EM and Mixture Models

1. † For $d$-dimensional data compare the computational cost of calculating the log-likelihood with a diagonal covariance matrix Gaussian distribution, a full covariance matrix Gaussian distribution and an $M$-component diagonal covariance matrix Gaussian mixture models. Clearly state any assumptions made.

2. For any convex function $f()$ show using a suitable diagram that

$$f((1 - t)a + tb) \geq (1 - t)f(a) + tf(b)$$

where $0 \leq t \leq 1$. Given a set of $M$ points $x_1, \ldots, x_M$ all lying in the interval $(a,b)$, and a set of $M$ numbers, $\lambda_j \geq 0$ such that $\sum_{m=1}^{M} \lambda_m = 1$, show that the quantity $\sum_{m=1}^{M} \lambda_m x_m$ also lies in the interval $(a,b)$. Starting from the stated inequality use induction to prove Jensen’s inequality

$$f\left(\sum_{m=1}^{M} \lambda_m x_m\right) \geq \sum_{m=1}^{M} \lambda_m f(x_m)$$

3. A series of $n$ independent, noisy, measurements are taken, $x_1, \ldots, x_n$. The noise is known to be Gaussian distributed with zero mean and unit variance. The “true” data is also known to be Gaussian distributed.

(a) Find the maximum likelihood estimates of the mean, $\mu$, and variance, $\sigma^2$, of the “true” data by equating the gradient to zero.

(b) A latent variable $z_i$ is introduced. It is the value of the noise for observation $x_i$. Show that the posterior probability of $z_i$ given the current model parameters is

$$p(z_i|x_i, \theta) = \mathcal{N}\left(z_i; \frac{(x_i - \mu)}{(1 + \sigma^2)}, \frac{\sigma^2}{(1 + \sigma^2)}\right)$$

Using the expectation-maximisation algorithm derive re-estimation formulae for the mean, $\mu$, and variance, $\sigma^2$. Show that the iterative estimation scheme for the mean converges to the correct answer, you may assume that the variance is of the true data is known and fixed at $\sigma^2$. 

Discuss the merits of the two optimisation schemes for this task and for optimisation tasks in general.

4. Consider an $M$ component mixture model of $d$-dimensional binary data $\mathbf{x}$ of the form

$$ p(\mathbf{x}) = \sum_{m=1}^{M} P(\omega_m)p(\mathbf{x}|\omega_m) $$

where the $j^{th}$ component PDF has parameters $\lambda_{j1}, \ldots, \lambda_{jd}$ and

$$ p(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} \lambda_{ji}^{x_{ji}}(1 - \lambda_{ji})^{1-x_{ji}} $$

A set of training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are used to train the mixture model. Using the standard form of EM with mixture models show that the maximum likelihood estimate for the “new” parameters, $\hat{\lambda}_{ji}$, is given by

$$ \hat{\lambda}_{ji} = \frac{\sum_{k=1}^{n} P(\omega_j|x_k)x_{ki}}{\sum_{k=1}^{n} P(\omega_j|x_k)} $$

where $P(\omega_j|x_k)$ is obtained using the “old” model parameters.

5. *(optional)* Expectation maximisation can be used in factor analysis. In factor analysis the observations $\mathbf{x}$ is assumed to be generated by a process of the form

$$ \mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{v} $$

where $\mathbf{z}$ is a $p$-dimensional zero-mean identity covariance matrix noise, $\mathbf{A}$ is a $d \times p$ matrix ($d > p$), and $\mathbf{v}$ is a $d$-dimensional. The parameters to estimate are the matrix $\mathbf{A}$ and the diagonal covariance matrix of $\mathbf{v}$. A set of $n$ independent training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are available. The mean of the data is known to be zero, $\mathbf{0}$.

Show that the PDF of the generative model may be written as

$$ p(\mathbf{x}_i|\mathbf{A}, \Sigma) = \mathcal{N}(\mathbf{x}_i; \mathbf{0}, \mathbf{A}\mathbf{A}' + \Sigma) $$

A set hidden variables, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are introduced. They indicate the values of $\mathbf{z}$ associated with each training sample. Show that

$$ p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{A}, \Sigma) = \mathcal{N}(\mathbf{x}_i; \mathbf{A}\mathbf{z}_i, \Sigma) $$

Using the previous two results, show that the posterior probability of the hidden variables given the current model parameters, $\mathbf{A}$ and $\Sigma$ is

$$ p(\mathbf{z}_i|\mathbf{x}_i, \mathbf{A}, \Sigma) = \mathcal{N}(\mathbf{z}_i; \beta\mathbf{x}_i, \mathbf{I} - \beta\mathbf{A}) $$
where
\[ \beta = A'(\Sigma + AA')^{-1} \]

Derive the re-estimation formulae for the new model parameters \( \hat{A} \) and \( \hat{\Sigma} \).

Describe why factor analysis is a useful tool for covariance modelling and data visualisation.

**Single Layer Perceptrons**

6. The standard single layer perceptron is used to discriminate between two classes. There are two simple techniques for generalising this to a \( K \) class problem. The first is to build a set of pairwise classifiers i.e. \( \omega_i \) versus \( \omega_j, j \neq i \). The second is to build a set of classifiers of each class versus all other classes i.e. \( \omega_i \) versus \( \{\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \omega_K\} \). Compare the two forms of classifier in terms of training and testing computational cost. By drawing a specific example with \( K = 3 \) show that both forms of classifier can result in an “ambiguous” region i.e. no decision can be made. Describe how multiple binaries classifiers may be trained so that no ambiguous regions exist.

**Multi-Layer Perceptrons**

7. † A multi-layer perceptron consists of \( d \) inputs, \( L \) hidden layers with \( M \) hidden units in each hidden layer, and \( K \) output nodes. Write down an expression for the total number of weights (including biases) in the network. Describe the factors that influence the number of hidden layers, the activation functions on the output layer, and the number of hidden units.

8. † For the logistic regression function, \( \phi(z) \), show that
\[ \frac{\partial}{\partial z} \phi(z) = \phi(z)(1 - \phi(z)) \]

How does the nature of the activation function affect the computational cost of the error-back propagation algorithm?

9. Consider the optimisation of a set of weights where the gradient of the error function with respect to the weight space is approximately constant. The following update rule is used
\[ w[\tau + 1] = w[\tau] + \Delta w[\tau] \]

where
\[ \Delta w[\tau] = -\eta \nabla E(w)_{w[\tau]} + \alpha \Delta w[\tau - 1] \]
Show that the effect of the momentum term is to increase the effective learning rate from $\eta$ to $\frac{\eta}{1-\alpha}$.

What is the effective learning rate for a region where the gradient descent scheme is oscillating about the real solution?

10. * The Hessian is a useful matrix for use in the optimisation of the weights of multi-layer perceptrons.

(a) Describe how the Hessian may be used for optimising the weights of a multi-layer perceptron. Discuss the limitations for the practical implementation of such schemes.

(b) For the least squares error function

$$ E = \frac{1}{2} \sum_{p=1}^{n} (y(x_p) - t(x_p))^2 $$

show that the elements of the Hessian matrix can be expressed as

$$ \frac{\partial^2 E}{\partial w_{ij} \partial w_{lk}} = \sum_{p=1}^{n} \frac{\partial y(x_p)}{\partial w_{ij}} \frac{\partial y(x_p)}{\partial w_{lk}} + \sum_{p=1}^{n} (y(x_p) - t(x_p)) \frac{\partial^2 y(x_p)}{\partial w_{ij} \partial w_{lk}} $$

For the case of well trained, sufficiently powerful, network, with an infinitely large training set, show that at the minimum the second term may be ignored. This is called the outer-product approximation.

(c) The Hessian after the $N^{th}$ data point is approximated by

$$ H_N = \sum_{p=1}^{N} g^{(p)}(g^{(p)})' $$

where

$$ g^{(p)} = \nabla E|_{w[t]} $$

By using the equality

$$ (A + BC)^{-1} = A^{-1} - A^{-1}B(I + CA^{-1}B)^{-1}CA^{-1} $$

where $I$ is the identity matrix, show that

$$ H_{N+1}^{-1} = H_{N}^{-1} - \frac{H_{N}^{-1}g^{(N+1)}(g^{(N+1)})'H_{N}^{-1}}{1 + (g^{(N+1)})'H_{N}^{-1}g^{(N+1)}} $$

Why is this a useful approximation to estimate the inverse Hessian during multi-layer perceptron training.
Radial Basis Functions

11. Consider the general form of a radial basis function

\[ y_k(x) = \sum_{m=1}^{M} w_{km} \phi_m(x) + w_{k0} \]

Show that by appropriately transforming the input space, \( x \), that a system with basis functions of the form

\[ \phi_j(x) = \exp \left( -\frac{1}{2}(x - \mu_j)'\Sigma^{-1}(x - \mu_j) \right) \]

may be written as

\[ \phi_j(x) = \exp \left( -\frac{||x - \mu_j||^2}{2\sigma^2} \right) \]

where \( x = Ax \). Describe how the transform parameters may be obtained.