SC4/SM8 Advanced Topics in Statistical Machine Learning
Problem Sheet 3

1. In lectures, we derived the M-step updates for fitting Gaussian mixtures with EM algorithm, for the mixing proportions and for the cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known.

   (a) What happens to the algorithm if we set σ^2 to be very small? How does the resulting algorithm as $\sigma^2 \to 0$ relate to K-means?

   (b) If σ^2 is in fact not known and is a parameter to be inferred as well, derive an M-step update for σ^2.

2. We are given a *labelled dataset* $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \{0,1\}^p$ and $y_i \in \{1, \ldots, K\}$ and the *naïve Bayes classifier model* which assumes that different dimensions/features in vector X_i are independent given the class label $Y_i = k$, resulting in the joint probability

$$p(x_i, y_i; \{\pi_k\}, \{\phi_{kj}\}) = \sum_{k=1}^{K} \left\{ \mathbf{1}\left(y_i = k\right) \pi_k \prod_{j=1}^{p} \left[\left(\phi_{kj}\right)^{x_i^{(j)}} \left(1 - \phi_{kj}\right)^{1 - x_i^{(j)}} \right] \right\}.$$

   where $\pi_k = \mathbb{P}(Y_i = k)$ are the marginal class probabilities and ϕ_{kj} is the probability of feature j being present in the class k, i.e., of $x_i^{(j)} = 1$ for an item x_i belonging to class k).

   (a) Derive the maximum likelihood estimates for π_k and ϕ_{kj}.

   (b) Assume that we are also given an additional set of *unlabelled data items* $\{x_i\}_{i=n+1}^{n+m}$. Using the same naïve Bayes model, and by treating missing labels as latent variables, describe an EM algorithm that makes use of this unlabelled dataset and give the E-step update for the variational distribution q and the M-step updates for parameters π_k and ϕ_{kj}. Discuss the difference of these results to those in part (a).

3. Verify that in the probabilistic PCA model from the lectures, E-step of the EM algorithm at iteration $t+1$ can be written as

$$q^{(t+1)}(y_i) = \mathcal{N}\left(y_i; b_i^{(t)}, R^{(t)}\right)$$

   where

$$b_i^{(t)} = \left((L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I\right)^{-1} (L^{(t)})^\top x_i, \tag{1}$$

$$R^{(t)} = (\sigma^2)^{(t)} \left((L^{(t)})^\top L^{(t)} + (\sigma^2)^{(t)} I\right)^{-1}. \tag{2}$$

4. Consider a collaborative filtering model with "implicit feedback" observations y_{ij} which indicate not the rating but some form of frequency of interaction of user j with item i (for example, a user may watch a TV series every week, but that does not necessarily mean that she would rate it higher than a film she has seen only once). We convert the implicit feedback into binary $b_{ij} = \mathbf{1}\{y_{ij} > 0\}$ and also introduce confidence measures $c_{ij} = 1 + \alpha y_{ij}$ for $\alpha > 0$ (note that we do not treat $y_{ij} = 0$ as missing - we simply have a lower confidence in those observations). For user j, we are then solving the weighted least squares problem:

$$\min_{\psi_j} \sum_{i=1}^{n_1} c_{ij}(b_{ij} - \phi_i^\top \psi_j)^2 + \lambda_\psi \|\psi_j\|_2^2, \quad j = 1, \ldots, n_2. \tag{3}$$

By expressing the criterion in matrix form, derive a closed form solution of ψ_j.

5. Consider a collaborative filtering model on binary ratings -1 and $+1$ with a *probit likelihood*

$$p\left(y_{ij}=1|a_i,b_j\right)=\Phi(a_i^\top b_j)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a_i^\top b_j}\exp\left(-t^2/2\right)dt, \tag{4}$$

where y_{ij} is the rating of item i by user j, $a_i \in \mathbb{R}^k$ is the feature vector of item i, b_j is the preference vector of user j and Φ is the standard normal cdf.

Consider an alternative model with additional latent variables z_{ij}, given by

$$z_{ij}|a_i,b_j \sim \mathcal{N}(a_i^\top b_j,1), \quad p(y_{ij}=1|z_{ij})=\mathbf{1}\{z_{ij}>0\}.$$

   (a) Show that these two models are equivalent, i.e. that $p\left(y_{ij}=1|a_i,b_j\right)$ still takes the form in (4).

   (b) Derive $p(z_{ij}|a_i,b_j,y_{ij}=\pm 1)$.

   (c) Now consider the model that treats feature vectors and preference vectors as model parameters $\theta = (\{a_i\}_{i=1}^{n_1},\{b_j\}_{j=1}^{n_2})$ with latents $\mathbf{Z}=(\{z_{ij}\}_{e_{ij}=1})$. Describe the resulting EM algorithm.

6. Consider the model $p(r|\lambda)=\frac{e^{-\lambda}\lambda^r}{r!}$ with $\lambda>0$ and the improper prior $p(\lambda)\propto\frac{1}{\lambda}$. Derive the Laplace approximation to the posterior $p(\lambda|r)$. Then change the parametrisation to $\theta=\log\lambda$, so that the prior is $p(\theta)\propto 1$, and find the Laplace approximation to the posterior $p(\theta|r)$. Which version of the Laplace approximation is better?

7. Suppose we have a model $p(\mathbf{X},\mathbf{z}|\theta)$ where $\mathbf{X}$ is the observed dataset and $\mathbf{z}$ are the latent variables. We would like to take a Bayesian approach to learning, treating the parameter θ to be a random variable as well, with some prior $p(\theta)$.

   (a) Suppose that $q(\mathbf{z},\theta)$ is a distribution over both $\mathbf{z}$ and θ. Explain why the following is a lower bound on $p(\mathbf{X})$:

$$\mathcal{F}(q)=\mathbb{E}_q[\log p(\mathbf{X},\mathbf{z},\theta)-\log q(\mathbf{z},\theta)]$$

   (b) Show that the optimal $q(\mathbf{z},\theta)$ is simply the posterior $p(\mathbf{z},\theta|\mathbf{X})$.

   (c) Typically the posterior is intractable. Consider a factorised distribution $q(\mathbf{z},\theta)=q_{\mathbf{z}}(\mathbf{z})q_\theta(\theta)$. In other words we assume that $\mathbf{z}$ and θ are independent. Derive the optimal $q_{\mathbf{z}}$ given a q_θ, and hence describe an algorithm to optimise $\mathcal{F}(q)$ subject to assumption of independence between $\mathbf{z}$ and q.

8. Verify steps (2) and (3) in the CAVI updates for the Latent Dirichlet Allocation model.