\end{aligned}$$