Supervision Assignments in MLBI
Lent Term 2018
Set 3
due 9 March

Supervisor : Dionysis Manousakas
dm754@cam.ac.uk

February 24, 2018

1. *(Latent Variable Models.)*

   Describe a real-world data set which you believe could be modelled using a mixture model. Argue why a mixture model is a sensible model for your real world data set. What do you expect the mixture components to represent? How many components (or clusters) do you think there would be? What parametric form would each component have?

2. *(EM for Binary Data.)*

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

   (a) Write down the likelihood for a model consisting of a mixture of K multivariate Bernoulli distributions. Use the parameters $\pi_1, ..., \pi_K$ to denote the mixing proportions ($0 \leq \pi_k \leq ; \sum_k \pi_k = 1$) 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.

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

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

   (c) 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 code and a high level explanation of what you algorithm does.

   (d) 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.

(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.

3. (*Zero-temperature EM*)

   In the automatic speech recognition community, HMMs are sometimes trained by using the Viterbi algorithm instead of the forwardbackward 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: $p(s_{1:T}|x_{1:T}, \theta)$, the sufficient statistics are computed using the single most probable hidden state sequence: $s^*_{1:T} = \text{argmax}_{s_{1:T}} p(s_{1:T}|x_{1:T}, \theta)$.

   (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.

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

3. (*Machine Learning Methods*) Answer to 2016, paper 8, question 2.