KTH Electrical Engineering

Exam in Pattern Recognition EN2202

Date: Thursday Oct 30, 2014, 14:00 – 19:00

Place: V11, V21.

Allowed: Beta, calculator with empty memory, one page handwritten note.

Grades: A: 31p; B: 27p; C: 23p; D: 20p; E: 17; of max 25p + 10p project bonus.

Language: English.

Results: Friday, Nov 14, 2014.

Review: Via scanned version

Contact: Saikat: 0738913581; Martin: 0702521154; Arun: 0727895538

Good Luck!

1 A signal source randomly selects a state $S = 1$ or $S = 2$ with equal probability. In both states the signal source generates stationary zero mean two-dimensional Gaussian source $X(n) = (X_1(n),\, X_2(n))^\top$ with independent elements over n, that is,

$$\mathcal{E}(X(n_1)X(n_2)^\top) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},\ n_1 \neq n_2.$$

In state 1 the covariance of every source element is given by C_1 and in state 2 the covariance is C_2. A two-category classifier receives and stores L consecutive source samples $(X(0), \cdots, X(L-1))$ and then guesses the state of the signal source.

(a) Design this classifier for minimum error probability, with $C_1 = \mathcal{E}(X(n)X(n)^\top|1) = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$ and $C_2 = \mathcal{E}(X(n)X(n)^\top|2) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. (3p)
Hint: Design a discriminant function to produce a single design variable, such that optimal classification can be obtained by a simple threshold mechanism with this feature variable as input.
Solution: As both source alternatives are equally probable, we use the Maximum Likelihood decision rule. We can define a single discriminant function simply as

$$\begin{aligned} g(\underline{\mathrm{x}}) &= \ln f_{\underline{\mathrm{X}}|S}(\underline{\mathrm{x}}|2) - \ln f_{\underline{\mathrm{X}}|S}(\underline{\mathrm{x}}|1) \\ &= \frac{1}{2}\sum_{n=0}^{L-1} \mathbf{x}(\mathbf{n})C_1^{-1}\mathbf{x}(\mathbf{n}) - \frac{1}{2}\sum_{n=0}^{L-1} \mathbf{x}(\mathbf{n})C_2^{-1}\mathbf{x}(\mathbf{n}) + \frac{1}{2}L\ln\left(\frac{\det(C_1)}{\det(C_2}\right) \\ &= \frac{1}{2}\sum_{n=0}^{L-1}\left(\frac{4}{15}x_1(n)^{\,2} + \frac{4}{15}x_2(n)^2 - \frac{2}{15}x_1(n)x_2(n)\right) \\ &- \frac{1}{2}\sum_{n=0}^{L-1}\left(\frac{2}{3}x_1(n)^{\,2} + \frac{2}{3}x_2(n)^2 - \frac{2}{3}x_1(n)x_2(n)\right) + \frac{1}{2}L\ln\left(\frac{15}{3}\right) \\ &= -\sum_{n=0}^{L-1}\left(\frac{1}{5}x_1(n)^{\,2} + \frac{1}{5}x_2(n)^2 - \frac{4}{15}x_1(n)x_2(n)\right) + \frac{1}{2}L\ln 5 \end{aligned} \tag{1}$$

(b) What is the probability that source $S = 1$ was active given observations $X(0) = (0,0)^\top$, $X(1) = (1,0)^\top$, and $X(2) = (0,2)^\top$, that is, $P(S = 1|\underline{\mathrm{X}} = ((0,0)^\top, (1,0)^\top, (0,1)^\top)$? (2p)
Solution: We compute the log likelihood values:

$$L_i = \ln f_{\underline{\mathrm{X}}|S}(\underline{\mathrm{x}}|i) = -\frac{1}{2}\sum_{n=0}^{L-1} \mathbf{x}(\mathbf{n})C_i^{-1}\mathbf{x}(\mathbf{n}) - \frac{L}{2}\ln\det C_i + const \tag{2}$$

Omitting common factors and substituting values, we have

$$L_1 = -(2/15)(0+1+0) - (2/15)(0+0+1) - (3/2)\ln 15 = -4.3287 \tag{3}$$

and

$$L_2 = -(1/3)(0+1+0) - (1/3)(0+0+1) - (3/2)\ln 3 = -2.3146 \tag{4}$$

then the conditional probability for $S = 1$ is:

$$P(S = 1|\underline{\mathrm{X}} = ((0,0)^\top, (1,0)^\top, (0,1)^\top) = \frac{e^{L1}}{e^{L_1} + e^{L_2}} \approx 0.118.$$

2 Determine for each of the following statements whether it is *true* or *false*, and give a brief argument for your choice: (1p each) (5p)

(a) If $\mathbf{x} \in \mathcal{R}^N$ is a multivariate Gaussian random variable, then $\mathbf{Ax} + \mathbf{b}$, where $\mathbf{A} \in \mathcal{R}^{N \times N}$ and $\mathbf{b} \in \mathcal{R}^N$, is also a Multivariate Gaussian random variable.

Solution: TRUE

(b) Let X_1 and X_2 be two zero-mean Gaussian variables with the same variance, if the variables are uncorrelated, i.e.

$$E[X_1 X_2] = 0,$$

then they are independent.

Solution: TRUE. For Gaussian variables, uncorrelated variables are independent. However, uncorrelated variables are not independent in general.

(c) The Backward algorithm gives a more accurate estimate of the initial state probability than the forward algorithm.

Solution: TRUE. The forward algorithm gives the probability $P(S_1 = i|x_1)$ while the backward algorithm gives the probability $P(S_1 = i|x_1, x_2, \ldots, x_T)$. Since the backward algorithm uses more data it gives a more accurate estimate.

(d) The Jeffries prior (the most non-informative prior) is always flat, i.e. the probability distribution is constant where it is non-zero.
Solution: FALSE. The Jeffries prior is given by $p(w) \propto \sqrt{I(w)}$, where $I(w)$ is the Fisher information. When $w = \sigma^2$ is the variance of a gaussian variable, then the Jeffires prior becomes $p(w) \propto 1/w$. So the Jeffries prior is not always flat.

(e) An irreducible Markov chain with a finite number of states cannot have a finite duration.
Solution: FALSE. A Markov chain is irreducible if one can go from any state to any other state in a finite number of steps with non-zero probability. The Markov chain with

$$A = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.4 & 0.5 & 0.1 \end{pmatrix},$$

is both finite duration and irreducible.

3 (GMM) Consider the distribution of a signal vector $\mathbf{x} \in \mathcal{R}^N$ assumed to be a Gaussian mixture model (GMM) as $p(\mathbf{x}) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(\mathbf{x}; \mathbf{m}_i, \mathbf{C}_i)$, where α_i, $\mathbf{m}_i$ and $\mathbf{C}_i$ have usual meanings.

(a) Evaluate the mean and covariance of $p(\mathbf{x})$ in terms of $\mathbf{m}_i$ and $\mathbf{C}_i$. (2p)
Solution: We solve $\mathcal{E}(\mathbf{x}) = \mathbf{m} = \int_{\mathbf{x}} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{I} \alpha_i\, \mathbf{m}_i$. Then we solve

$$\begin{aligned}
\mathcal{E}((\mathbf{x} - \mathcal{E}(\mathbf{x}))(\mathbf{x} - \mathcal{E}(\mathbf{x}))^\top) &= \mathcal{E}(\mathbf{x}\mathbf{x}^\top) - \mathbf{m}\mathbf{m}^\top \\
&= \int_{\mathbf{x}} \mathbf{x}\mathbf{x}^\top \sum_{i=1}^{I} \alpha_i \mathcal{N}(\mathbf{x}; \mathbf{m}_i, \mathbf{C}_i) d\mathbf{x} - \mathbf{m}\mathbf{m}^\top \\
&= \sum_{i=1}^{I} \alpha_i \int_{\mathbf{x}} \mathbf{x}\mathbf{x}^\top \mathcal{N}(\mathbf{x}; \mathbf{m}_i, \mathbf{C}_i) d\mathbf{x} - \mathbf{m}\mathbf{m}^\top \\
&= \sum_{i=1}^{I} \alpha_i (\mathbf{C}_i + \mathbf{m}_i \mathbf{m}_i^\top) - \mathbf{m}\mathbf{m}^\top \\
&= \sum_{i=1}^{I} \alpha_i (\mathbf{C}_i + \mathbf{m}_i \mathbf{m}_i^\top) - (\sum_{i=1}^{I} \alpha_i\, \mathbf{m}_i)(\sum_{i=1}^{I} \alpha_i\, \mathbf{m}_i)^\top
\end{aligned}$$

(b) Assume that there are J independent sources where each source distribution is a Gaussian mixture as $p(\mathbf{x}|j) = \sum_{i=1}^{I} \alpha_{ij} \mathcal{N}(\mathbf{m}_{ij}, \mathbf{C}_{ij})$, $j = 1, 2, \ldots, J$. For example, we may consider a task of speaker classification where each speaker is an independent source and each source has a Gaussian mixture distribution. Assume the prior probability of jth source is denoted by Pr(j). Using second order statistics of Gaussian mixture distributions, find the analytical expression of MAP classification rule. (2p)
Hint: Second order statistics are mean and covariance of any distribution. Think which distribution is fully expressed via mean and covariance, and use that distribution for second order statistics based modeling of true Gaussian mixtute distributions. The final analytical expression of MAP rule should be in terms of all parameters.

Solution: Gaussian distribution is fully expressed via means and covariances. So we need to use Gaussian distribution to model true GMM distribution as follows

$$p_{\mathcal{M}}(\mathbf{x}|j) = \mathcal{N}\left(\mathbf{x}; \sum_{i=1}^{I} \alpha_{ij}\, \mathbf{m}_{ij}, \sum_{i=1}^{I} \alpha_{ij} \left(\mathbf{C}_{ij} + \mathbf{m}_{ij}\mathbf{m}_{ij}^\top\right) - (\sum_{i=1}^{I} \alpha_{ij}\, \mathbf{m}_{ij}) \left(\sum_{i=1}^{I} \alpha_{ij}\, \mathbf{m}_{ij}\right)^\top\right).$$

Then the MAP classification rule is

$$d(\mathbf{x}) = \arg\max_j \; p_{\mathcal{M}}(\mathbf{x}|j) \mathrm{Pr(j)}.$$

4 (Expectation Maximization) We observe samples $\underline{x} = (x_1, x_2, \ldots, x_N)$ from the mixture distribution

$$p(x) = q\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2} + (1-q)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2},$$

where σ^2 is known but μ_1, μ_2 and q are unknown. We have that $0 \leq q \leq 1$. Use expectation maximization principle to find analytical expressions for calculating μ_1, μ_2 and q which will be further used in iterations. Note that the task is to find the analytical expressions. (4p)
Hint: You need to write EM help function in terms of all necessary parameters and then maximize the help function via a standard approach.

Solution: Introduce the latent variable S_t and set $\lambda = (\mu_1, \mu_2, q)$. If $S_t = 1$, then x_t comes from the first Gaussian distribution (with mean μ_1), if $S_t = 2$ the x_t comes from the second distribution (with mean μ_2). We find that

$$\begin{aligned}
p(S_t = 1|x_t, q) &= \frac{p(x_t|S_t = 1, q)p(S_t = 1|q)}{p(x_t|q)} \\
&= \frac{qe^{-\frac{1}{2\sigma^2}(x_t-\mu_1)^2}}{qe^{-\frac{1}{2\sigma^2}(x_t-\mu_1)^2} + (1-q)e^{-\frac{1}{2\sigma^2}(x_t-\mu_2)^2}} = \gamma_t, \\
p(\underline{x}, \underline{S}|\lambda) &= \prod_{t=1}^{N} p(x_t|S_t, \lambda)p(S_t|\lambda).
\end{aligned}$$

Using this, we find that

$$\begin{aligned}
&Q(\lambda', \lambda) = E[\log p(\underline{x}, \underline{S}|q')|\underline{x}, q] = \text{constant} \\
&+ \sum_{t=1}^{N}\left[\gamma_t\left(-\frac{1}{2\sigma^2}(x_t - \mu_1')^2 + \log(q')\right) + (1-\gamma_t)\left(-\frac{1}{2\sigma^2}(x_t - \mu_2')^2 + \log(1-q')\right)\right]
\end{aligned}$$

Maximizing $Q(\lambda', \lambda)$ with respect to λ' gives us the iterative update equations

$$\begin{aligned}
\mu_1' &= \left(\sum_{t=1}^{N}\gamma_t x_t\right) / \left(\sum_{t=1}^{N}\gamma_t\right), \\
\mu_2' &= \left(\sum_{t=1}^{N}(1-\gamma_t)x_t\right) / \left(\sum_{t=1}^{N}(1-\gamma_t)\right), \\
q' &= \frac{1}{N}\sum_{t=1}^{N}\gamma_t.
\end{aligned}$$

5 (HMM) We observe a sequence $\underline{x} = (1, 3, 2)$ from an HMM with

$$q = \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}, \; A = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}, \; B = \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0.8 & 0.2 \end{pmatrix},$$

where q, A and B have usual meanings. Determine the most probable sequence of states (S_1, S_2, S_3) given the observations $\underline{x}$. (3p)
Hint: Use the Viterbi algorithm. Viterbi algorithm has two important parameters to compute.

Viterbi Probability Vector : $\chi_{j,t} = \max_{i_1 \ldots i_{t-1}} P[S_1 = i_1, \ldots S_{t-1} = i_{t-1}, S_t = j, x_1, \ldots, x_t]$,
Viterbi Backpointer Matrix : $\zeta_{j,t} = \arg\max_i \chi_{i,t-1} a_{ij}$.

Solution: Using the Viterbi algorithm we find that

	State 1	State 2
q	0.3	0.7
$b(x_1)$	0.5	0
χ_1	0.15	0
$\max_i \chi_{i,1} a_{ij}$	0.12	0.03
ζ_2	1	1
$b(x_2)$	0.5	0.2
χ_2	0.06	0.006
$\max_i \chi_{i,2} a_{ij}$	0.048	0.012
ζ_2	1	1
$b(x_3)$	0	0.8
χ_2	0	0.0096

Backtracking we find that $\hat{i}_3 = 2$, $\hat{i}_2 = 1$ and $\hat{i}_1 = 1$. The Viterbi algorithm thus gives that the most probable sequence is $(1, 1, 2)$.

6 (Bayesian Learning) Consider a two-dimensional Gaussian distribution $p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$ with known mean

$$\boldsymbol{\mu} = \begin{pmatrix} -1.7 \\ 3.2 \end{pmatrix}.$$

The covariance matrix is unknown, but is believed to be close to the matrix

$$\mathbf{C}_0 = 1.41 \cdot \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}.$$

We model this prior knowledge using the inverse-Wishart distribution

$$p(\mathbf{C}) = \frac{(\nu-3)^{\nu} \det(\mathbf{C}_0)^{\nu/2}}{2^{\nu}\Gamma_2(\nu/2)\det(\mathbf{C})^{(\nu+3)/2}} e^{-\frac{1}{2}(\nu-3)\mathrm{tr}(\mathbf{C}_0\mathbf{C}^{-1})},$$

as a prior distribution, where $\Gamma_2(x) = \pi^{1/2}\Gamma(x)\Gamma(x-1/2)$, $\nu > 3$, $\Gamma(.)$ is standard Gamma function and the distribution is defined for all positive definite $\mathbf{C} \in \mathcal{R}^{2\times 2}$ ($\mathbf{C} \succ \mathbf{0}$). The inverse Wishart distribution has the mean

$$\mathcal{E}[\mathbf{C}] = \mathbf{C}_0.$$

What is the posterior distribution of $\mathbf{C}$ after measuring datapoints $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$? What is the posterior mean of $\mathbf{C}$ (the mean with respect to the posterior distribution)? (4p)
Hint: It is useful to introduce the *sample covariance matrix*

$$\hat{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})^{\top},$$

and use the relation $\mathbf{z}^{\top}\mathbf{A}\mathbf{z} = \mathrm{tr}(\mathbf{A}\mathbf{z}\mathbf{z}^{\top})$.

Solution: Denote the observations by $\underline{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. The posterior distribution becomes

$$\begin{aligned}
p(\mathbf{C}|\underline{x}) &= \frac{p(\underline{x}|\mathbf{C})p(\mathbf{C})}{p(\underline{x})} \propto p(\underline{x}|\mathbf{C})p(\mathbf{C}) \\
&\propto \det(\mathbf{C})^{-(\nu+3)/2-T/2} \exp\left(-\frac{1}{2}\left[\mathrm{tr}(\mathbf{C}^{-1}(\nu-3))\mathbf{C}_0) + \sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{x}_t - \boldsymbol{\mu})\right]\right) \\
&= \det(\mathbf{C})^{-(\nu+T+3)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}((\nu-3)\mathbf{C}_0 + T\hat{\mathbf{R}})\right) \\
&= \det(\mathbf{C})^{-(\nu+T+3)/2} \exp\left(-\frac{1}{2}(\nu+T-3)\mathrm{tr}(\mathbf{C}^{-1}\mathbf{C}_T)\right),
\end{aligned}$$

where

$$\mathbf{C}_T = \frac{(\nu-3)\mathbf{C}_0 + T\hat{\mathbf{R}}}{\nu+T-3}.$$

Using that

$$\int_{\mathbf{C}\succ\mathbf{0}} \frac{(\nu-3)^{\nu}\det(\mathbf{C}_0)^{\nu/2}}{2^{\nu}\Gamma_2(\nu/2)\det(\mathbf{C})^{(\nu+3)/2}} e^{-\frac{1}{2}(\nu-3)\mathrm{tr}(\mathbf{C}_0\mathbf{C}^{-1})} d\mathbf{C} = 1,$$

we find that

$$\begin{aligned}&\int_{\mathbf{C}\succ\mathbf{0}} \det(\mathbf{C})^{-(\nu+T+3)/2} \exp\left(-\frac{1}{2}(\nu+T-3)\mathrm{tr}(\mathbf{C}^{-1}\mathbf{C}_T)\right) d\mathbf{C}\\ &= \frac{2^{\nu+T}\Gamma_2((\nu+T)/2)}{(\nu+T-3)^{\nu+T}\det(\mathbf{C}_T)^{-(\nu+T)/2}}.\end{aligned}$$

So

$$p(\mathbf{C}|\underline{x}) = \frac{(\nu+T-3)^{\nu+T}\det(\mathbf{C}_T)^{(\nu+T)/2}}{2^{\nu+T}\Gamma_2(\nu/2)\det(\mathbf{C})^{(\nu+T+3)/2}} e^{-\frac{1}{2}(\nu+T-3)\mathrm{tr}(\mathbf{C}_T\mathbf{C}^{-1})}.$$

The mean thus becomes

$$\mathcal{E}[\mathbf{C}] = \mathbf{C}_T.$$

We also see that the inverse Wishart is the conjugate prior for the covariance matrix.