Version JMHL/2

EGT3
ENGINEERING TRIPOS PART IIB

Monday 28 April 2025 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) Define and compare generative and discriminative classifiers in the context of supervised learning and provide examples of each type of classifier. [15%]

(b) Consider a binary classification problem where the data for each class is generated by multivariate Gaussian distributions with different means μ_1, μ_2 and a shared covariance matrix Σ. The prior probabilities of the two classes are $P(C_1)$ and $P(C_2)$.

(i) Derive the decision boundary for this problem. [15%]

(ii) Assuming the true class-conditional distributions are known, write an expression for the probability of classification error as a function of $P(C_1)$ and $P(C_2)$ and the parameters of the Gaussian distributions. [15%]

(c) A discriminative model is used instead for the same classification problem, represented as a logistic regression model:

$$P(C_1|\mathbf{x}) = \frac{1}{1 + \exp\{-\mathbf{w}^{\mathrm{T}}\mathbf{x} - b\}}$$

where $\mathbf{w}$ is a vector of parameters, $\mathbf{x}$ is an input feature vector and b is a scalar bias parameter.

(i) Given a dataset $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$ with pairs of input feature vectors $\mathbf{x}_n$ and corresponding class labels $y_n \in \{C_1, C_2\}$, derive the the log-likelihood function used to estimate the parameters $\mathbf{w}$ and b of this model. [15%]

(ii) The cost of a false positive (predicting C_1 when the true label is C_2) is C_{FP} The cost of a false negative (predicting C_2 when the true label is C_1) is C_{FN}. There is no cost associated with making a correct prediction. The decision rule classifies a sample $\mathbf{x}$ as class C_1 when $p(C_1|\mathbf{x}) \geq \theta$. Derive the optimal decision threshold θ that minimises the expected cost. [20%]

(d) Assume that, in addition to the labelled dataset $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, an extra set with K unlabelled inputs $\{\mathbf{x}_k^U\}_{k=1}^{K}$ is available. Show how these extra inputs can be used to improve the estimation of the generative classifier from part (b). [20%]

2 An M-component Gaussian mixture model (GMM) is to be used to model N data points $X = \{x_1, \ldots, x_N\}$, where each data point x_n is in $\mathbb{R}^d$. The model parameters, $\boldsymbol{\theta}$, are the mean vectors $\mu = \{\mu_1, \ldots, \mu_M\}$, covariance matrices $\Sigma = \{\Sigma_1, \ldots, \Sigma_M\}$, and prior probabilities $\pi = \{\pi_1, \ldots, \pi_M\}$ of the M components.

(a) $Z = \{z_1, \ldots, z_N\}$ are latent variables where $z_n = m$ if x_n is sampled from the m-th component of the GMM. The joint distribution of X and Z is given by

$$p(X, Z; \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{m=1}^{M} [\pi_m \mathcal{N}(x_n; \mu_m, \Sigma_m)]^{\mathbb{I}(z_n = m)}$$

where $\mathbb{I}(\cdot)$ is the indicator function taking value 1 when its input is true and 0 otherwise. Use this expression to derive the form of the marginal log-likelihood function $\log p(X; \boldsymbol{\theta})$. [20%]

(b) Assume the covariance matrices Σ are isotropic with $\Sigma_m = \sigma^2 I$, where σ^2 is known and fixed. The goal is to estimate the means μ using the Expectation-Maximisation (EM) algorithm.

(i) Starting with the general auxiliary function

$$Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \mathbb{E}_{Z \sim P(Z|X;\boldsymbol{\theta}^{(k)})} \left[\log p(X, Z; \boldsymbol{\theta}^{(k+1)}) \right]$$

derive the simplified form of $Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$ with respect to the means μ. Clearly define all quantities and variables involved. [20%]

(ii) Obtain the update equation for $\mu_m^{(k+1)}$ by maximising $Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$. Express the result in terms of the responsibilities $\gamma_{nm} = P(z_n = m \mid x_n, \boldsymbol{\theta}^{(k)})$. [20%]

(c) Suppose the data is incomplete, and instead of X, only noisy binary observations $Y = \{y_1, \ldots, y_N\}$ are available, where $y_n = 1$ if x_n belongs to a specific region $R \subset \mathbb{R}^d$, and $y_n = 0$ otherwise.

(i) Write the modified joint distribution $p(X, Z, Y; \boldsymbol{\theta})$ incorporating the binary observations. Note that y_n is conditionally independent of other variables given x_n. [10%]

(ii) Write the distribution $p(Y; \boldsymbol{\theta})$ obtained by integrating out the missing variables X and summing out the latent variables Z. Write your expression in terms of the probability mass that each Gaussian component assigns to the region R. [10%]

(iii) Write down the new form for the responsibilities $\gamma_{nm} = P(z_n = m \mid y_n, \boldsymbol{\theta}^{(k)})$ in this missing data scenario. [20%]

3 (a) Explain the concept of a support vector machine (SVM), emphasising the role of support vectors and the significance of the margin in classification tasks. [15%]

(b) Consider a binary classification problem using an SVM with a linear kernel. The optimisation problem for the primal form is given as

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n\,,$$

subject to

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 - \xi_n\,, \quad \xi_n \geq 0\,, \quad n = 1,\ldots,N$$

where $\mathbf{w}$ and b are the weights and bias, C is a regularisation parameter, and ξ_n are slack variables for each data point n formed by a feature vector $\mathbf{x}_n \in \mathbb{R}^d$ and corresponding class label $y_n \in \{-1, 1\}$.

(i) Write the Lagrangian for this optimisation problem, identifying all the Lagrange multipliers. [15%]

(ii) State the KKT conditions for this problem and explain how they can be used to identify the data points that are support vectors. [15%]

(c) Now consider the dual formulation of the SVM problem with kernel $k(\mathbf{x}_n, \mathbf{x}_m)$:

$$\max_{\alpha_1,\ldots,\alpha_N} \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m k(\mathbf{x}_n, \mathbf{x}_m)$$

subject to

$$\sum_{n=1}^{N}\alpha_n y_n = 0 \quad 0 \leq \alpha_n \leq C\,, \quad n = 1,\ldots,N$$

(i) Explain how the optimal values of $\alpha_1,\ldots,\alpha_N$ determine the classifier's output and the decision boundary. [15%]

(ii) Show that the dual problem is a convex optimisation problem. Use properties of the kernel matrix K and the objective function to support your argument. [20%]

(iii) Assume a dataset with $N = 3$ one-dimensional inputs, labelled $(x_1 = -1, y_1 = 1)$, $(x_2 = 0, y_2 = -1)$, $(x_3 = 1, y_3 = 1)$. A linear kernel $k(x, y) = xy$ is used with regularisation parameter $C = 1$. Solve for the dual variables α_1, α_2 and α_3. [20%]

4 A neural network will be used to translate sentences from English to French. The input is a sequence of T English words, with corresponding word embeddings $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The output is a sequence of T' French words, with corresponding word embeddings $\mathbf{y}_1, \ldots, \mathbf{y}_{T'}$. All the embedding vectors are in $\mathbb{R}^d$. $\mathbf{x}_T$ and $\mathbf{y}_{T'}$ are special embeddings representing the end of English and French sentences. The model is trained by sampling pairs of English and French sentences and predicting a word from the French sentence given the English words and the previous French words. There are a total of V_{French} different French words in the data.

(a) The transformer model is proposed for this task.

(i) What output layer should be used and how many parameters will it have? [10%]

(ii) Explain the roles of positional encoding and self-attention in this model. [20%]

(b) In the encoder:

(i) Given input embeddings, the self-attention mechanism computes new embeddings by calculating attention scores from query and key matrices and then using these to linearly combine a transformation of the original embeddings. Calculate the computational cost of the self-attention mechanism in terms of scalar multiplication operations for a pair of English and French sequences of length T and T' and embedding dimension d. Ignore the cost of computing non-linearities, e.g. softmax. Discuss the implications for very long sequences. [20%]

(ii) Multi-head attention is used to improve the performance of self-attention. Explain the purpose of multi-head attention and describe how the outputs of multiple attention heads are combined. [15%]

(iii) Describe how masking is used to ensure autoregressive generation during training. [15%]

(c) A recurrent neural network (RNN) architecture is considered as an alternative model. Compare and contrast transformers with RNNs for this task. [20%]

END OF PAPER

THIS PAGE IS BLANK