Version MJFG/3

EGT3
ENGINEERING TRIPOS PART IIB

Monday 24 April 2023 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 Gaussian mixture model (GMM) with diagonal covariance matrices is to be used as the emission probability associated with the states of a J emitting state Hidden Markov Model (HMM). The parameters of the emission distributions for all J states are constrained to be the same. The feature vector is d-dimensional. A long sequence of N training vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, is available to estimate the model parameters. 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 component priors, $c_1, \ldots, c_M$, the known HMM transition matrix, $\mathbf{A}$, and the component parameters. For this model does the use of an HMM give any additional information over using a GMM? [15%]

(b) Expectation-Maximisation (EM) is to be used to find the Gaussian component means. The auxiliary function for this problem can be expressed as

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m=1}^{M} P(\mathtt{s}_j, \omega_m | \mathbf{x}_i, \boldsymbol{\theta}) \log(p(\mathbf{x}_i | \mathtt{s}_j, \omega_m, \hat{\boldsymbol{\theta}}))$$

where $\boldsymbol{\theta}$ is the set of all the model parameters and $\hat{\boldsymbol{\theta}}$ the parameters to be estimated. Constant terms, and terms related to estimating the HMM transition matrix, have been ignored in this expression.

(i) Show that the update formula for the mean of the m-th component is

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} P(\mathtt{s}_j, \omega_m | \mathbf{x}_i, \boldsymbol{\theta}) \mathbf{x}_i}{\sum_{i=1}^{N} \sum_{j=1}^{J} P(\mathtt{s}_j, \omega_m | \mathbf{x}_i, \boldsymbol{\theta})}$$

Ensure that your notation is clearly defined in the derivation. [20%]

(ii) The diagonal covariance matrices for all components of the model are restricted to be the same. Derive an expression to estimate this single covariance matrix $\boldsymbol{\Sigma}$. [20%]

(c) The form of the model is now changed so that the means of all the components are restricted to be the same, $\boldsymbol{\mu}$. Additionally the covariance matrices are scalar multiples of a single, diagonal, covariance matrix $\boldsymbol{\Sigma}$. Thus the mean for component m is $\boldsymbol{\mu}$ and the covariance matrix is $\alpha_m \boldsymbol{\Sigma}$.

(i) Derive the update formula for the value of α_m. Discuss any issues that need to be considered when estimating α_m. [30%]

(ii) Discuss the differences between using this form of model compared to using the form discussed in Part (b). [15%]

2 A machine translation system is to be trained to translate from French into German. Supervised training data is provided for this task, where the German translation for each French sentence is provided.

(a) Initially the vocabulary of the system is limited so that the system is trained to predict one of the K possible translated sentences, $\omega_1, \ldots, \omega_K$. Thus for each training example there is the pair of the French word-sequence and the translated sentence label.

(i) An attention-based architecture is to be used for this task. Briefly discuss a suitable form of network that can be used. You should include equations to illustrate the translation of the L-length French word sequence $\omega^{\mathsf{f}}_{1:L} = \omega^{\mathsf{f}}_1, \ldots, \omega^{\mathsf{f}}_L$ into the German sentence label, ω_k. [25%]

(ii) Define a suitable training criterion to train the model parameters, justifying your answer and clearly defining all symbols. [15%]

(b) The system is now extended so that it can handle more general French to German translations. The training data is now modified to comprise translation pairs of French word-sequences and the translated German word-sequence. An *auto-regressive* translation process is to be used where for the German translation $\omega^{\mathsf{g}}_{1:J} = \omega^{\mathsf{g}}_1, \ldots, \omega^{\mathsf{g}}_J$

$$P(\omega^{\mathsf{g}}_{1:J}|\omega^{\mathsf{f}}_{1:L}) = \prod_{i=1}^{J} P(\omega^{\mathsf{g}}_i|\omega^{\mathsf{g}}_{1:i-1}, \omega^{\mathsf{f}}_{1:L})$$

An encoder for the French words is provided so that the word-sequence $\omega^{\mathsf{f}}_1, \ldots, \omega^{\mathsf{f}}_L$ is mapped to embeddings $\mathbf{h}_1, \ldots, \mathbf{h}_L$. An attention-based approach is to be used to propagate information from the encoded word sequence to the decoder.

(i) Briefly describe how the encoded French word sequence can be used to predict the i-th translated German word. You should clearly describe, including equations, how information is propagated from the encoded French word sequence and the previous $i-1$ translated German words, $\hat{\omega}^{\mathsf{g}}_{1:i-1}$, and a suitable attention mechanism that can be used. You should define all symbols in the equations. [40%]

(ii) The model parameters are to be estimated by maximising the probability of each German word given the correct (from the reference) translation for the previous words. Give the expression that should be maximised in this case and comment on the advantages and disadvantages of such an approach. [20%]

3 (a) Give a definition of *support vector* in an SVM with a hard margin (i.e. C, the parameter that controls the trade-off between the slack variable penalty and the margin, is set to ∞). [15%]

(b) A data scientist fits a hard margin SVM classifier to a 2-D dataset using the following non-linear mapping of the data into a 5-D feature space:

$$\mathbf{x} = [x_1, x_2]^{\mathrm{T}} \rightarrow \phi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2]^{\mathrm{T}} . \quad (1)$$

The resulting SVM classifier is $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b$, with $\mathbf{w} = [1, 2, 0, 2, 1]^{\mathrm{T}}$ and $b = 0$.

(i) What is the distance of the data point $\phi(\mathbf{x}) = [1, 0.5, -2, 0.5, 1]^{\mathrm{T}}$ to the classifier's decision boundary in feature space? [15%]

(ii) What is the magnitude of the margin in feature space for this classifier? [15%]

(iii) Write down the kernel function $k_1(\mathbf{x}, \mathbf{x}')$ associated with the mapping in (1). [15%]

(iv) Instead of $k_1(\mathbf{x}, \mathbf{x}')$, the data scientist considers using the Gaussian kernel

$$k_2(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{1}{2s}(\mathbf{x} - \mathbf{x}')^{\mathrm{T}}(\mathbf{x} - \mathbf{x}')\right\} , \quad (2)$$

where s is tuned using validation data. Assuming that plenty of data is available, when will this kernel be preferred to the original one? [15%]

(c) The data scientist trains a hard margin linear SVM on a dataset comprising point, $\mathbf{x}_i$, and target, t_i, pairs. There are four training pairs with values: $\mathbf{x}_1 = [1, 4]^{\mathrm{T}}$, $t_1 = +1$; $\mathbf{x}_2 = [2, 3]^{\mathrm{T}}$, $t_2 = +1$; $\mathbf{x}_3 = [4, 5]^{\mathrm{T}}$, $t_3 = -1$; and $\mathbf{x}_4 = [5, 6]^{\mathrm{T}}$, $t_4 = -1$. The points are shown in Fig. 1.

(i) What is the weight vector $\mathbf{w}$ and the bias term b of the resulting classifier? Hint: the slope of the decision boundary is -1, so $\mathbf{w}$ is orthogonal to $[1, -1]^{\mathrm{T}}$ [15%]

(ii) The data scientist switches to a soft margin SVM with a linear kernel and $C = 0.1$. At the optimal solution, $\mathbf{w} = [-0.36, -0.28]^{\mathrm{T}}$ and $b = 2.48$. What are the associated values for the slack variables, $\{\xi_n\}_{n=1}^4$, and Lagrange multipliers, $\{\mu_n\}_{n=1}^4$ and $\{a_n\}_{n=1}^4$? You should ensure that you clearly define all symbols. [10%]

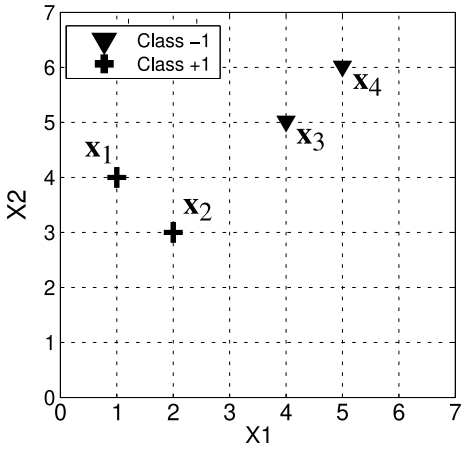
Class –1
Class +1
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
X1
X2
0
1
2
3
4
5
6
7

Fig. 1

4 A classifier is to be constructed using Bayes' decision rule for a binary classification problem. A d-dimensional observation feature-vector is used. The true feature vector for classes ω_1 and ω_2 are always $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ respectively. However, due to the measurement process there is a significant level of noise on any measurements taken from these feature vectors. The noise from each measurement is independent of the noise from other measurements and is known to have zero mean for both classes. The prior probabilities for the two classes are known to be equal.

(a) State Bayes' decision rule for binary classification tasks. [10%]

(b) Initially it is assumed that the noise on the measurements is multivariate Gaussian distributed. Furthermore the covariance matrix for the noise, $\boldsymbol{\Sigma}$, is assumed to be a scaled identity matrix with scaling factor α. A generative model classifier is trained on a large quantity of supervised training data.

(i) For a test measurement $\boldsymbol{x}^\star$ derive an expression for the posterior probability of assigning this measurement to class ω_1 for this classifier. You should simplify the form of your expression where possible. [25%]

(ii) To improve the accuracy of the classifier, measurements for each test sample are repeated five times, all five samples are known to come from the same class. The noise for each of these measurements is independent. By considering the probability of generating all five samples, $\boldsymbol{x}_1^\star, \ldots, \boldsymbol{x}_5^\star$, from classes ω_1 or ω_2, derive an expression for the posterior probability of class ω_1. Compare the derived posterior probability to the form derived in Part (b)(i). [20%]

(c) An ensemble of M deep-learning based classifiers is proposed to further improve the performance. Each classifier is trained to directly predict the posterior probabilities of the two classes for a test measurement $\boldsymbol{x}^\star$, $P(\omega_1|\boldsymbol{x}^\star)$ and $P(\omega_2|\boldsymbol{x}^\star)$.

(i) Briefly describe *one* approach for generating an ensemble of deep classifiers from the available training data. Discuss any limitations of the approach. [15%]

(ii) How can predictions from the ensemble be combined together to predict the class posterior for a single measurement $\boldsymbol{x}^\star$? You should motivate your answer. [10%]

(iii) If the ensemble now has five measurements to predict the class posterior, $\boldsymbol{x}_1^\star, \ldots, \boldsymbol{x}_5^\star$, how can you use these multiple samples with the deep-learning ensemble? You should clearly motivate your approach and discuss any limitations. [20%]

END OF PAPER