Version MJFG/4

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 5 May 2021 1.30 to 3.10

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 and at the top of each answer sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
You are allowed access to the electronic version of the Engineering Data Books.

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

The time taken for scanning/uploading answers is 15 minutes.

Your script is to be uploaded as a single consolidated pdf containing all answers.

1 An M-component Gaussian mixture model (GMM) is to be used to model N data points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of dimension d. The model parameters, $\boldsymbol{\theta}$, are the mean vectors $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M\}$, covariance matrices $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M\}$ and prior probabilities $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_M\}$ of the M components.

(a) $\mathbf{Z} = \{z_1, \ldots, z_N\}$ are latent variables associated with $\mathbf{X}$, where $z_n = m$ if $\mathbf{x}_n$ was sampled from the m-th component in the GMM. The joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is

$$p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{m=1}^{M} \pi_m^{\mathbf{1}(z_n=m)} \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)^{\mathbf{1}(z_n=m)}$$

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

(b) The mean vectors, $\boldsymbol{\mu}$, and covariance matrices, $\boldsymbol{\Sigma}$, are known and fixed so that the only unknown parameters are the prior probabilities $\boldsymbol{\pi}$. The priors are to be estimated using the Expectation-Maximisation (EM) algorithm.

(i) Show that the EM auxiliary function can be expressed as

$$Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{n=1}^{N} \left[\sum_{m=1}^{M} a_{nm} \log(\pi_m^{(k+1)}) \right] + \text{constant}$$

The value of a_{nm} should be clearly defined, as well as the meaning of the variables. You can use the fact that the EM auxiliary function for a model with latent variables $\mathbf{Z}$, observed data $\mathbf{X}$ and parameters θ is given by

$$Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}^{(k)}) \log(p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}^{(k+1)}))$$

All elements of the auxiliary function that are not dependent on the component priors are combined into the constant term. [20%]

(ii) Obtain the EM update equation for $\pi_m^{(k+1)}$ by optimising the Lagrangian function

$$\mathcal{L}(\boldsymbol{\pi}^{(k+1)}, \lambda) = Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) + \lambda \left(\sum_{m=1}^{M} \pi_m^{(k+1)} - 1 \right)$$

Write your update equation in terms of a_{nm} from part (b)(i) . [20%]

(cont.

(c) For a particular task the dimensionality of the observation is 1, $d = 1$, so that $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_M\}$ and $\boldsymbol{\Sigma} = \{\sigma_1^2, \ldots, \sigma_M^2\}$ contain scalars. Instead of directly observing the random variables $\mathbf{X}$ only the output of a classifier is given for each observation, $\mathbf{S} = \{s_1, \ldots, s_N\}$, where s_n is 1 if $x_n \geq t$ and 0 otherwise for a particular threshold t. The joint distribution can now be expressed as

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{S}; \boldsymbol{\theta}, t) = \prod_{n=1}^{N} \left[\left(\prod_{m=1}^{M} \pi_m^{\mathbf{1}(z_n=m)} \mathcal{N}(x_n; \mu_m, \sigma_m^2)^{\mathbf{1}(z_n=m)} \right) s_n^{\mathbf{1}(x_n \geq t)} (1 - s_n)^{\mathbf{1}(x_n < t)} \right]$$

Note $0^0 = 1$.

(i) Derive the log-likelihood function $\log(P(\mathbf{S}; \boldsymbol{\theta}, t))$ using the expression above. Note that, in this case, only $\mathbf{S}$ is observed and $\mathbf{X}$ and $\mathbf{Z}$ are unknown. [20%]

(ii) Discuss how EM can be used to estimate the component priors $\boldsymbol{\pi}$ for this task. [20%]

2 A classifier is to be constructed using generative models and Bayes' decision rule. The training data comprises d-dimensional feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and corresponding labels $y_1, \ldots, y_N$, where each label indicates one of K classes, $\omega_1, \ldots, \omega_K$. For class ω_k the class-conditional probability distribution is given by $p(\mathbf{x}|\omega_k)$ and the prior by $P(\omega_k)$.

(a) All the parameters of the classifier are known. A test observation $\mathbf{x}^\star$ is to be labelled using the classifier. Decision rules are to be derived for two forms of loss functions. Any approximations and assumptions in defining the rules should be stated.

(i) The loss when the classifier makes an error is 1, and 0 when the classifier is correct. Give an expression for the decision rule for the test sample that minimises the expected loss. You should express the decision rule in terms of the class priors and class-conditional probability distributions. [15%]

(ii) The loss now depends on the correct class. When the classifier makes an error and the correct class was ω_k the loss is l_k, and again 0 when the classifier is correct. Derive a new decision rule that again minimises the expected loss in terms of l_k, the class priors and class-conditional probability distributions. [15%]

(b) The classifier partitions the feature space into K regions $\Omega_1, \ldots, \Omega_K$, such that the classifier labels an observation $\mathbf{x}$ that is in region Ω_k as ω_k. Derive an expression for the expected loss of the classifier using the loss in part (a)(ii). You should clearly state any approximations being made. [15%]

(c) Multivariate Gaussian distributions are to be used as the class-conditional probability distributions.

(i) Show that the posterior of class ω_k given test observation $\mathbf{x}^\star$ can be written in the form of a softmax function using quadratic functions of $\mathbf{x}^\star$. [30%]

(ii) Discuss the advantages and disadvantages of estimating the prior and parameters of the multivariate Gaussian distribution for each class ω_k using the available training data and either: maximum likelihood estimation of the class-conditional probability distributions and priors; or directly minimising the expected loss from part (b). [25%]

3 A neural network system is to be designed to classify a speaker's emotion from a spoken sentence. The input is an utterance represented by a sequence of d-dimensional vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_T$, with each vector in the sequence representing 10 milliseconds of audio. The length of the sequence, T, varies from utterance to utterance. The output of the system, y, is one of 5 emotions (neutral, angry, sad, frustrated, happy) describing the speaker's emotion. It is proposed to use a network incorporating a single recurrent network layer, which is connected to a final classification layer. Two forms of activation function, $\phi(z)$, are considered for the recurrent layer ($\alpha > 0$):

$$(1)\ \phi(z) = \max(0, \alpha z); \qquad (2)\ \phi(z) = \begin{cases} \max(0, \alpha z); & z \leq 1 \\ \alpha; & z > 1 \end{cases}$$

(a) Describe an overall network structure for the emotion classification task incorporating a single unidirectional layer of recurrent units. You should clearly describe the input, recurrent layer and any further layers. Give *two* options for how the input to the final classification layer can be obtained from the recurrent layer, stating the advantages and limitations of your choice. [30%]

(b) The parameters of the network are to be trained using gradient-descent based optimisation.

(i) Sketch the two forms of recurrent layer activation function given. For the form of activation function in (1) compute the activation function output mean and variance if z is Gaussian distributed with mean of zero and variance of σ^2. Note $\int_0^\infty x\mathcal{N}(x; 0, \sigma^2)dx = \frac{\sigma}{2}\sqrt{\frac{2}{\pi}}$.

[20%]

(ii) Discuss how the results from part (b)(i) can be used to initialise the recurrent layer network parameters. You should give an appropriate form of parameter initialisation for each case, clearly motivating the form that you have selected.

[20%]

(iii) Which of the two forms of activation function is expected to be more sensitive to initialisation, justifying your answer? [15%]

(c) The system is required to operate in scenarios where there are significant levels of background noise. A separate network is available that generates a vector $\mathbf{n}$ that describes the background noise for a speaker's utterance. Describe how this vector can be used in the emotion classification system to improve performance. [15%]

4 (a) Give a definition of the margin of a linear classifier and indicate why classifiers with maximum margin are typically preferred. [10%]

(b) An alternative to the max margin classifier is the nearest-neighbour classifier, which assigns a new input vector $\mathbf{x}$ to the same class as that of the nearest input vector from the training set, where in the simplest case, the distance between two vectors $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ and $\mathbf{x}' = (x'_1, \ldots, x'_D)^{\mathrm{T}}$ is defined by their squared Euclidean distance, given by $\sum_{d=1}^{D}(x_d - x'_d)^2$.

(i) Write the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$ in terms of dot products between vectors. [15%]

(ii) Use the kernel trick to obtain the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$ in a non-linear feature space specified by a kernel function k. [15%]

(iii) When do you expect the nearest-neighbour classifier with kernel function k to outperform the original nearest-neighbour classifier? [15%]

(c) Let $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$ be a hard-margin SVM classifier trained on a dataset formed by input features $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^d$ and corresponding target variables $t_1, \ldots, t_N$ in $\{-1, 1\}$, such that $y(\mathbf{x}_+) = 1$ and $y(\mathbf{x}_-) = -1$ for positive and negative support vectors $\mathbf{x}_+$ and $\mathbf{x}_-$ in $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, respectively.

(i) Derive the magnitude M of the margin of this classifier as a function of $\mathbf{w}$. Justify your answer. [15%]

(ii) Describe how minimising $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ relates to maximising the margin. [10%]

(iii) Show that the margin magnitude M is also given by $M = 1/\sqrt{\sum_{n=1}^{N} a_n}$, where $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is the solution to the dual problem

$$\max_{\mathbf{a}} \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m \quad \text{s.t.} \quad \sum_{n=1}^{N} a_n t_n = 0, \quad \{a_n \geq 0\}_{n=1}^{N}.$$

Recall that the solution for $\mathbf{a}$, b and $\mathbf{w}$ satisfies that either $a_n = 0$ or $t_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) - 1 = 0$, for $n = 1, \ldots, N$, so that the objective function for the dual problem and the objective function for the original problem take the same value. [20%]

END OF PAPER