Version MJFG/3

EGT3
ENGINEERING TRIPOS PART IIB

Monday 25 April 2022 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 An M-component mixture model is to be used as the emission probability associated with the states of a J-state Hidden Markov Model (HMM). Each of the component distributions has the same form, a Bernoulli distribution. For the m-th component, ω_m, of state j, $\mathtt{s}_j$, the distribution may be written as

$$P(x|\mathtt{s}_j, \omega_m, \lambda_m) = \lambda_m^x (1-\lambda_m)^{(1-x)}$$

The component priors are the same for all states. A long sequence of n binary values, $x_1, \ldots, x_n$, is used to estimate the HMM parameters. The transition matrix, $\mathbf{A}$, for the HMM and the component priors, $c_1, \ldots, c_M$, are known and fixed. The parameters of the model are to be estimated using Maximum Likelihood (ML) estimation.

(a) Find an expression for the log-likelihood of the training data in terms of the model parameters, $\lambda_1, \ldots, \lambda_M$, and transition matrix $\mathbf{A}$. [15%]

(b) Expectation-Maximisation (EM) is to be used. The auxiliary function for this task may be written as

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = K + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{m=1}^{M} P(\mathtt{s}_j, \omega_m | x_i, \boldsymbol{\lambda}) \log(P(x_i|\mathtt{s}_j, \omega_m, \hat{\lambda}_m))$$

where $\boldsymbol{\lambda}$ is the set of all model parameters, $\lambda_1, \ldots, \lambda_M$. K is a constant including all terms that do not depend on $\boldsymbol{\lambda}$.

(i) Describe how EM is used to estimate the model parameters and the part played by the auxiliary function. Why is EM often used for this form of model? [15%]

(ii) Derive the update formula for finding the parameter estimates. [30%]

(c) A second binary variable, z, is now observed, so that there is a sequence of n observation pairs $\{x_1, z_1\}, \ldots, \{x_n, z_n\}$. The joint distribution for x and z for state j component m is given by

$$P(x, z|\mathtt{s}_j, \omega_m, \lambda_m, \boldsymbol{\alpha}) = \lambda_m^x (1-\lambda_m)^{(1-x)} \alpha_j^z (1-\alpha_j)^{(1-z)}$$

$\boldsymbol{\alpha}$ are the J additional parameters of the HMM. EM is again to be used to estimate an M-component mixture model.

(i) Briefly discuss how this modified distribution alters the HMM log-likelihood and ability of the HMM to model sequences. [15%]

(ii) Derive the update formula for $\boldsymbol{\alpha}$. [25%]

2 The parameters of a deep neural network are to be trained using a quadratic approximation to the error surface. The set of parameters associated with the network are denoted as the vector $\boldsymbol{\theta}$. The available training data are $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$ where $\boldsymbol{x}^{(i)}$ is the d-dimensional observation vector and $y^{(i)}$ is the continuous target value. The cost function with model parameters $\boldsymbol{\theta}$ is $E(\boldsymbol{\theta})$ where

$$E(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N}\left(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) - y^{(i)}\right)^2$$

and $f(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$ is the prediction of the network with parameters $\boldsymbol{\theta}$ and observation $\boldsymbol{x}^{(i)}$.

(a) The following quadratic approximation is to be used to estimate the weights

$$E(\boldsymbol{\theta}) \approx E(\boldsymbol{\theta}^{(\tau)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(\tau)})^{\mathsf{T}}\mathbf{b} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\tau)})^{\mathsf{T}}\mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\tau)})$$

(i) By considering a second-order Taylor series expansion about the point $\boldsymbol{\theta}^{(\tau)}$, the estimate of the model parameters at iteration τ, find expressions for $\mathbf{b}$ and $\mathbf{A}$. [15%]

(ii) Derive an expression for the value of $\boldsymbol{\theta}$ that will minimise this quadratic approximation. Hence describe how this approximation can be used to train the network parameters. [25%]

(b) An approximation to $\mathbf{A}$ in Part (a) is proposed of the form

$$\mathbf{A} \approx \sum_{i=1}^{N}\left(\nabla f(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)\left(\nabla f(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)^{\mathsf{T}}; \qquad \nabla f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{x}^{(i)};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

(i) By considering the exact form of $\mathbf{A}$ in Part (a) for the cost function $E(\boldsymbol{\theta})$ used to train the network, when is this a good approximation? [20%]

(ii) Show that the inverse of $\mathbf{A}$ can be expressed using the following recursion

$$\tilde{\mathbf{A}}_n^{-1} = \tilde{\mathbf{A}}_{n-1}^{-1} - \frac{\tilde{\mathbf{A}}_{n-1}^{-1}\left(\nabla f(\boldsymbol{x}^{(n)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)\left(\nabla f(\boldsymbol{x}^{(n)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)^{\mathsf{T}}\tilde{\mathbf{A}}_{n-1}^{-1}}{1 + \left(\nabla f(\boldsymbol{x}^{(n)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)^{\mathsf{T}}\tilde{\mathbf{A}}_{n-1}^{-1}\left(\nabla f(\boldsymbol{x}^{(n)};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{(\tau)}}\right)}; \qquad \mathbf{A}^{-1} \approx \tilde{\mathbf{A}}_N^{-1}$$

The following matrix equality may be useful

$$(\mathbf{D} + \mathbf{CF})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{C}(\mathbf{I} + \mathbf{FD}^{-1}\mathbf{C})^{-1}\mathbf{FD}^{-1}$$

where $\mathbf{I}$ is an Identity matrix and all matrix inverses exist. [20%]

(c) Briefly discuss the issues with using the form of quadratic approximation in Part (a) and how the approximation in Part (b) can be helpful. [20%]

3 A system to confirm the identity of a speaker, *speaker verification*, is to be implemented using Support Vector Machines (SVMs). The speech data from a speaker is parameterised using a d-dimensional feature vector extracted every 10 milliseconds. A separate SVM is trained for each of the S speakers in the database. For each speaker multiple utterances that are known to come from the speaker, the *enrolment data*, are available in the training database.

(a) The system is to use a *Fisher kernel* based on an M-component Gaussian mixture model (GMM). The GMM is trained on data from a large number of speakers. The following form of the Fisher kernel is used

$$k(\mathbf{X}^{(i)}, \mathbf{X}^{(s)}) = \left[\nabla_\mu \log(p(\mathbf{X}^{(i)}|\boldsymbol{\theta}))\right]^{\mathsf{T}} \left[\nabla_\mu \log(p(\mathbf{X}^{(s)}|\boldsymbol{\theta}))\right]$$

where $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(s)}$ are $T^{(i)}$ and $T^{(s)}$ length sequences of feature vectors, and ∇_μ indicates the derivative with respect to all the GMM component mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M$. $p(\mathbf{X}^{(i)}|\boldsymbol{\theta})$ is the likelihood of sequence $\mathbf{X}^{(i)}$ using the GMM with parameters $\boldsymbol{\theta}$

$$p(\mathbf{X}^{(i)}|\boldsymbol{\theta}) = \prod_{j=1}^{T^{(i)}} p(\mathbf{x}_j^{(i)}|\boldsymbol{\theta})$$

where $\mathbf{x}_j^{(i)}$ is the j^{th} vector in the $T^{(i)}$ length sequence $\mathbf{X}^{(i)}$.

(i) Briefly describe how an SVM with the Fisher kernel can be used in a speaker verification task. [20%]

(ii) Derive an expression for the feature-space associated with the Fisher kernel in terms of the component means and variances. What is the dimensionality of the feature-space? [30%]

(b) The Fisher kernel is replaced by a *sequence kernel* which has the form

$$k(\mathbf{X}^{(i)}, \mathbf{X}^{(s)}) = \sum_{j=1}^{T^{(i)}} \sum_{k=1}^{T^{(s)}} k^{\mathsf{s}}\left(\mathbf{x}_j^{(i)}, \mathbf{x}_k^{(s)}\right)$$

$k^{\mathsf{s}}()$ is either a linear kernel, or a Gaussian kernel.

(i) Under what conditions will this sequence kernel with $k^{\mathsf{s}}()$ being a linear kernel yield the same classifier as a Fisher kernel? [25%]

(ii) Give the form of the Gaussian kernel. Compare the sequence kernel using this Gaussian kernel and the Fisher kernel for speaker verification. You should discuss the computational cost and strengths and weaknesses of the two forms. [25%]

4 A neural network based sentiment classification system is to be trained on a corpus of film reviews. There are N film reviews and each review is labelled with one of K sentiment labels. Each of the words in a review is mapped to a d-dimensional vector representation, so the ith, L length, review will be mapped to the L length sequence $\mathbf{X}^{(i)} = \{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_L^{(i)}\}$. Thus the available training data, $\mathcal{D}$, are $\mathcal{D} = \{\{\mathbf{X}^{(1)}, y^{(1)}\}, \ldots, \{\mathbf{X}^{(N)}, y^{(N)}\}\}$ where $y^{(i)} \in \{\omega_1, \ldots, \omega_K\}$ and ω_j indicates the sentiment class. It is decided to train the network in the following configuration. The classification of an unknown mapped review, $\mathbf{X}^\star$, is based on

$$\mathbf{h} = \mathbf{f}(\mathbf{X}^\star); \qquad \mathbf{t}^\star = \mathbf{g}(\mathbf{h})$$

where $\mathbf{h}$ is an l-dimensional, fixed-length, vector, and the estimate of the sentiment, $\hat{\omega}^\star$, is obtained from the K-dimensional vector $\mathbf{t}^\star$.

(a) Describe an appropriate form of neural network for the function $\mathbf{g}()$. For this network, what is a suitable training criterion to estimate the model parameters based on $\mathcal{D}$? You should clearly describe all terms in the training criterion and why you think the form is appropriate. [25%]

(b) An attention mechanism is used for the function $\mathbf{f}()$. The query that is used to determine the relevance of a mapped word to sentiment classification, for word j of the review $\mathbf{X}^{(i)}$, is the mapped word vector $\mathbf{x}_j^{(i)}$. Describe, including appropriate equations, how an attention mechanism can be used for the function $\mathbf{f}()$ using review $\mathbf{X}^{(i)}$ as an example. You should describe *two* forms of possible attention that can be used, contrasting the two forms. [25%]

(c) To improve the performance of the system the attention mechanism of Part (b) is replaced by a *multi-head attention mechanism*.

(i) Describe how a multi-head attention mechanism can be used for $\mathbf{f}()$ again using review $\mathbf{X}^{(i)}$ as an example. You should include equations for the multi-head attention mechanism and the form of $\mathbf{h}$ that results from this process. [20%]

(ii) Compare this form of multi-head attention mechanism with the attention mechanism in Part (b). What is the advantage of using a multi-head attention mechanism? [15%]

(iii) It is proposed to use *dropout* when training the parameters of the multi-head attention mechanism. Briefly describe how dropout could be applied and whether you think it will improve the system performance. [15%]

END OF PAPER

THIS PAGE IS BLANK