Version MJFG/4

EGT3
ENGINEERING TRIPOS PART IIB

Monday 27 April 2026 14.00 to 15.40

Module 4F10

DEEP LEARNING AND STRUCTURED DATA

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 An interesting family of probability distributions for one-dimensional data may be described by the following equation

$$p(x|\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{f}(x)\right)$$

where $\boldsymbol{\alpha}$ is the vector of parameters associated with the distribution and $\mathbf{f}(x)$ is a function of the data point x that returns a vector of the same dimension as $\boldsymbol{\alpha}$.

(a) Show that if

$$\mathbf{f}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

then a univariate Gaussian distribution may be expressed in this form. Find expressions for Z and the elements of the vector $\boldsymbol{\alpha}$ in terms of the mean, μ, and variance, σ^2, of the Gaussian distribution in this case. [25%]

(b) Rather than using a single distribution, a mixture of distributions is to be used based on the function $\mathbf{f}(x)$ in Part (a). To reduce the total number of parameters one of the elements of the parameter vector is constrained to be the same for all components. Thus

$$p(x|\boldsymbol{\theta}) = \sum_{m=1}^{M} c_m \frac{1}{Z_m}\exp\left(\begin{bmatrix} \alpha_m \\ \lambda \end{bmatrix}^{\mathsf{T}}\mathbf{f}(x)\right)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M, \lambda\}$ and the component priors, c_1 to c_M, are known and fixed. The parameters of the distribution, $\boldsymbol{\theta}$, are to be trained on N independent samples of data, $x_1, \ldots, x_N$. Maximum Likelihood (ML) training is used to estimate the model parameters.

(i) Derive an expression for Z_m in terms of α_m and λ. [15%]

(ii) The first element of the parameter vector for component ω_m, α_m, is to be trained using Expectation Maximisation (EM). The following form of auxiliary function is to be used

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{m=1}^{M}\sum_{i=1}^{N} P(\omega_m|x_i, \boldsymbol{\theta}) \log(p(x_i|\omega_m, \hat{\boldsymbol{\theta}}))$$

Derive an update formula to find the value of α_m. [30%]

(iii) All of the model parameters (excluding the priors), $\boldsymbol{\theta}$, are now to be estimated. Derive an update formula for λ using the auxiliary function in Part (b)(ii). [20%]

(iv) Comment on the update formulae derived in Parts (b)(ii) and (b)(iii) compared to using EM with the component distributions parameterised using the mean and variance. [10%]

2 A Support Vector Machine (SVM) classifier is to be built for a two class problem. There are a total of N, d-dimensional, training samples $\mathbf{x}_1$ to $\mathbf{x}_N$ with associated labels y_1 to y_N where $y_i \in \{-1, 1\}$.

(a) The decision boundary for SVMs with a linear kernel has the form

$$\mathbf{w}^\mathsf{T}\mathbf{x} + b = 0$$

How does this expression change if a general kernel function $k(\mathbf{x}_i, \mathbf{x})$ is to be used? Why are SVMs suited to classification with kernel functions? [15%]

(b) Show that if a kernel can be expressed as the sum of kernels then the decision boundary can also be expressed as a sum of kernels. [10%]

(c) A polynomial kernel can be written in the following forms

$$k(\mathbf{x}_i, \mathbf{x}) = \left(\mathbf{x}_i^\mathsf{T}\mathbf{x} + a\right)^c = \sum_{i=0}^{c} \left(\frac{c!}{(c-i)!i!}\right) a^{c-i} \left(\mathbf{x}_i^\mathsf{T}\mathbf{x}\right)^i$$

where $n!$ indicates the factorial of the non-negative integer n, and $0! = 1$.

(i) Show that if $d = 2$, $c = 3$ and $a = 0$, then the feature-space associated with the polynomial kernel, $\boldsymbol{\phi}(\mathbf{x})$, may be written in the form

$$\boldsymbol{\phi}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^3 & \sqrt{3}x_1^2x_2 & \sqrt{3}x_1x_2^2 & x_2^3 \end{bmatrix}^\mathsf{T}$$

[25%]

(ii) Show that the dimensionality of the feature-space for the polynomial kernel when $d = 2$ and $a = 1$ is $\left(\frac{(2+c)!}{2!c!}\right)$. [20%]

(d) The polynomial kernel is replaced by a Gaussian kernel.

(i) Give the equation for the Gaussian kernel. [10%]

(ii) How can you find the effective dimensionality of the feature-space for the Gaussian kernel? How do you expect this dimensionality to vary as you change the parameter associated with the Gaussian kernel? [20%]

3 A deep neural network is to be trained for a K-class problem. The activation function for the output layer is a softmax activation function, `soft`(). ReLU activation functions, `relu`(), are used for all L hidden layers. The number of nodes for layer l is $N^{(l)}$. None of the layers of the network make use of a bias vector. Thus for input $\mathbf{x}_i$ layer l can be expressed as

$$\mathbf{z}_i^{(l)} = \mathbf{A}^{(l)}\mathbf{x}_i^{(l)}, \quad \mathbf{x}_i^{(l+1)} = \texttt{relu}\left(\mathbf{z}_i^{(l)}\right)$$

and $\mathbf{x}_i^{(1)} = \mathbf{x}_i$. The network is to be trained using gradient descent with supervised training data $\mathcal{D} = \{\{\mathbf{x}_1, y_1\}, \ldots, \{\mathbf{x}_N, y_N\}\}$ where $\mathbf{x}_i$ is the observation vector for the i-th sample and $y_i \in \{\omega_1, \ldots, \omega_K\}$. The model is to be trained using a cross-entropy criterion, `ce`(). A batch size of one is used.

(a) Give expressions for the two activation functions being used in this network, `soft`() and `relu`(). Why are these activation functions appropriate for this task? [15%]

(b) The optimisation of the network parameters is to be based on the use of the Vector Jacobian Product (VJP).

(i) Show that the VJP for a softmax activation function, `soft_vjp`() can be expressed in the form

$$\texttt{soft\_vjp}(\mathbf{a}, \mathbf{b}) = \texttt{soft}(\mathbf{a}) \odot \mathbf{b} - \texttt{soft}(\mathbf{a})\texttt{sum}\left(\texttt{soft}(\mathbf{a}) \odot \mathbf{b}\right)$$

where `sum`($\mathbf{x}$) yields the sum of the elements of vector $\mathbf{x}$, and $\odot$ yields element-wise multiplication. You should clearly define $\mathbf{a}$ and $\mathbf{b}$. [35%]

(ii) Give the expression for the VJP, `mul_vjp`(), for the matrix multiplication associated with layer l, $\mathbf{z}_i^{(l)} = \mathbf{A}^{(l)}\mathbf{x}_i^{(l)}$. You should clearly define the inputs and the outputs for this function. [25%]

(iii) Assume that the VJP function for the ReLU activation function, `relu_vjp`(), is given. Briefly describe how the three VJP functions, `relu_vjp`(), `soft_vjp`() and `mul_vjp`(), can be used in the estimation of the model parameters ensuring that it is clear how the outputs from each of the VJPs is used in the estimation process. [25%]

4 A language model (LM) is to be trained using a set of N training sentences. For training sentence n consisting of T tokens, $\omega_{1:T}^{(n)} = \omega_1^{(n)}, \ldots, \omega_T^{(n)}$, the probability of the token sequence is computed using the following conditional probabilities

$$P(\omega_{1:T}^{(n)}) = P(\omega_1^{(n)}) \prod_{i=2}^{T} P(\omega_i^{(n)} | \omega_{1:i-1}^{(n)})$$

For the vocabulary of V tokens, an embedding layer that maps each token $\omega_i^{(n)}$ to a d-dimensional vector $\mathbf{x}_i^{(n)}$ is provided.

(a) Give the equation for an appropriate form of training criterion based on the N training sentences. You should be explicit about exactly what the LM is predicting and why this form of prediction is appropriate for LMs. Discuss any special token that should be included in the training data sequences. [20%]

(b) How can the trained LM be used to generate sentences? [15%]

(c) The LM is to be based on the transformer architecture, where the form of the transformer selected can be used as an encoder for the sequence of d-dimensional vectors $\mathbf{x}_{1:T}^{(n)} = \mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_T^{(n)}$ for training sentence n. L transformer blocks are connected together. The L-th transformer block is used to yield the language model prediction.

(i) Sketch the form of the transformer block that can be used. For each stage of the transformer block briefly describe, including equations, the purpose of that stage. [40%]

(ii) Describe, including equations, how the L-th transformer block can be used to yield the required language model probability. You should be clear how variable length sentences are handled. [15%]

(d) Briefly discuss any limitations in the training criterion described in Part (a) for training this form of model. [10%]

END OF PAPER

THIS PAGE IS BLANK