4 Consider the k-means clustering algorithm which seeks to minimise the cost function

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} \|x_n - m_k\|^2$$

where m_k is the mean (centre) of cluster k, x_n is data point n, $s_{nk} = 1$ signifies that data point n is assigned to cluster k, and there are N data points and K clusters.

(a) Assume that the cluster assignments s_{nk} have all been determined, under the constraint that each data point must be assigned to one cluster, that is, $\sum_k s_{nk} = 1$ for all n, and $s_{nk} \in \{0,1\}$ for all n and k. Now derive the value of the means $\{m_k\}$ which minimise the cost C above, and give an interpretation in terms of the k-means algorithm. [30%]

(b) Give an interpretation of the k-means algorithm in terms of a probabilistic model. Describe three generalisations based on this probabilistic model. [40%]

(c) You are applying the k-means algorithm to a large collection of images, where most of the images are not labelled, but you have labels for a few of the images (e.g. "cat", "dog", "person", "car"). You would like to modify your k-means algorithm so that images with the same label are always in the same cluster, and images with different labels are never in the same cluster. Describe a modified version of the algorithm that would achieve this. [30%]

END OF PAPER

4 Consider a binary classification problem where the data $\mathscr{D}$ consists of N data points, $\mathscr{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, x_n is a real scalar and $y_n \in \{0, 1\}$, and the goal is to predict class labels y for new x.

Assume a very simple logistic classification model in which the class labels were produced independently and identically from the following model:

$$P(y_n = 1 | x_n, a, b) = \sigma(ax_n + b)$$

where σ is the logistic function, $\sigma(z) = \frac{1}{1+\exp(-z)}$, and a and b are the parameters of the classifier.

(a) Write down the likelihood of a and b for the data $\mathscr{D}$ and describe an algorithm to optimise this likelihood as a function of a and b. [40%]

(b) Consider a data set consisting of only two data points, $\mathscr{D} = \{(-2, 0), (3, 1)\}$. For this data set, describe the set of parameters which classify both data points correctly with probability greater than 0.5. Furthermore, what is the maximum achievable likelihood value? Describe the set of parameters which achieve this maximum. [40%]

(c) Explain how Bayesian learning of the parameters might give more reasonable inferences about a and b from the data set in part (b) than maximum likelihood (ML) and how the Bayesian predictions about future labels differ from the ML predictions. [20%]

END OF PAPER

Version 4

4 Consider a regression problem where the data $\mathscr{D}$ consists of N data points, $\mathscr{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, and you are trying to predict y given x.

(a) Assume a very simple linear regression model:

$$y_n = ax_n + \varepsilon_n$$

where the noise term ε_n is Gaussian with mean zero and variance σ^2. Derive the maximum likelihood (ML) estimates of a and σ^2. [40%]

(b) Now consider a more complex model:

$$y_n = ax_n + bx_n^2 + c + \varepsilon_n$$

with parameters a, b, c and noise variance σ^2 as before. Do you expect the maximum of the likelihood for this model to be lower or higher than for the model in part (a)? Explain your answer. Do you expect the ML estimate of the value of σ^2 to be lower, higher, or the same as in part (a)? Explain your answer. [30%]

(c) Now consider a model like in part (a) but where the noise terms ε_n have a Laplacian distribution,

$$p(\varepsilon_n) = \frac{1}{2\sigma} \exp\left\{-\frac{1}{\sigma}|\varepsilon_n|\right\}$$

instead of a Gaussian distribution. Explain when using such a Laplacian noise model might be a good idea. [30%]

END OF PAPER

4 Consider the k-means clustering algorithm which seeks to minimise the cost function

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} \|x_n - m_k\|^2$$

where m_k is the mean (centre) of cluster k, x_n is data point n, $s_{nk} = 1$ signifies that data point n is assigned to cluster k, and there are N data points and K clusters.

(a) Given all the means m_k, and the constraint that each data point must be assigned to one cluster (that is, $\sum_{k=1}^{K} s_{nk} = 1$ for all n, and $s_{nk} \in \{0, 1\}$ for all n and k), derive the value of the assignments $\{s_{nk}\}$ which minimise the cost C and give an interpretation in terms of the k-means algorithm. [30%]

(b) You would like to automatically learn the number of clusters K from data. One possibility is to minimise the cost C as a function of K. Explain whether this is a good idea or not, and what the solution to this minimisation is. [30%]

(c) Consider an algorithm for clustering high-dimensional data which first performs a principal components analysis (PCA) dimensionality reduction on the data, and then runs k-means on the low dimensional projection of the data. Will this result in the same clustering of the data as running k-means on the original high-dimensional data? Explain your answer. [40%]

END OF PAPER

4 Consider a dataset of observations $\mathscr{D} = \{(\mathbf{x}_n, y_n)\}$ where $n = 1, \ldots, N$, and N is the total number of data points. $\mathbf{x}_n$ is a two dimensional vector. A regression model of the following form is to be trained using the following form of regression

$$y_n = \mathbf{a}^{\mathrm{T}}\mathbf{x}_n + \varepsilon_n$$

where ε_n is independent zero-mean Gaussian noise with variance σ^2.

(a) Write down the log-likelihood $\log(p(y_1, \ldots, y_N | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{a}, \sigma^2))$ in terms of $y_1, \ldots, y_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{a}, \sigma^2$. [20%]

(b) Show that the maximum likelihood estimate of the regression parameters, $\hat{\mathbf{a}}$, can be expressed in the following form

$$\hat{\mathbf{a}} = \mathbf{C}^{-1}\mathbf{B}$$

You should clearly state the forms of the two matrices $\mathbf{C}$ and $\mathbf{B}$. The following equality may be useful

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^{\mathrm{T}}\mathbf{A} + \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

for any square matrix $\mathbf{A}$ and vector $\mathbf{x}$. [50%]

(c) A non-linear transformation $\phi(\mathbf{x}_n)$ is applied to the observations $\mathbf{x}_n$. The size of the resulting vector $\phi(\mathbf{x}_n)$ is d. Regression based on these transformed data points is then performed. Now

$$y_n = \mathbf{a}^{\mathrm{T}}\phi(\mathbf{x}_n) + \varepsilon_n$$

where ε_n is again independent zero-mean Gaussian noise with variance σ^2. Briefly discuss how the performance of the regression process and the estimation of the regression parameters may be impacted as the size of the transformed features, d, increases. [30%]

END OF PAPER

4 Consider a binary classification problem with scalar real-valued observations x, and class labels $y \in \{0, 1\}$. Assume a model with parameters θ where

$$p(y = 1|x, \theta) = \frac{1}{1 + e^{(-\theta x + 1/2)}}$$

(a) Describe the online learning rule for learning the parameter θ assuming the learning algorithm receives one data point at a time. [40%]

(b) Consider a data set $\mathscr{D}$ consisting of three data points: $(x_1 = 0,\ y_1 = 1)$, $(x_2 = -1,\ y_2 = 0)$, and $(x_3 = 1,\ y_3 = 1)$. Compute the likelihood for the parameters θ given this data set $\mathscr{D}$. [30%]

(c) Characterise the solution(s) to the maximum likelihood estimate of θ in part (b) above. Discuss properties of these solution(s), indicating any problems with the result and possible ways of resolving those problems. [30%]

END OF PAPER

4 Consider a binary classification problem with scalar real-valued observations x, and class labels $y \in \{0, 1\}$. Assume that $p(x|y = 0)$ is a Gaussian distribution with mean 0 and variance 2, and $p(x|y = 1)$ is a Gaussian distribution with mean 1 and variance 2. Furthermore, assume that $p(y = 0) = p(y = 1) = 1/2$.

(a) Compute the probability that given an observation $x = 2$, its corresponding class label is $y = 1$. [30%]

(b) Derive the general expression for $p(y = 0|x)$ as a function of x, and discuss how this relates to logistic classification. [40%]

(c) Now assume that you fit a maximum likelihood Gaussian distribution $p(x|y = 0)$ with mean μ_0 and variance σ_0^2 to the observed data with label $y = 0$, and similarly you fit a separate maximum likelihood Gaussian distribution $p(x|y = 1)$ with mean μ_1 and variance σ_1^2 to the observed data with label $y = 1$.

Describe several ways in which the above procedure differs from maximum likelihood logistic classification, paying particular attention to the role of the variances and likelihood that is being optimised. [30%]

END OF PAPER

4 Consider the k-means clustering algorithm which seeks to minimise the cost function

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} \|x_n - m_k\|^2$$

where m_k is the mean (centre) of cluster k, x_n is data point n, $s_{nk} = 1$ signifies that data point n is assigned to cluster k, and there are N data points and K clusters.

(a) Given all the assignments $\{s_{nk}\}$, derive the value of m_k which minimises the cost C and give an interpretation in terms of the k-means algorithm. [30%]

(b) Give a probabilistic interpretation of k-means and describe how it can be generalised to unequal cluster sizes (number of data points per cluster) and non-spherical (elongated) clusters as shown in Fig. 1 below. [30%]

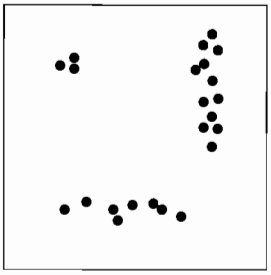

Fig. 1

(c) In many real-world applications, data points arrive sequentially and one wants to cluster them as they come in. Devise a sequential variant of the k-means algorithm which takes in one data point at a time and updates the means $\{m_1, \ldots, m_K\}$ sequentially without revisiting previous data points. Describe your sequential algorithm. [40%]

END OF PAPER

4 Consider a data set of pairs of observations $\mathscr{D} = \{(x_n, y_n)\}$ where $n = 1, \ldots, N$ and N is the total number of data points. Assume we wish to learn a regression model

$$y_n = ax_n + \varepsilon_n$$

where ε_n is independent zero-mean Gaussian noise with variance σ^2.

(a) Write down the log likelihood $\log p(y_1, \ldots, y_N | x_1, \ldots, x_N, a, \sigma^2)$ in terms of $y_1, \ldots, y_N, x_1, \ldots, x_N, a, \sigma^2$. [30%]

(b) Assume the following data set of $N = 4$ pairs of points

$$\mathscr{D} = \{(0,1), (1,2), (2,0), (3,4)\}$$

Solve for the maximum likelihood estimates of a and σ^2. [40%]

(c) Assume the same data set, but instead a regression model that predicts x given y:

$$x_n = by_n + \varepsilon_n$$

Is the maximum likehood estimate of b equal to $\frac{1}{a}$? Explain why or why not, giving a derivation if necessary. [30%]

END OF PAPER

4 Consider a data set of pairs of observations $\mathscr{D} = \{(x_n, y_n)\}$ where $n = 1, \ldots, N$ and N is the total number of data points. Assume we wish to learn a regression model

$$y_n = ax_n + \varepsilon_n$$

where ε_n is independent zero-mean Gaussian noise with variance σ^2.

(a) Write down the log likelihood $\log p(y_1, \ldots, y_N | x_1, \ldots, x_N, a, \sigma^2)$ in terms of $y_1, \ldots, y_N, x_1, \ldots, x_N, a, \sigma^2$. [30%]

(b) Assume the following data set of $N = 4$ pairs of points

$$\mathscr{D} = \{(0,1), (1,2), (2,0), (3,4)\}$$

Solve for the maximum likelihood estimates of a and σ^2. [40%]

(c) Assume the same data set, but instead a regression model that predicts x given y:

$$x_n = by_n + \varepsilon_n$$

Is the maximum likehood estimate of b equal to $\frac{1}{a}$? Explain why or why not, giving a derivation if necessary. [30%]

END OF PAPER