

CG168 final exam

Mark Johnson

Due midnight, Monday 8th December 2008

Please turn in this final exam electronically. A typeset pdf file or an ASCII file would be best, but a *legible* scanned pdf file would be fine also. You can use “LaTeX-like” notation if you want, e.g., you can write $\sum_y n_y(D')$ instead of $\sum_y n_y(D')$.

Imagine that we scraped data from the Web consisting of movie titles together an adjective describing that movie. Our data might begin something like the following:

<u>Adjective</u>	<u>Movie</u>
action-packed	Batman
dramatic	Casablanca
futuristic	Star Trek
classic	Casablanca

Your job is to figure out how to cluster the movies by the adjectives they appear with, and the adjectives by the movies they appear with. You could just concatenate the adjective and the movie to form a “sentence” and use the sentence clustering methods we’ve studied in this course but that probably wouldn’t be optimal: after all, “Scary Movie” might actually be better described as “funny” rather than “scary”. This same sort of *co-clustering problem* arises in lots of situations: e.g., given data pairing each user with the program they are currently using, we might want to cluster programs by the users that use them, and cluster users by the programs they use.

There are many ways this might be done. Let our data $D = ((x_1, y_1), \dots, (x_n, y_n))$, where each $x_i \in \mathcal{X}$ and each $y_i \in \mathcal{Y}$. In our movie clustering problem, \mathcal{X} is the set of adjectives and \mathcal{Y} is the set of movie titles. Let every data pair (x_i, y_i) be associated with a cluster or class $z_i \in \mathcal{Z}$, where $z_i \in \mathcal{Z}$ and $\mathcal{Z} = \{1, \dots, m\}$ is the set of clusters. We will assume that each (x, y, z) triple is generated by first generating z , then generating x and y independently given z as follows:

$$P(x, y, z) = P(z) P(x | z) P(y | z)$$

We parameterize these distributions as follows: $P(z) = \zeta_z$, $P(x | z) = \varphi_{z,x}$ and $P(y | z) = \theta_{z,y}$. In this exam your job will be to estimate the hidden clustering $\mathbf{z} = (z_1, \dots, z_n)$ as well as the model parameters ζ , φ and θ .

Questions:

1. Suppose that the clusters \mathbf{z} were in fact visible (e.g., after the expenditure of a great deal of money we managed to hand label each adjective-movie pair with the cluster they belong to), so our data D' consists of triples (x_i, y_i, z_i) . What are the maximum likelihood estimates for ζ , φ and θ ? Express your answer in terms of quantities like n , the total number of data items, or $n_{x,y}(D')$, which is the number of data items (x_i, y_i, z_i) in D' where $x_i = x$ and $y_i = y$.
2. Now suppose we put uniform Dirichlet priors on ζ , φ and θ with Dirichlet parameters α , β and γ on ζ , φ and θ respectively. (“Uniform” means that e.g., α does not vary with the cluster z). Again assuming visible data D' , what are the expected values of ζ , φ and θ under this prior distribution? Note: the number of adjectives is $|\mathcal{X}|$, the number of movies is $|\mathcal{Y}|$ and the number of clusters is $|\mathcal{Z}| = m$.
3. Now you will start to derive the EM algorithm for estimating \mathbf{z} , ζ , φ and θ from D . First, give the formula for calculating $P(z_i | x_i, y_i)$ given estimates for ζ , φ and θ .
4. Now explain what quantities need to be calculated in the E-step of the EM algorithm for this problem, and give the formulae for calculating these quantities.
5. Given the quantities you just described in Question 4., give the M-step, which calculates updated values for ζ , φ and θ .
6. Now give the corresponding Mean Field Variational Bayes M-step update formulae, assuming the Dirichlet priors specified in Question 2. (You can write $\Psi(v)$ as $\text{Digamma}(v)$ if that is easier).
7. In this question and the next, you will devise the sampling formula for a Gibbs sampler that generates samples from $P(\mathbf{z} | D, \alpha, \beta, \gamma)$. For this question, suppose D' is the set of (x_i, y_i, z_i) triples as in questions 1 and 2, where the clusters z_i are visible. Based on the expected values for ζ , φ and θ under uniform Dirichlet priors you gave in question 2, give the formula for $P(x, y, z | D', \alpha, \beta, \gamma)$ in terms of counts like $n_{x,y}(D')$ and α , β and γ . Then give the formula for $P(z | x, y, D', \alpha, \beta, \gamma)$ (you can give this in terms of $P(z, x, y | D', \alpha, \beta, \gamma)$).
8. Using this, now derive the Gibbs resampling formula for sampling $P(\mathbf{z} | D, \alpha, \beta, \gamma)$. In this question D' consists of triples (x_i, y_i, z_i) , where (x_i, y_i) is the i th pair values from D and z_i is the current sample for the cluster z_i for (x_i, y_i) . Let D'_{-j} contain all of the triples in D' except the j th one; i.e.,

$$D'_{-j} = ((x_1, y_1, z_1), \dots, (x_{j-1}, y_{j-1}, z_{j-1}), (x_{j+1}, y_{j+1}, z_{j+1}), \dots, (x_n, y_n, z_n)).$$

The collapsed Gibbs sampler iterates through the data items $j = 1, \dots, n$ in turn, and resamples z_j according to the distribution $P(z_j | x_j, y_j, D'_{-j}, \alpha, \beta, \gamma)$. Using the formulae you derived in the previous question, give the Gibbs sampling distribution

$P(z_j \mid x_j, y_j, D'_{-j}, \alpha, \beta, \gamma)$. (*Hint: this is a trivial modification to one of the formulae you derived in question 7*).