EM for HMMs a.k.a. The Baum-Welch Algorithm

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 11th, 2007

The general learning problem with missing data

- Marginal likelihood – **x** is observed, **z** is missing:

$$
\begin{aligned}
\ell(\theta : \mathcal{D}) &= \log \prod_{j=1}^{m} P(\mathbf{x}_j \mid \theta) \\
&= \sum_{j=1}^{m} \log P(\mathbf{x}_j \mid \theta) \\
&= \sum_{j=1}^{m} \log \sum_{\mathbf{z}} P(\mathbf{x}_j, \mathbf{z} \mid \theta)
\end{aligned}
$$

observed parts

sum over hidden vars
(marginalize out)

EM is coordinate ascent

$$\overbrace{\ell(\theta:\mathcal{D})}^{\text{marginal likelihood}} \;\geq\; \underset{\uparrow\,\text{max.}}{F}(\theta,Q) = \sum_{j=1}^{m}\sum_{\mathbf{z}} Q(\mathbf{z}\mid\mathbf{x}_j)\log\frac{P(\mathbf{z},\mathbf{x}_j\mid\theta)}{Q(\mathbf{z}\mid\mathbf{x}_j)}$$

- **M-step**: Fix Q, maximize F over θ (a lower bound on $\ell(\theta:\mathcal{D})$):

$$\ell(\theta:\mathcal{D}) \;\geq\; F(\theta,Q^{(t)}) = \sum_{j=1}^{m}\sum_{\mathbf{z}} Q^{(t)}(\mathbf{z}\mid\mathbf{x}_j)\log P(\mathbf{z},\mathbf{x}_j\mid\theta) + m.H(Q^{(t)})$$

Expected counts.

- **E-step**: Fix θ, maximize F over Q:

$$\ell(\theta^{(t)}:\mathcal{D}) \geq F(\theta^{(t)},Q) \;=\; \ell(\theta^{(t)}:\mathcal{D}) - \sum_{j=1}^{m} KL\left(Q(\mathbf{z}\mid\mathbf{x}_j)\|P(\mathbf{z}\mid\mathbf{x}_j,\theta^{(t)})\right)$$

  - "Realigns" F with likelihood:

$$F(\theta^{(t)},Q^{(t+1)}) \;=\; \ell(\theta^{(t)}:\mathcal{D})$$