Assignment 2

Unsupervised & Probabilistic Learning

Maneesh Sahani & Yee Whye Teh

Due: Thursday Nov 1, 2012

Note: all assignments for this course are to be handed in to the Gatsby Unit, **not** to the CS department. Assignments are due at the **beginning** of the lecture or tutorial on the due date. Late assignments (included those handed in later on the due day) will be penalised. If you are unable to attend, you may hand in your assignment to either lecturer or TA prior to the due time, or to Barry Fong in the Alexandra House 4th floor reception. Do not leave them with anyone else.

Please attempt the first questions before the bonus ones.

1. **[65 points] EM for Binary Data**.

   Consider the data set of binary (black and white) images used in the previous assignment. Each image is arranged into a vector of pixels by concatenating the columns of pixels in the image. The data set has N images $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ and each image has D pixels, where D is (number of rows $\times$ number of columns) in the image. For example, image $\mathbf{x}^{(n)}$ is a vector $(x_1^{(n)}, \ldots, x_D^{(n)})$ where $x_d^{(n)} \in \{0, 1\}$ for all $n \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$.

   (a) Write down the likelihood for a model consisting of a mixture of K multivariate Bernoulli distributions. Use the parameters $\pi_1, \ldots, \pi_K$ to denote the mixing proportions $(0 \leq \pi_k \leq 1; \sum_k \pi_k = 1)$ and arrange the K Bernoulli parameter vectors into a matrix P with elements p_{kd} denoting the probability that pixel d takes value 1 under mixture component k. Assume the images are iid under the model, and that the pixels are independent of each other within each component distribution. [5 points]

   Just as we can for a mixture of Gaussians, we can formulate this mixture as a latent variable model, introducing a discrete hidden variable $s^{(n)} \in \{1, \ldots, K\}$ where $P(s^{(n)} = k|\boldsymbol{\pi}) = \pi_k$.

   (b) Write down the expression for the responsibility of mixture component k for data vector $\mathbf{x}^{(n)}$, i.e. $r_{nk} \equiv P(s^{(n)} = k|\mathbf{x}^{(n)}, \boldsymbol{\pi}, \mathsf{P})$ [5 points]

   (c) Implement the EM algorithm for a mixture of K multivariate Bernoullis. The algorithm should take as input the number K, a matrix X containing the data set, and a maximum number of iterations to run. The algorithm should run for that number of iterations or until the log likelihood converges (does not increase by more than a very small amount). Beware of numerical problems as likelihoods can get very small, it is better to deal with log likelihoods. Also be careful with numerical problems when computing responsibilities — it might be necessary to multiply the top and bottom of the equation for responsibilities by some constant to avoid problems. Hand in code and a high level explanation of what you algorithm does. [30 points]

   (d) Run your algorithm on the data set for values of K in $\{2, 3, 4\}$. Verify that the log likelihood increases at each step of EM. Report the log likelihoods obtained (measured in *bits*) and display the parameters found. [15 points]

   (e) Comment on how well the algorithm works, whether it finds good clusters (look at the responsibilities and try to interpret them), and how you might improve the model. [10 points]

2. **[20 points] Modelling Data**.

   (a) Download the data file called geyser.txt from the course web site. This is a sequence of 295 consecutive measurements of two variables from Old Faithful geyser in Yellowstone National Park: the duration of the current eruption in minutes (rounded to the nearest second), and the waiting time since the last eruption in minutes (to the nearest minute).

   Examine the data by plotting the variables within (`plot(geyser(:,1),geyser(:,2),'o');`) and between (e.g. `plot(geyser(1:end-n,1),geyser(n+1:end,1 or 2),'o');` for various `n`) time steps. Discuss and justify based on your observations what kind of model might be most appropriate for this data set. Consider each of the models we have encountered in the course through week 3: a multivariate normal, a mixture of Gaussians, a Markov chain, a hidden Markov model, an observed stochastic linear dynamical system and a linear-Gaussian state-space model. Can you guess how many discrete states or (continuous) latent dimensions the model might have? [10 points]

   (b) Consider a data set consisting of the following string of 160 symbols from the alphabet $\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$:

   ```
   AABBBACABBBACAAAAAAAAABBBACAAAAABACAAAAAABBBBACAAAAAAAAAAAABACABACAABBACAAABBBBA
   CAAABACAAAABACAABACAAABBACAAAABBBBACABBACAAAAAABACABACAAABACAABBBACAAAABACABBACA
   ```

   Look *carefully* at the above string. Having analysed the string, describe an HMM model for it. Your description should include the number of states in the HMM, the transition matrix including the values of the elements of the matrix, the emission matrix including the values of its elements, and the intial state probabilities. You need to provide some description/justification for how you arrived at these numbers. We do **not** expect you to implement the Baum-Welch algorithm—you should be able to answer this question just by examining the sequence carefully. [10 points]

3. **[15 points] Zero-temperature EM**. In the automatic speech recognition community, HMMs are sometimes trained by using the Viterbi algorithm in place of the forward–backward algorithm. In other words, in the E step of EM (Baum–Welch), instead of computing the expected sufficient statistics from the posterior distribution over hidden states: $p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \theta)$, the sufficient statistics are computed using the single *most probable* hidden state sequence: $\mathbf{s}^*_{1:T} = \arg\max_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \theta)$.

   (a) Is this algorithm guaranteed to converge (in the sense that the free-energy reaches an asymptote)? To answer this you might want to consider the proof for the EM algorithm and what happens if we constrain $q(s)$ to put all its mass on one setting of the hidden variables. Support your arguments. [10 points]

   (b) If it converges, will it converge to a maximum of the likelihood? If not, will it oscillate? Support your arguments. [5 points]

   (c) [Bonus] Why do you think this question is labelled "Zero-temperature EM" Hint: think about where temperature would appear in the the free-energy. [5 points]

4. **[Bonus: 20 points] SSMs**. Consider the following nonlinear system:

$$z_{t+1} = 0.9\, z_t - \frac{z_{t-1}^2}{2(z_{t-1}^2 + 1)} + e \qquad\qquad x_t = z_t + v$$

   where $e \sim N(0, 1)$ and $v \sim N(0, 0.1)$. Using Matlab, generate 300 data points from this system starting with $z_1 = z_2 = 0$.

   (a) Write it as a nonlinear state-space model: i.e. using a state vector $\mathbf{y}_t$ which depends only on the previous state $\mathbf{y}_{t-1}$ and noise; and with x_t depending on $\mathbf{y}_t$ and noise. [10 points]

   (b) With enough state variables, can you model the dynamics in the above system perfectly with a linear dynamical system? Argue why or why not. [5 points]

   (c) Fred says that if you give him a linear-Gaussian state-space model he can convert it into an equivalent one with the same number of states but with the covariance of the state-noise $Q = I$, the identity matrix. Is he right? Justify your answer. [5 points]