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?

   Answer: In the E-step, the posterior probabilities are:

   $$q_{\sigma^2}(z_i = k) \propto \pi_k f(x_i|\mu_k, \sigma^2) \propto \pi_k \exp(-\frac{1}{2\sigma^2}\|x_i - \mu_k\|_2^2)$$

   As σ^2 approaches zero, the exponentiated term will be dominated by the k such that μ_k is closest to x_i by Euclidean distance. Thus, if there is a unique such k, as $\sigma^2 \to 0$:

   $$q_{\sigma^2}(z_i = k) \to \begin{cases} 1 & \text{for } k = \text{argmin}_{k'} \|x_i - \mu_{k'}\|, \\ 0 & \text{otherwise.} \end{cases}$$

   If there is another $\mu_{k'}$ at same distance to x_i, $q(z_i)$ will spread probability mass equally among all such components. This looks exactly like the cluster assignment step of K-means (also note: π_k values have no effect on the cluster assignment in this limit). The M-step is exactly the mean update step, thus K-means can be understood as an EM algorithm for a mixture of Gaussians with infinitesimally small σ^2.

   (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.

   Answer: Differentiating the free energy with respect to $\nu = \sigma^{-2}$ (you can also differentiate with respect to σ or σ^2, just involves a bit more algebra) and setting to 0 gives:

   $$\begin{aligned} \nabla_\nu \mathcal{F}(\theta, q) &= \sum_{i=1}^n \sum_{k=1}^K q(z_i = k) \nabla_\nu \left(-\frac{p}{2} \log(2\pi/\nu) - \nu \frac{1}{2} \|x_i - \mu_k\|_2^2 \right) \\ &= \sum_{i=1}^n \sum_{k=1}^K q(z_i = k) \left(\frac{p}{2} \frac{1}{\nu} - \frac{1}{2} \|x_i - \mu_k\|_2^2 \right) \\ &= \frac{np}{2\nu} - \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K q(z_i = k) \|x_i - \mu_k\|_2^2 = 0 \\ &\Rightarrow \sigma^2 = \frac{\sum_{i=1}^n \sum_{k=1}^K q(z_i = k) \|x_i - \mu_k\|_2^2}{np}. \end{aligned}$$

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

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

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}.

Answer: Denoting $n_k = \sum_{i=1}^n \mathbf{1}(y_i = k)$, direct differentiation of the log-likelihood (with the Lagrange multiplier for π_k as before) gives

$$\hat{\pi}_k = \frac{n_k}{n}, \qquad \hat{\phi}_{k,j} = \frac{\sum_{i=1}^n \mathbf{1}(y_i = k) x_i^{(j)}}{n_k}$$

(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).

Answer:

We model the missing labels as latent random variables $z_i = y_{n+i}$, for $i = 1, \ldots, m$. E-step then sets the variational distribution $q(\mathbf{z})$ to the posterior of the missing labels given data and current parameters:

$$q^{(t+1)}(z_i = k) \propto \pi_k^{(t)} \prod_{j=1}^p \left[\left(\phi_{kj}^{(t)}\right)^{x_{n+i}^{(j)}} \left(1 - \phi_{kj}^{(t)}\right)^{(1 - x_{n+i}^{(j)})} \right]$$

and M-step finds the optimal parameters given q. The M-step update is:

$$\pi_k^{(t)} = \frac{n_k + \sum_{i=1}^m q^{(t)}(z_i = k)}{n + m}, \qquad \phi_{kj}^{(t)} = \frac{\sum_{i=1}^n \mathbf{1}(y_i = k) x_i^{(j)} + \sum_{i=1}^m q^{(t)}(z_i = k) x_{n+i}^{(j)}}{n_k + \sum_{i=1}^m q^{(t)}(z_i = k)}.$$

This is an example of a semisupervised problem. In part (a), all labels are observed (fully supervised setting) which is like having $q(z_i = k) = \mathbf{1}\{y_i = k\}$ and there are no latent variables.

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

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

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

Answer:

Follows from Gaussian conditioning, i.e., completing the square in the exponent. We omit superscript $\cdot^{(t)}$) on parameters $\theta = (L, \sigma^2)$ for simplicity.

$$\begin{aligned} p(y_i|x_i,\theta) &\propto p(y_i)p(x_i|y_i,\theta) \\ &\propto \exp\left(-\frac{1}{2}y_i^\top y_i\right)\exp\left(-\frac{1}{2\sigma^2}(x_i - Ly_i)^\top(x_i - Ly_i)\right) \\ &\propto \exp\left(-\frac{1}{2}\left[y_i^\top \underbrace{\left(I + \frac{1}{\sigma^2}L^\top L\right)}_{R^{-1}} y_i - 2y_i^\top \underbrace{\left(\frac{1}{\sigma^2}L^\top x_i\right)}_{R^{-1}b_i}\right]\right) \\ &\propto \mathcal{N}(y_i; b_i, R) \end{aligned}$$

where

$$R = \sigma^2\left(L^\top L + \sigma^2 I\right)^{-1}, \tag{3}$$

$$b_i = \left(L^\top L + \sigma^2 I\right)^{-1} L^\top x_i. \tag{4}$$

as required.

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

   Answer: We can write
   $$\mathcal{F}(q) = \mathbb{E}_q[\log p(\mathbf{z}, \theta|\mathbf{X}) - \log q(\mathbf{z}, \theta)] + \log p(\mathbf{X})$$
   with the first term being the negative of KL divergence, so is non-positive. Thus $\mathcal{F}(q)$ is a lower bound on $\log p(\mathbf{X})$.

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

   Answer: The optimal q is the one that minimises KL divergence to the posterior, i.e. the posterior $p(\mathbf{z}, \theta|\mathbf{X})$ itself.

   (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.

   Answer: Making the factorization assumption,
   $$\mathcal{F}(q) = \int q_{\mathbf{z}}(\mathbf{z})q_\theta(\theta)\{\log p(\mathbf{X}, \mathbf{z}, \theta) - \log q_{\mathbf{z}}(\mathbf{z}) - \log q_\theta(\theta)\}d\theta d\mathbf{z}$$
   Differentiate wrt $q_{\mathbf{z}}(\mathbf{z})$, including a Lagrange multiplier to make sure $q_{\mathbf{z}}(\mathbf{z})$ integrates to 1, we get:
   $$\nabla_{q_{\mathbf{z}}(\mathbf{z})}\mathcal{F}(q) = \int q_\theta(\theta)\log p(\mathbf{X}, \mathbf{z}, \theta)d\theta - \log q_{\mathbf{z}}(\mathbf{z}) - 1 + \lambda = 0$$
   $$q_{\mathbf{z}}(\mathbf{z}) \propto \exp\left(\int q_\theta(\theta)\log p(\mathbf{X}, \mathbf{z}, \theta)d\theta\right)$$

By symmetry, the optimal q_θ given $q_{\mathbf{z}}$ is:

$$q_\theta(\theta) \propto \exp\left(\int q_{\mathbf{z}}(\mathbf{z}) \log p(\mathbf{X}, \mathbf{z}, \theta) d\mathbf{z}\right)$$

We can alternate between optimizing q_θ given $q_{\mathbf{z}}$ and vice versa to maximise the lower bound. This is similar to the EM algorithm.

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

   Answer: For (2), we note that

$$\log p\left(\theta_d | z_d\right) = \text{const} + \sum_{k=1}^{K} \left(\alpha_k + \sum_{n=1}^{N_d} z_{dn}\left[k\right] - 1\right) \log \theta_{dk},$$

   so that

$$\exp\left[\mathbb{E}_{z_d \sim q} \log p\left(\theta_d | z_d\right)\right] \propto \prod_{k=1}^{K} \theta_{dk}^{\alpha_k + \sum_{n=1}^{N_d} \phi_{dn}[k] - 1},$$

   which is proportional to the Dirichlet distribution with parameter vector $\gamma_d = \alpha + \sum_{n=1}^{N_d} \phi_{dn}$ and similarly for (3).

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:

$$\theta \sim \mathcal{N}(0, \sigma_0^2 I_p) \tag{5}$$

$$y_i | x_i, \theta \sim \mathcal{N}(\cdot | f_\theta(x_i), \sigma_y^2) \tag{6}$$

   where f_θ is the output of the NN with parameters θ.

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

   Answer: The posterior is intractable because the functional dependence of f_θ on θ is non-linear, typically.

   An example with tractable posterior is a feedforward network with no hidden units. This is then just Bayesian linear regression.

   (b) How is the prior over θ related to weight decay?

   Answer: If we do MAP inference, the prior's role is exactly that of L_2 regularisation. The gradient then corresponds 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.

Answer:

$$\begin{aligned}
&\mathsf{KL}(\mathcal{N}(\mu_1, \sigma_1^2) \| \mathcal{N}(\mu_2, \sigma_2^2)) \\
=&\mathbb{E}_{\mathcal{N}(\mu_1,\sigma_1^2)} \left[\log \frac{(2\pi\sigma_1^2)^{-1/2} \exp(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2)}{(2\pi\sigma_2^2)^{-1/2} \exp(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2)} \right] \\
=&-\frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} - \frac{1}{2} + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1-\mu_2)^2) \\
=&-\frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} - \frac{1}{2} + \frac{1}{2}\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2\sigma_2^2}(\mu_1-\mu_2)^2
\end{aligned}$$

The first three terms can be shown to be non-negative, and 0 when $\sigma_1^2 = \sigma_2^2$. The last term is non-negative, and 0 when $\mu_1 = \mu_2$.

(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 μ and σ^2 are vectors of the same length as θ, and we assume that the covariance matrix is diagonal with entries given by σ^2.

Describe an amortised variational inference approach to optimising the variational parameters μ and σ^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.

Answer: The ELBO objective is:

$$\mathcal{F}(\mu, \sigma^2) = \mathbb{E}_{\theta \sim q_{\mu,\sigma^2}} \left[\sum_{i=1}^{n} |y_i - f_\theta(x_i)|^2 \right] - \mathsf{KL}(\mathcal{N}(\mu, \sigma^2) \| \mathcal{N}(0, \sigma_0^2))$$

Using the reparameterisation trick, we can write a sample $\theta \sim q_{\mu,\sigma^2}$ as $\theta = \mu + \epsilon\sigma$ where $\epsilon \sim \mathcal{N}(0, I)$. The objective is then:

$$\mathcal{F}(\mu, \sigma^2) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\sum_{i=1}^{n} |y_i - f_{\mu+\epsilon\sigma}(x_i)|^2 \right] - \mathsf{KL}(\mathcal{N}(\mu, \sigma^2) \| \mathcal{N}(0, \sigma_0^2))$$

We can now use a single Monte Carlo sample from ϵ, plus a minibatch, to form an unbiased estimate of the derivative of the first term. The derivatives of the second can be obtained directly.