Engineering Tripos Part IIB FOURTH YEAR

Module 4F10: STATISTICAL PATTERN RECOGNITION

Examples Paper 1

Straightforward questions are marked †
Tripos standard (but not necessarily Tripos length) questions are marked ∗

Bayes Risk

1. In many pattern classification problems, one has the option either to assign the pattern to one of the c classes, or to reject it as being unrecognizable. If the cost to reject is not too high, rejection may be a desirable action. Let the cost of classification be defined as

$$\lambda(\omega_i|\omega_j) = \begin{array}{lll} 0 & \omega_i = \omega_j & (i.e.\ \text{(Correct classification)} \\ \lambda_r & \omega_i = \omega_0 & (i.e.\ \text{Rejection)} \\ \lambda_s & \text{Otherwise} & (i.e.\ \text{Substitution Error)} \end{array}$$

   Show that for minimum risk classification, the decision rule should associate a test vector $\boldsymbol{x}$ with class ω_i, if $P(\omega_i|\boldsymbol{x}) \geq P(\omega_j|\boldsymbol{x})$ for all j **and** $P(\omega_i|\boldsymbol{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise.

EM and Mixture Models

2. † For d-dimensional data compare the computational cost of calculating the log-likelihood with a diagonal covariance matrix Gaussian distribution, a full covariance matrix Gaussian distribution and an M-component diagonal covariance matrix Gaussian mixture models. Clearly state any assumptions made.

3. A 1-dimensional 2-component mixture distribution has a common fixed known variance = 1 and initial mean values $\mu_1 = 0$ $\mu_2 = 2$ and mixture weights $c_1 = c_2 = 0.5$.

   There is a data set of 9 training data points provided

$$-1.5,\ -0.5,\ 0.1,\ 0.3,\ 0.9,\ 1.3,\ 1.9,\ 2.3,\ 3.0$$

   (a) Calculate the log likelihood of the training data for the mixture distribution with the initial parameters.

   (b) Calculate updated values for the mean and mixture weights for 1 iteration of the E-M algorithm.

4. Consider an M component mixture model of d-dimensional binary data $\mathbf{x}$ of the form

$$p(\mathbf{x}) = \sum_{m=1}^{M} P(\omega_m) p(\mathbf{x}|\omega_m)$$

where the j^{th} component PDF has parameters $\lambda_{j1}, \ldots, \lambda_{jd}$ and

$$p(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} \lambda_{ji}^{x_i}(1-\lambda_{ji})^{1-x_i}$$

A set of training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are used to train the mixture model. Using the standard form of EM with mixture models show that the maximum likelihood estimate for the "new" parameters, $\hat{\lambda}_{ji}$, is given by

$$\hat{\lambda}_{ji} = \frac{\sum_{k=1}^{n} P(\omega_j|\mathbf{x}_k)x_{ki}}{\sum_{k=1}^{n} P(\omega_j|\mathbf{x}_k)}$$

where $P(\omega_j|\mathbf{x}_k)$ is obtained using the "old" model parameters.

5. $*$ A series of n independent, noisy, measurements are taken, $x_1, \ldots, x_n$. The noise is known to be Gaussian distributed with zero mean and unit variance. The "true" data is also known to be Gaussian distributed.

   (a) Find the maximum likelihood estimates of the mean, μ, and variance, σ^2, of the "true" data by equating the gradient to zero.

   (b) A latent variable z_i is introduced. It is the value of the noise for observation x_i. Show that the posterior probability of z_i given the current model parameters is

   $$p(z_i|x_i, \theta) = \mathcal{N}\left(z_i; \frac{(x_i-\mu)}{(1+\sigma^2)}, \frac{\sigma^2}{(1+\sigma^2)}\right)$$

   Using the expectation-maximisation algorithm derive re-estimation formulae for the mean, μ, and variance, σ^2. Show that the iterative estimation scheme for the mean converges to the correct answer, you may assume that the variance of the true data is known and fixed at σ^2.

   Discuss the merits of the two optimisation schemes for this task and for optimisation tasks in general.

   Product of Experts

6. $*$ *For parts of this question it is useful to use matlab/octave* A product of experts system is to be used for speech synthesis. The data is known to be generated from two classes ω_1 and ω_2. Four Gaussian experts are to be used. These experts are:

$$\begin{aligned} p(x_t|\omega_1) &= \mathcal{N}(x_t; 1, 1) \textbf{ Expert 1} \\ p(x_t - x_{t-1}|\omega_1) &= \mathcal{N}(x_t - x_{t-1}; 1, 1) \textbf{ Expert 2} \\ p(x_t|\omega_2) &= \mathcal{N}(x_t; 2, 1) \textbf{ Expert 3} \\ p(x_t - x_{t-1}|\omega_2) &= \mathcal{N}(x_t - x_{t-1}; -1, 1) \textbf{ Expert 4} \end{aligned}$$

   A sequence of 3 samples are to be generated. The first two are known to come from class ω_1, the final sample from class ω_2. The data is known to start in silence, which has a value of 0.

(a) Show that the overall sequence of observarions can be written in the following form

$$\mathbf{A}\boldsymbol{x} = \mathbf{A}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 - 0 \\ x_2 \\ x_2 - x_1 \\ x_3 \\ x_3 - x_2 \end{bmatrix}$$

(b) The transformed data, $\mathbf{A}\boldsymbol{x}$ is Gaussian distributed, so

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{1}{Z} p(\mathbf{A}\boldsymbol{x}|\boldsymbol{\theta}) \\ &= \frac{1}{Z}\mathcal{N}(\mathbf{A}\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

where Z is the appropriate normalisation term to ensure a valid PDF. Find expressions for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

(c) By using the following expression (or otherwise)

$$\exp\left(-\frac{1}{2}(\mathbf{A}\boldsymbol{x} - \boldsymbol{\mu})'(\mathbf{A}\boldsymbol{x} - \boldsymbol{\mu})\right) = \exp\left(-\frac{1}{2}\left(\boldsymbol{x}'\mathbf{A}'\mathbf{A}\boldsymbol{x} - 2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{x} + \boldsymbol{\mu}'\boldsymbol{\mu}\right)\right)$$

find the mean of the distribution of $\boldsymbol{x}$. How can this approach be used for speech synthesis? What does $\boldsymbol{x}$ look like if experts 2 and 4 are not used, set $\mathbf{A} = \mathbf{I}$ (an identity matrix)

Restricted Boltzmann Machine

7. A restricted Boltzmann machine is to be built where the observations, $\boldsymbol{x}$, are continuous variables and the hidden units, $\boldsymbol{h}$, are binary. The energy function has the following form:

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\theta}) = \sum_{i=1}^{d} \frac{(x_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{J} b_j h_j - \sum_{i,j} \frac{x_i}{\sigma_i} h_j w_{ij}$$

Show that the posterior probabilty of the hidden and observed variables can be expressed as

$$\begin{aligned} P(h_j = 1|\boldsymbol{x}, \boldsymbol{\theta}) &= \frac{1}{1 + \exp(-b_j - \sum_{i=1}^{d} \frac{x_1}{\sigma_i} w_{ij})} \\ p(x_i|\boldsymbol{h}, \boldsymbol{\theta}) &= \mathcal{N}(x_i; a_i + \sigma_i \sum_j h_j w_{ij}, \sigma_i^2) \end{aligned}$$

Why is this form of expression important when training Restricted Boltzmann machines?

Single Layer Perceptrons

8. The standard single layer perceptron is used to discriminate between two classes. There are two simple techniques for generalising this to a K class problem. The first is to build a set of pairwise classifiers i.e. ω_i versus ω_j, $j \neq i$. The second is to build a set of classifiers of each class versus all other classes i.e. ω_i versus $\{\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \omega_K\}$. Compare the two forms of classifier in terms of training and testing computational cost. By drawing a specific example with $K = 3$ show that both forms of classifier can result in an "ambiguous" region i.e. no decision can be made. Describe how multiple binaries classifiers may be trained so that no ambiguous regions exist.

Answers

3. (a) total log-likelihood of data (natural log (ln)) -15.302 (likelihood 2.262e-07); (b) $\hat{\mu}_1 = -0.0426$; $\hat{\mu}_2 = 1.878$; $\hat{c}_1 = 0.5266$; $\hat{c}_2 = 0.4734$

M.J.F. Gales
P.C. Woodland
Oct 2003 - Jan 2007