Assignment 2

Unsupervised Learning

Maneesh Sahani & Yee Whye Teh

Due: Mon Nov 3, 2008

Note: all assignments for this course are to be handed in to the Gatsby Unit, **not** to the CS department. Please hand in all assignments at the beginning of lecture on the due date to the lecturer. Late assignments will be penalised. If you are unable to come to class, you can also hand in assignments to Rachel Howes in the Alexandra House 4th floor reception.

Please attempt the first questions before the bonus ones. You will not receive any credit for the bonus questions if you don't show credible effort on the first ones.

1. **[55 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 $(x_1^{(n)}, \ldots, x_D^{(n)})^T$ 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. [5 points]

   Just like in a mixture of Gaussians we can think of this model as a latent variable model, with 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) Derive the M-steps needed to update the parameters $\boldsymbol{\pi}$ and P. [5 points]

   (d) Implement the EM algorithm for a mixture of K multivariate Bernoullis. The algorithm should take as input K, a matrix X containing the data set, and a number of iterations. 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 a listing of your code. [30 points]

   (e) Run your algorithm on the data set for varying $K = 2, 3, 4$. Verify that the log likelihood increases at each step of EM. Report the log likelihoods obtained (measured in *bits*) and display the parameters found. [5 points]

   (f) 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. [5 points]

2. **[25 points] Time series**. 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 (to nearest 0.1 minute), and the waiting time until the next eruption in minutes (to nearest minute).

   (a) Examine the data by plotting the variables within and between consecutive time steps. E.g. `plot(geyser(1:end-1,1),geyser(2:end,1),'o');`. Discuss and justify based on your observations what kind of model might be most appropriate for this data set: e.g. a mixture of Gaussians, a hidden Markov model, a linear dynamical system, etc. [10 points]

   Download the data file called data3d.txt from the course web site. This is a data set of 100 observations of 3-dimensional vectors. Examine this data set and answer the following questions. Clearly justify your answers with arguments and plots. Unjustified answers will not get credit.

   (b) Is this a time series? Justify your answer. [5 points]

   (c) What kind of probabilitic model would you use to model this data? Why? Some choices are probabilistic PCA, factor analysis, mixtures of Gaussians, hidden Markov models, linear dynamical systems, nonlinear dynamical systems. Support your answer with plots and arguments. [10 points]

3. **[20 points] Hidden Markov models**. Consider a data set consisting of the following string of 160 symbols from the alphabet {A, B, C}:

```
AABBBACABBBACAAAAAAAAAABBBACAAAAABACAAAAAABBBBACAAAAAAAA
AAAABACABACAABBACAAABBBBACAAABACAAAABACAABACAAABBACAAAAA
BBBBACABBACAAAAAAABACABACAAABACAABBBACAAAABACABBACA
```

   (a) 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. I am **not** expecting you to code the HMM algorithm—you should be able to answer this question just by examining the sequence carefully. [10 points]

   (b) Do the same thing (build and justify an HMM model) for the following string of 60 symbols from the alphabet {A, B, C, D, E, F}:

   ```
   CEFBAFCEFBAFBAFCEFCADCAFBEFCEFBEFBEFCAFCEDCAFBEDCADBAFCAFCEF
   ```

   [10 points]

4. **[Bonus: 15 points] Zero-temperature EM**. In the automatic speech recognition community, HMMs are sometimes trained by using the Viterbi algorithm instead 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? 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. [10 points]

5. **[Bonus: 15 points] Nonlinear SSMs**. Consider the following nonlinear system:

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

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]

6. **[Bonus: 5 points] SSM identifiability**. 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.