CG168 Homework 7

Mark Johnson

due 2nd December 2008

Please modify your EM clusterer to use Mean Field Variational Bayes for Dirichlet-Multinomials. Despite the long name, this is actually very easy. The E-step is exactly the same as in regular EM; compute $n_y^{(t)}$, the expected number of times cluster y is seen at iteration t, and $n_{y,x}^{(t)}$, the expected number of times word x is seen in cluster y at iteration t, just as before.

The Mean Field Variational Bayes M-step is only slightly more complicated: it is:

$$\begin{aligned}
\theta_y^{(t+1)} &= \exp\left(\Psi(\alpha + n_y^{(t)}) - \Psi(\sum_{y' \in \mathcal{Y}} \alpha + n_{y'}^{(t)})\right) \\
\varphi_{y,x}^{(t+1)} &= \exp\left(\Psi(\beta + n_{y,x}^{(t)}) - \Psi(\sum_{x' \in \mathcal{X}} \beta + n_{y,x'}^{(t)})\right)
\end{aligned}$$

where α and β are the parameters of the Dirichlet priors for the $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ respectively. The web site contains code for computing the Digamma function $\Psi(x)$, and the slides and notes in class directory `/course/cog168/` contain more details. (As those notes explain, these iterations in general do *not* increase the log likelihood of the data, so don't worry if that doesn't increase at each iteration).

Homework:

1. *Modify your EM clustering algorithm to use the M-step just described. Run your algorithm on* `motherese.txt` *with Dirichlet prior parameters* $\alpha = \beta = 1$. *Because VB is usually slower than EM, you should perform at least 50 iterations in a VB run. Using* `prominent-members.py` *or a similar program, report the 10 most salient words in each cluster. How do you think the clustering found by VB compares with the one found by EM?*