UNIVERSITY OF
CAMBRIDGE

ENGINEERING TRIPOS PART IIB

Tuesday 26 April 2005 2.30 to 4

Module 4F10

STATISTICAL PATTERN PROCESSING

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.*

There are no attachments.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

(TURN OVER

1 An interesting family of probability distributions for one-dimensional data may be described by the following equation

$$p(x|\boldsymbol{\alpha}) = \frac{1}{Z}\exp\left(\boldsymbol{\alpha}'\mathbf{f}(x)\right)$$

where $\boldsymbol{\alpha}$ is the vector of parameters associated with the distribution and $\mathbf{f}(x)$ is a function of the data point x that returns a vector of the same dimension as $\boldsymbol{\alpha}$.

(a) What expression must be satisfied by Z for this expression to be a valid probability density function? [10%]

(b) Show that if

$$\mathbf{f}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

then a univariate Gaussian distribution may be expressed in this form. Find expressions for Z and the elements of the vector $\boldsymbol{\alpha}$ in terms of the mean, μ, and variance, σ^2, of the Gaussian distribution in this case. [25%]

(c) Rather than using a single distribution, a mixture of distributions is to be used. This has the form

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

The parameters of the distribution, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m$, are to be trained on N independent samples of data, $x_1, \ldots, x_N$. The priors, c_1 to c_M, are known and not re-estimated. Maximum Likelihood (ML) training is used to estimate the model parameters.

(i) Write down an expression for the log-likelihood of the training data using this mixture distribution. [15%]

(ii) Differentiate the log-likelihood expression in part (c)(i) with respect to the parameters of component m, $\boldsymbol{\alpha}_m$. This should be expressed in terms of the component posterior $P(m|x_i)$ and the differential of Z_m. [25%]

(iii) The parameters of the model are to be trained using Expectation Maximisation (EM). Write down the form of auxiliary function that could be used in this case. Hence find the statistics that must be extracted from the training data to allow the model parameters to be estimated. [25%]

2 A linear classifier with parameter a of the form

$$y(x) = ax$$

is to be trained for a one-dimensional, two-class, problem. The data for each of the two classes, ω_1 and ω_2, is Gaussian distributed. For class ω_1 the mean is 0 and variance is 1. For class ω_2 the mean is 2 and the variance is 2. The priors for the two classes are known to be equal. There are N training examples for each of the classes.

(a) What is the general form of Bayes' decision rule for a two class problem? [10%]

(b) The linear classifier is to be trained using least squares estimation with target values of 0 for class ω_1 and 1 for class ω_2. A very large number of training examples, N, are available to estimate the classifier parameter. Calculate the value of a. [35%]

(c) A threshold of 0.5 on $y(x)$ is used to classify the data. Using the value of a estimated in part (b) calculate the probability of misclassifying a sample in terms of the Gaussian cumulative density function $F(x)$ where

$$F(x) = \int_{-\infty}^{x} \mathcal{N}(z; 0, 1) dz$$

[25%]

(d) What expression is satisfied by a point x that lies on the optimal decision boundary specified by the Bayes' decision rule? Using this expression obtain a new estimate of a that will reduce the probability of error with the threshold given in part (c). [30%]

3 A multilayer perceptron is to be trained using gradient-descent based algorithms. The set of weights associated with the network are denoted as the vector $\boldsymbol{\theta}$. An iterative procedure is commonly used to update the weight vector where at iteration $\tau+1$

$$\boldsymbol{\theta}^{(\tau+1)} = \boldsymbol{\theta}^{(\tau)} + \Delta\boldsymbol{\theta}^{(\tau)}$$

where $\boldsymbol{\theta}^{(\tau)}$ is the estimate of the model parameters at iteration τ. The value of the cost function with model parameters $\boldsymbol{\theta}$ is $E(\boldsymbol{\theta})$.

(a) What expression is used for $\Delta\boldsymbol{\theta}^{(\tau)}$ in the standard form of gradient descent? Briefly discuss the issues that must be considered when using this standard form. [20%]

(b) In order to improve the performance of the gradient descent optimisation, a second-order approximation is used. Here

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

(i) By considering a second-order Taylor series expansion about the point $\boldsymbol{\theta}^{(\tau)}$ 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. [20%]

(iii) What are the practical issues in using this form of optimisation approach? [15%]

(c) An alternative approach is to make a second-order approximation and assume that all the elements of $\boldsymbol{\theta}$ are independent. In addition only the gradient at each point is used. Using this form of approximation show that a suitable update for element i at iteration $\tau+1$ would use

$$\Delta\theta_i^{(\tau)} = \left(\frac{g_i^{(\tau)}}{g_i^{(\tau-1)} - g_i^{(\tau)}}\right)\Delta\theta_i^{(\tau-1)}$$

where

$$g_i^{(\tau)} = \left.\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}^{(\tau)}}$$

[30%]

4 A linear classifier is to be trained for a d-dimensional, two-class, problem. The available training data consists of N independent samples, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and associated class labels, $y_1, \ldots, y_N$, where $y_i \in \{-1, 1\}$. The data is known to be linearly separable.

(a) Briefly discuss the differences between the solutions that would be obtained for this problem using a support vector machine (SVM) with a linear kernel, and the perceptron algorithm for training the classifier. [15%]

(b) An SVM is to be trained on the data. The optimisation criterion to train the decision boundary is usually expressed in the form

$$\hat{\mathbf{w}} = \min_{\mathbf{w},b} \left\{ \frac{1}{2} ||\mathbf{w}||^2 \right\}$$

subject to all points satisfying a constraint.

(i) What constraint must be satisfied by all training examples when the maximum margin classifier has been obtained for this linearly separable case? [10%]

(ii) Derive an expression for the margin of a linear classifier for linearly separable data and show how minimising the optimisation criterion given above relates to maximising the margin. [25%]

(iii) Hence using Lagrange multipliers derive the conditions that must be satisfied at the solution to training the SVM. [20%]

(iv) How is b estimated for an SVM? [10%]

(c) A Gaussian kernel is to be used for both the SVM and perceptron algorithm trained classifier. Briefly discuss the form of the classification rule that should be used for both classifiers in this case. [20%]

5 A decision tree is to be built for a classification problem.

(a) As part of the training process for a decision tree a *node impurity* function is required.

(i) What general attributes should be satisfied by any node impurity function? How are node impurity functions used in building a decision tree? [20%]

(ii) For a K-class problem, the Gini impurity measure may be expressed in either of the two following ways:

$$\sum_{i \neq j} P(\omega_i)P(\omega_j); \quad \text{or} \quad 1 - \sum_{i=1}^{K} (P(\omega_i))^2$$

Describe how $P(\omega_i)$ should be calculated for a particular decision tree node. Show that these two expressions are equivalent and satisfy the attributes described in part (a)(i). [25%]

(b) The Gini impurity measure is to be used to train a decision tree for a two-class problem with 4-dimensional, binary valued, data. The training data for the two classes, ω_1 and ω_2, is shown below.

$$\begin{array}{lllll} \omega_1: & [0,0,0,0]' & [1,0,1,0]' & [1,1,0,0]' & [0,0,1,1]' \\ \omega_2: & [1,1,0,0]' & [1,1,1,1]' & [1,1,1,0]' & [0,1,1,1]' \end{array}$$

(i) Using changes in the Gini impurity measure, determine which of the first two elements of the feature-vector would be better used for the initial split. What is the change in the impurity measure in this case? [20%]

(ii) What is the lowest impurity measure that can be obtained using this data? [15%]

(c) For a particular task, the cost of misclassifying class ω_2 is twice that of class ω_1. How would this alter the decision tree training process? [20%]

END OF PAPER

ENGINEERING TRIPOS PART IIB

Tuesday 26 April 2005 9 to 10.30

Module 4F11

SPEECH PROCESSING

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.*

There are no attachments.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

(TURN OVER

1 The parameters of a Linear Prediction (LP) model are to be estimated from a speech signal $s(n)$.

(a) Write the p^{th} order linear prediction in terms of the predictor coefficients $a_1, \ldots, a_p$, and write the expression for the prediction error sequence, $e(n)$. [10%]

(b) Show how the filter transfer function $\frac{S(z)}{E(z)}$ describes a system in which the speech signal is modelled as the output of a linear system excited by the prediction error sequence. Draw the corresponding linear filter. [20%]

(c) Referring to the Source-Filter model of speech production, explain the role of the prediction error sequence. [10%]

(d) The speech signal is bandpass filtered to a 20-4000 Hz range. Assuming that four formants are present in the signal, suggest an appropriate value of p and justify your answer. [10%]

(e) Explain how the LP spectrum is derived from the LP coefficients. [10%]

(f) Describe how to extract formant information from the LP spectrum. Describe how to extract formant information directly from the LP polynomial. [25%]

(g) Discuss how formant extraction based on LP analysis differs from formant extraction based on a smoothed discrete Fourier transform of speech. [15%]

2 (a) The analysis stage of a channel vocoder is shown in Fig. 1.

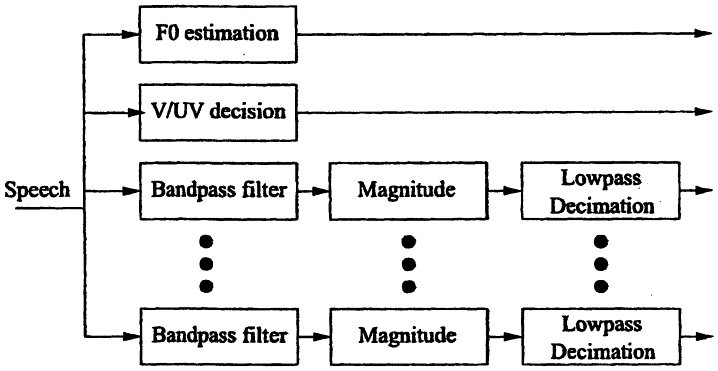

Fig. 1

(i) Draw the components of the synthesis stage. [15%]

(ii) Discuss the vocoder analysis and synthesis operations. [15%]

(iii) Discuss how overall synthesis quality could be improved by incorporating Mel frequency-scale filterbanks and exploiting auditory masking. [20%]

(b) Multipulse excitation is to be used for a linear prediction synthesis of speech. A single frame of N speech samples $s(n)$ is analysed to obtain a linear prediction filter with impulse response $h(n)$. The filter, and a single excitation impulse placed k samples from the start of the frame, are to be used to reconstruct the frame. By first finding the optimal amplitude of the impulse, show that the minimum sum squared reconstruction error is [35%]

$$\epsilon(k) = \left(\sum_{n=0}^{N-1} s^2(n)\right) - \frac{(\hat{r}_{sh}(k))^2}{\hat{r}_{hh}(0)}$$

$$\text{where} \quad \hat{r}_{sh}(k) = \sum_{n=0}^{N-1} s(n)h(n-k)$$

$$\text{and} \quad \hat{r}_{hh}(0) = \sum_{n=0}^{N-1} h^2(n-k)$$

(c) Compare multipulse excitation with the excitation in a CELP coder. [15%]

3 A training utterance, $\mathbf{y}_1 \dots \mathbf{y}_T$, will be used in estimating the parameters of an N-state left-to-right Hidden Markov Model (HMM), $\mathcal{M} = \{N, \{a_{i,j}\}, \{b_j(\cdot)\}\}$, where each state has a single Gaussian observation distribution.

(a) Suggest a procedure to initialise the HMM parameters prior to training. [10%]

(b) Following initialisation, Viterbi training will be performed.

(i) Explain how the Viterbi algorithm can be used to find

$$\Phi_j(t) = \max_{s_1,\dots,s_{t-1}} p(\mathbf{y}_1, \dots, \mathbf{y}_t, s_t = j|\mathcal{M})$$

where s_t is the state occupied at time t. State how the algorithm is initialised. [15%]

(ii) Explain how to obtain the most likely state sequence $s_1, \dots, s_T$. [10%]

(iii) Give the estimation equations for the parameters of the Gaussian distributions in terms of the segmented acoustic sequence. [20%]

(iv) Draw a flow chart describing the iterative application of the Viterbi training procedure. Suggest two different stopping criteria. [15%]

(c) After the Viterbi training of the single-Gaussian-per-state system has converged, the state output distributions are replaced by a mixture of Gaussians.

(i) Describe how Viterbi training can be used to estimate the parameters of the components of the mixture distributions. [20%]

(ii) Suggest a method to initialise this training procedure. [10%]

4 (a) Triphones are widely used in large-vocabulary speech recognition systems.

(i) Discuss why triphones are a good unit for use in a speech recognition system. [10%]

(ii) What issues must be addressed in using triphones in practical systems? [10%]

(b) Describe methods of state-level parameter tying for constructing triphones using

(i) Phonetic decision trees; [30%]

(ii) Bottom-up clustering. [20%]

In each case describe the basic operation of the technique, and list the advantages and disadvantages of the method for constructing triphone-based recognition systems.

(c) As part of the decision tree clustering procedure, a node splitting and merging cost must be defined. Single Gaussian distributions are used to model the data for the parent node p, and children nodes r and s. The change in likelihood for such a model can be computed from just the covariance matrices and the number of samples of data associated with p, r and s.

(i) Show how the covariance matrix at any node in the tree can be computed from knowledge of the state mean, covariance matrices and state occupation count of all triphone contexts, stating the assumptions that are made during this process. [20%]

(ii) Discuss the implications of using Gaussian mixtures to model each tree node during the decision tree clustering process. [10%]

5 (a) Define the term *perplexity* in statistical language modelling and show how perplexity may be computed given a language model and some test text. Explain why perplexity is a reasonable measure of language model quality. [20%]

(b) Describe a back-off N-gram language model, including parameter estimation using discounting and the computation of back-off weights. [30%]

(c) An N-gram language model has been trained on a corpus of text data. The training set perplexity is to be estimated solely from the language model parameters.

(i) Show how the perplexity of the training data can be computed for a unigram model. [15%]

(ii) Show how the training set perplexity calculation can be extended to a bigram model. [15%]

(d) Discuss how training set perplexity and test set perplexity differ. How would they vary with training set size, the number of language model parameters, and the use of discounting in language model estimation? [20%]

END OF PAPER