# 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 $\sigma^2$ to be very small? How does the resulting algorithm as $\sigma^2 \to 0$ relate to K-means?

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

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 $\phi_{kj}$ 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 $\pi_k$ and $\phi_{kj}$.

   (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 $\pi_k$ and $\phi_{kj}$. 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. 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 $\theta$ 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 $\theta$. 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 $\theta$ are independent. Derive the optimal $q_{\mathbf{z}}$ given a $q_\theta$, and hence describe an algorithm to optimise $\mathcal{F}(q)$ subject to assumption of independence between $\mathbf{z}$ and $q$.

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

6. Consider a Bayesian approach to neural networks, where we are interested in working with a posterior distribution over parameters rather than just a point estimate. Say we have a Gaussian prior for parameters, and a likelihood parameterised by a neural network. For simplicity, consider:

$$\theta \sim \mathcal{N}(0, \sigma_0^2 I_p) \tag{3}$$

$$y_i | x_i, \theta \sim \mathcal{N}(\cdot | f_\theta(x_i), \sigma_y^2) \tag{4}$$

where $f_\theta$ is the output of the NN with parameters $\theta$.

(a) Why is the posterior typically intractable? Give an example of a NN with a tractable posterior.

(b) How is the prior over $\theta$ related to weight decay?

(c) Derive the formula for the KL divergence between two 1D Gaussians $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. Verify that the KL divergence is non-negative in this case, and 0 only when the parameters of the Gaussians match.

(d) We can take a variational approach to approximating the posterior, say with a Gaussian variational distribution $q_{\mu,\sigma^2}(\theta) = \mathcal{N}(\theta; \mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are vectors of the same length as $\theta$, and we assume that the covariance matrix is diagonal with entries given by $\sigma^2$.

Describe an amortised variational inference approach to optimising the variational parameters $\mu$ and $\sigma^2$. Write down the objective function, and describe how the reparameterisation trick can be used to provide unbiased estimates of the gradient of the objective with respect to the variational parameters.