Version MJFG/2

EGT3
ENGINEERING TRIPOS PART IIB

Monday 22 April 2024 2 to 3.40

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 A classifier is to be built for a K-class problem. There are n, d-dimensional, training samples, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with class labels, $y_1, \ldots, y_n$. A 1-of-K coding for the class label is used so that each training example, $\mathbf{x}_i$, has a K-dimensional vector, $\mathbf{t}_i$, associated with it. Initially a single layer network with a softmax activation function is used so that the output for node j, $\phi_j(\mathbf{x})$, is

$$\phi_j(\mathbf{x}) = \frac{\exp(\mathbf{w}_j^{\mathsf{T}}\mathbf{x})}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathsf{T}}\mathbf{x})}$$

where the parameters of the classifier, $\boldsymbol{\lambda}$, are $\mathbf{w}_1, \ldots, \mathbf{w}_K$.

(a) The parameters of the classifier, $\boldsymbol{\lambda}$, are to be trained using the following criterion

$$L(\boldsymbol{\lambda}) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \log(\phi_k(\mathbf{x}_i))$$

where t_{ik} is element k of vector $\mathbf{t}_i$.

(i) Why is this criterion appropriate for training this form of network? [15%]

(ii) Derive an expression for the derivative of $L(\boldsymbol{\lambda})$ with respect to $\mathbf{w}_j$. How can this derivative be used to find the model parameters? [30%]

(b) A regularisation term is added to the criterion in Part (a). The parameters are now estimated based on the following function

$$F(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}) - a \sum_{k=1}^{K} \mathbf{w}_k^{\mathsf{T}}\mathbf{w}_k$$

where a is a fixed scalar value.

(i) Discuss why this form of expression may yield an estimate of λ that generalises better. How does the value of a influence the estimate of $\boldsymbol{\lambda}$? [15%]

(ii) Derive an expression for the derivative of $F(\boldsymbol{\lambda})$ with respect to $\mathbf{w}_j$. [15%]

(c) Instead of adding a regularisation term, an additional layer is added to the network between the input and the existing layer. This additional layer has a linear activation function, $\phi(x) = x$. The number of nodes in the additional layer is b.

(i) Discuss how this additional layer might improve the generalisation of the network, and any constraints on the value of b for this layer to be useful. [15%]

(ii) Discuss how this additional layer will alter the training of the network, if a similar criterion to Part (a) is used. [10%]

2 A large language model (LLM) 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

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

Two forms of LLM are considered, one based on Recurrent Neural Networks (RNNs) and the second using self-attention mechanisms. 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. The average number of tokens per sentence in the training data is μ_T and the variance in the number of tokens per sentence is σ_T^2.

(a) How can the trained LLM be used to generate sentences? You should include the equation for how the LLM conditional probability distribution is used. Discuss any special token that should be included in the training data sequences. [15%]

(b) Give the form of training criterion, in terms of the LLM conditional probability distribution, that can be used to train the LLM on the N training sequences. [5%]

(c) Initially the LLM based on RNNs is analysed.

(i) Briefly describe an RNN-based architecture that can be used for the LLM. You should discuss in detail how the token history is represented and the form of the output layer. [20%]

(ii) Derive an estimate of the approximate computational cost of training the RNN. You need only consider a single layer RNN architecture and should clearly define the size of any network layer that you use. [15%]

(d) The LLM based on self-attention mechanisms is to be analysed.

(i) Briefly describe an architecture that can be used for the LLM. You should discuss in detail how the token history is represented. [20%]

(ii) Give an estimate of the computational cost of training this LLM. You need only consider a single attention mechanism and should clearly define the size of any layers that you use. [15%]

(e) State any other considerations beyond computational cost that should be taken into account when selecting between the two LLM architectures. [10%]

3 (a) Describe the concept of kernel functions, indicating why they are useful for a range of classification tasks. [10%]

(b) A data scientist considers using a max-margin classifier with kernel

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathsf{T}}\mathbf{y}/(||\mathbf{x}||\ ||\mathbf{y}||)$$

where $\mathbf{x}$ and $\mathbf{y}$ are feature vectors and $||\cdot||$ denotes the Euclidean norm.

(i) Prove that k is a valid kernel. [10%]

(ii) The data scientist receives the following dataset with 4 points shown in Fig.1:

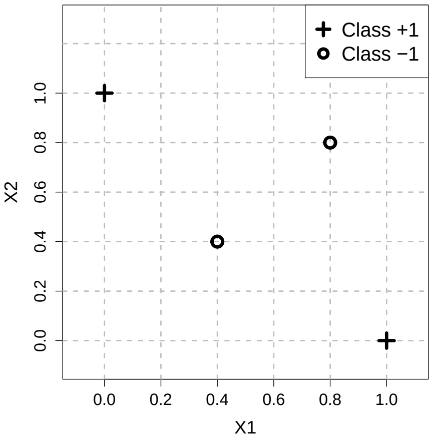

Fig. 1

Draw a plot of this dataset after mapping each original data-point to the feature space representation implied by the kernel k. On the plot that you just made, draw the decision boundary of a linear hard-margin SVM classifier. A sketch will be enough. Justify your answers. [20%]

(iii) Draw the original data, as in the figure above, together with the resulting decision boundary from the previous classifier (a sketch will be enough). Justify your decisions.
Hint: Consider where the feature map $\phi(\cdot)$ implied by the kernel $k(\mathbf{x}, \mathbf{y})$ above maps points to in 2D space. [25%]

(c) The data scientist considers then using a Gaussian kernel. For the one-dimensional case, a Gaussian kernel with unit scale is given by

$$k'(x, y) = \exp\left\{-(x-y)^2/2\right\}$$

where x and y are scalar input features.

(i) Show that the corresponding two-dimensional Gaussian kernel with unit scale can be obtained from the product of two one-dimensional kernels of the form of k'. [15%]

(ii) Show that k' has an associated feature map of infinite dimension given by

$$\boldsymbol{\phi}(x) = f(x)\left[1, \frac{x}{\sqrt{1!}}, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \ldots\right]$$

where $f(x)$ is a non-linear function of x that outputs a scalar. What is the form of $f(x)$? [20%]

4 A mixture of members of the exponential family of probability distributions is to be trained. The distribution for sample $\mathbf{x}$ has the form

$$p(\mathbf{x}|\alpha) = \sum_{m=1}^{M} c_m p(\boldsymbol{x}|\alpha_m) = \sum_{m=1}^{M} c_m \left(\frac{1}{Z_m}\exp\left(\boldsymbol{\alpha}_m^{\mathsf{T}}\mathbf{f}(\mathbf{x})\right)\right)$$

where $\boldsymbol{\alpha}_m$ is the vector of parameters associated with the m-th component of the distribution and $\mathbf{f}(\mathbf{x})$ is a function of the d-dimensional data point $\mathbf{x}$ that returns a vector of the same dimension as $\boldsymbol{\alpha}_m$. The parameters of this distribution, $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_M$, are to be trained on n independent samples of data, $\mathbf{x}_1,\ldots,\mathbf{x}_n$. The prior probabilities, c_1 to c_M, are known and not re-estimated. Maximum Likelihood (ML) training is used to estimate the model parameters.

(a) What expression must be satisfied by Z_m for $p(\mathbf{x}|\alpha_m)$ to be a valid probability density function? [10%]

(b) What is the training data log-likelihood using this mixture distribution. [10%]

(c) The parameters of the model, $\alpha_1,\ldots,\alpha_M$, are to be estimated using Expectation Maximisation (EM). The following form of auxiliary function is to be used

$$Q(\alpha,\hat{\alpha}) = \sum_{m=1}^{M}\sum_{i=1}^{n} P(m|\mathbf{x}_i,\alpha)\log(p(\mathbf{x}_i|m,\hat{\alpha}_m))$$

(i) Derive the set of statistics, in terms of $P(m|\mathbf{x}_i,\alpha)$ and $\mathbf{f}(\mathbf{x}_i)$, that must be extracted from the training data to allow the model parameters to be estimated. The set of statistics should be selected to minimise the number of values needed to store them. [25%]

(ii) Discuss, including equations, how the auxiliary function can be maximised. [20%]

(d) A member of the exponential family, a diagonal covariance matrix multivariate Gaussian distribution, is to be used for the component distributions.

(i) Show that the multivariate Gaussian distribution is a member of the exponential family. You should give expressions for the form of $\mathbf{f}(\mathbf{x})$, and the model parameters $\boldsymbol{\alpha}_m$ and Z_m in terms of the mean vector, $\boldsymbol{\mu}_m$, and diagonal covariance matrix, $\boldsymbol{\Sigma}_m$. [20%]

(ii) Compare finding the model parameters using EM and parameterising the distribution using $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ to the form based on $\boldsymbol{\alpha}_m$ in Part (c). [15%]

END OF PAPER