KTH Electrical Engineering

Solutions to Exam in Pattern Recognition EN2200

Date:	Tuesday, Oct 23, 2007, 14.00 - 19.00
Place:	Q24,Q26.
Allowed:	Beta (or corresponding), calculator with empty memory.
Grades:	A: at least 31p; B: 27p; C: 23p; D: 20p; E: 17p (incl. project results).
Language:	Optional: Swedish or English.
Solutions:	To be published on the course web page.
Results:	Tuesday, Nov 6, 2007.
Review:	At KTH-S3/ STEX, Osquldas v. 10.

Good Luck!

Please do the **Course Evaluation**! See the course web page.

1 It is known that about 1 out of 3000 new-born children has severely impaired hearing. It is important to detect this problem early, but test methods give rather unreliable results during the first days after birth. Therefore, two different test methods are used together, to give early indications of the problem.

Each test gives a real-valued result. After suitable transformation and scaling, the test results have approximately Gaussian (Normal) probability distributions with means and standard deviations (SD) shown in the following table. The two test results are statistically independent of each other, within any tested person.

Hearing	**Test A**		**Test B**	
	Mean	SD	Mean	SD
Normal	0	1	0	2
Impaired	1	1	3	2

(a) Design a classifier to guess, with minimum error probability, if a tested child has a hearing loss, given any combination of test results. (4p)
Note: For full credit, you must use the formal mathematical language of probability theory and show that an optimal decision can be made using only a scalar *linear* combination of the test results.

Solution: We regard the child's possible hearing loss as a hidden discrete source state S with two possible outcomes: $S = 0$, for normal hearing, or $S = 1$ indicating a hearing loss. The *a priori* state probabilities are given as $p_0 = P(S = 0) \approx 1-1/3000$ and $p_1 = P(S = 1) \approx 1/3000$.

We regard the test results as an outcome, $\boldsymbol{x} = (x_A, x_B)^T$, of a feature vector $\boldsymbol{X}$.

Criterion: As the a priori source probabilities are not equal we must use the MAP classification criterion.

Discriminant function: The conditional probability densities have identical variances, regardless of source, but different means, and can be expressed as (with short-hand notation)

$$P(\boldsymbol{X} = \boldsymbol{x}|S = 0) = \frac{1}{\sigma_A\sqrt{2\pi}} e^{-\frac{(x_A-\mu_{0A})^2}{2\sigma_A^2}} \cdot \frac{1}{\sigma_B\sqrt{2\pi}} e^{-\frac{(x_B-\mu_{0B})^2}{2\sigma_B^2}}$$
$$P(\boldsymbol{X} = \boldsymbol{x}|S = 1) = \frac{1}{\sigma_A\sqrt{2\pi}} e^{-\frac{(x_A-\mu_{1A})^2}{2\sigma_A^2}} \cdot \frac{1}{\sigma_B\sqrt{2\pi}} e^{-\frac{(x_B-\mu_{1B})^2}{2\sigma_B^2}}$$

The conditional probability that the tested child has a hearing loss is

$$P(S = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}|S = 1)P(S = 1)}{P(\boldsymbol{X} = \boldsymbol{x})}$$

and similarly for $P(S = 0|\boldsymbol{X} = \boldsymbol{x}) = 1 - P(S = 1|\boldsymbol{X} = \boldsymbol{x})$.

We can define a single discriminant function as

$$\begin{aligned}
g_1(\boldsymbol{x}) =& \ln P(S=1|\boldsymbol{X}=\boldsymbol{x}) - \ln P(S=0|\boldsymbol{X}=\boldsymbol{x}) = \\
=& -\frac{(x_A-\mu_{1A})^2}{2\sigma_A^2} - \frac{(x_B-\mu_{1B})^2}{2\sigma_B^2} + \frac{(x_A-\mu_{0A})^2}{2\sigma_A^2} - \frac{(x_B-\mu_{0B})^2}{2\sigma_B^2} + \ln\frac{p_1}{p_0} = \\
=& \frac{x_A(\mu_{1A}-\mu_{0A})}{\sigma_A^2} + \frac{x_B(\mu_{1B}-\mu_{0B})}{\sigma_B^2} - \frac{\mu_{1A}^2-\mu_{0A}^2}{2\sigma_A^2} - \frac{\mu_{1B}^2-\mu_{0B}^2}{2\sigma_B^2} + \ln\frac{p_1}{p_0} = \\
=& \frac{(x_A-m_A)d_A}{\sigma_A^2} + \frac{(x_B-m_B)d_B}{\sigma_B^2} + \ln\frac{p_1}{p_0}
\end{aligned}$$

where

$$\begin{aligned}
d_A =& \mu_{1A}-\mu_{0A}; & d_B =& \mu_{1B}-\mu_{0B} \\
m_A =& (\mu_{1A}+\mu_{0A})/2; & m_B =& (\mu_{1B}+\mu_{0B})/2
\end{aligned}$$

As the quadratic terms cancelled, this discriminant function is obviously just a linear combination of the observed test results $\boldsymbol{x}$.

Decision: The classifier must guess that the child has a hearing loss if $g_1(\boldsymbol{x}) > 0$.

(b) The test results for one child were 1.5 in test A and 3.5 in test B. What is the decision of your classifier in this case? (1p)

Solution: For the individual test result $\boldsymbol{x} = (1.5, 3.5)^T$ and the given density parameters, we obtain

$$\begin{aligned}
g_1(\boldsymbol{x}) =& \frac{(x_A-m_A)d_A}{\sigma_A^2} + \frac{(x_B-m_B)d_B}{\sigma_B^2} + \ln\frac{p_1}{p_0} = \\
=& \frac{(1.5-0.5)1}{1^2} + \frac{(3.5-1.5)3}{2^2} + \ln\frac{1}{2999} \approx \\
\approx& 1.0 + 1.5 - 8.0 < 0
\end{aligned}$$

Thus, the optimal MAP decision is to guess *normal hearing* in this case. Apparently, the test-result evidence for a hearing loss was not sufficiently strong to overcome the very low *a priori* probability for a hearing loss. Perhaps it would have been better to apply a Minimum-Risk criterion with a much higher cost for a *miss* than for a *false alarm*.

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

(a) To design an optimal classifier with N_d possible decisions, for a source with N_s source states, using K observed features, it is always sufficient to define at least K discriminant functions.

Solution: FALSE. The optimal classifier needs N_d (or $N_d - 1$) discriminant functions, so K is certainly not enough if $K << N_d$.

(b) Given an observed output sequence $\underline{\boldsymbol{x}} = (x_1, \ldots, x_T)$ from a Hidden Markov Model λ, we can run the complete Viterbi algorithm and then use the final result of one of the recursive

algorithm variables to determine the probability of the observed sequence, i.e. $P(\underline{\boldsymbol{x}}|\lambda)$. **Solution:** FALSE. The Viterbi algorithm is designed to evaluate probabilities only for the *single* best state sequence, but the total probability of the observed sequence may include non-zero contributions from many state sequences.

(c) A hidden Markov model with the following initial state probabilities and state transition probabilities produces a *stationary* and *ergodic* random sequence.

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.8 \\ 0.1 \\ 0.1 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 & 0 \\ 0.08 & 0.91 & 0.01 \\ 0 & 0.01 & 0.99 \end{pmatrix};$$

Solution: TRUE. Stationary, because $q = A^T q$. Ergodic because the Markov chain is irreducible and aperiodic.

(d) An optimal MAP classifier is designed for $N_s = 2$ source states with non-zero a priori probabilities, using a K-dimensional Gaussian feature vector with different means and equal covariance matrices for the two source states. This classifier defines exactly 2 *connected decision regions* in K-dimensional feature space.

Solution: TRUE. Decision regions need not be connected in general. However, if covariance matrices are *equal*, a linear discriminant function can be used. Then the decision boundary is a hyperplane in K-dimensional space, so there can be only two connected decision regions separated by this hyperplane.

(e) A Hidden Markov Model with N states is used to characterise a sequence of vector-valued observations $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots)$. Each observed $\boldsymbol{x}_t$ is a vector with K real-valued elements. In this case, the output probability density functions $b_j(\boldsymbol{x})$ must have values between 0 and 1 for any observed $\boldsymbol{x}$.

Solution: FALSE. As the functions calculate a probability *density*, the output can have any value $b_j(\boldsymbol{x}) \geq 0$.

3 You can observe the output sequence $\boldsymbol{x} = (x_1, \ldots, x_t, \ldots)$ from a discrete Hidden-Markov source, but you do not know the corresponding internal state sequence $\boldsymbol{S} = (S_1, \ldots, S_t, \ldots)$ in the source.
The initial state probability vector is

$$q = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}, \text{with elements } P(S_1 = i).$$

The state transition probability matrix is

$$A = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix}, \text{with elements } a_{ij} = P(S_{t+1} = j | S_t = i).$$

The output probability matrix is

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(X_t = k | S_t = i).$$

(a) Calculate $P(X_2 = 3 \cap S_2 = 1)$. (2p)

Solution:

$$\begin{aligned} P(S_2 = 1 \cap X_2 = 3) =& P(S_1 = 1 \cap S_2 = 1 \cap X_2 = 3) + P(S_1 = 2 \cap S_2 = 1 \cap X_2 = 3) = \\ =& q_1 a_{11} b_{13} + q_2 a_{21} b_{13} = \\ =& 0.8 \cdot 0.6 \cdot 0.5 + 0.2 \cdot 0.1 \cdot 0.5 = 0.25 \end{aligned}$$

(b) Calculate $P(S_2 = 1 | S_1 = 1 \cap X_1 = 3 \cap X_2 = 1 \cap S_3 = 1 \cap X_3 = 3)$. (3p)

Solution: S_2 is conditionally independent of X_1 and X_3, given S_1 and S_3, because of the Markov property. Therefore,

$$\begin{aligned} & P(S_2 = 1 | S_1 = 1 \cap X_1 = 3 \cap X_2 = 1 \cap S_3 = 1 \cap X_3 = 3) = \\ =& P(S_2 = 1 | S_1 = 1 \cap X_2 = 1 \cap S_3 = 1) = \\ =& \frac{P(S_2 = 1 \cap X_2 = 1 \cap S_3 = 1 | S_1 = 1)}{P(X_2 = 1 \cap S_3 = 1 | S_1 = 1)} = \\ =& \frac{a_{11} b_{11} a_{11}}{P(X_2 = 1 \cap S_3 = 1 | S_1 = 1)} \end{aligned}$$

where

$$P(S_2 = i \cap X_2 = 1 \cap S_3 = 1 | S_1 = 1) = a_{1i} b_{i1} a_{i1}$$

and

$$\begin{aligned} P(X_2 = 1 \cap S_3 = 1 | S_1 = 1) =& \sum_{i=1}^{2} P(S_2 = i \cap X_2 = 1 \cap S_3 = 1 | S_1 = 1) = \\ =& a_{11} b_{11} a_{11} + a_{12} b_{21} a_{21} \end{aligned}$$

Thus,

$$\begin{aligned} & P(S_2 = 1 | S_1 = 1 \cap X_2 = 1 \cap S_3 = 1) = \\ =& \frac{a_{11} b_{11} a_{11}}{a_{11} b_{11} a_{11} + a_{12} b_{21} a_{21}} = \\ =& \frac{0.6 \cdot 0.1 \cdot 0.6}{0.6 \cdot 0.1 \cdot 0.6 + 0.4 \cdot 0.7 \cdot 0.1} = 0.5625 \end{aligned}$$

4 (Sinusoidally Amplitude-modulated Noise) Two signal sources, called $S = 0$ and $S = 1$, both generate random sequences $\boldsymbol{X} = (X_1, \ldots, X_t, \ldots, X_T)$ with scalar random variables X_t. One of these sources is initially chosen at random, with equal probability, and all samples are then generated by this same source. You have observed an output sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$.

The first signal source, $S = 0$, is a reference that generates *steady* zero-mean white noise with constant variance σ^2. The second source generates *amplitude-modulated* zero-mean white noise with the same *average* variance across the the observed interval $1 \ldots T$, but here the variance (power) varies sinusoidally during this interval. More precisely,

$$\begin{aligned} &\text{If } S = 0, \quad X_t = W_t, && t = 1, \ldots, T \\ &\text{If } S = 1, \quad X_t = \sqrt{1 + a \sin(2\pi t / T)} W_t, && t = 1, \ldots, T \end{aligned}$$

Here, the modulation depth a is a known small positive constant, i.e. $0 < a << 1$. Regardless of the source state, each noise sample W_t is a Gaussian random variable with zero mean and known variance σ^2, and all W_t samples at different t are statistically independent of each other.

Note that the modulation for $S = 1$ gives a time-varying signal variance, with

$$\text{var}\left[X_t|S=1\right] = \sigma_{1t}^2 = \sigma^2\left(1 + a\sin\left(2\pi t/T\right)\right)$$
$$\frac{1}{T}\sum_{t=1}^{T}\sigma_{1t}^2 = \sigma^2$$

(a) Design an optimal (non-linear) classifier that can guess, with minimum error probability, which of the two sources, $S = 0$ or $S = 1$, generated the observed sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$. (4p)

Solution:

Decision criterion: Both sources are equally probable, so we use the ML criterion.

Discriminant function: We use a single discriminant function, defined as

$$\begin{aligned} g_1'(\boldsymbol{x}) &= \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|1) - \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|0) = \\ &= \ln\prod_{t=1}^{T}\frac{1}{\sigma_{1t}\sqrt{2\pi}}e^{-\frac{(x_t-0)^2}{2\sigma_{1t}^2}} - \ln\prod_{t=1}^{T}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_t-0)^2}{2\sigma^2}} = \\ &= \sum_{t=1}^{T} -\frac{1}{2}\ln\left(1 + a\sin\left(2\pi t/T\right)\right) - \frac{x_t^2}{2\sigma^2\left(1 + a\sin\left(2\pi t/T\right)\right)} + \frac{x_t^2}{2\sigma^2} = \\ &= \sum_{t=1}^{T}\frac{x_t^2 a\sin\left(2\pi t/T\right)}{2\sigma^2\left(1 + a\sin\left(2\pi t/T\right)\right)} - \frac{1}{2}\ln\left(1 + a\sin\left(2\pi t/T\right)\right) \end{aligned}$$

We can easily scale the discriminant function to get rid of the factor $1/2$, and redefine

$$g_1(\boldsymbol{x}) = \sum_{t=1}^{T}\frac{x_t^2 a\sin\left(2\pi t/T\right)}{\sigma^2\left(1 + a\sin\left(2\pi t/T\right)\right)} - \ln\left(1 + a\sin\left(2\pi t/T\right)\right)$$

Decision: Guess that $S = 1$, iff $g_1(\boldsymbol{x}) > 0$

(b) Simplify the classifier by using first-order approximations for $|a| << 1$, for example $1+a \approx 1$, and $\ln(1+a) \approx a$. Argue briefly why the classifier performance is approximately independent of σ^2. (1p)

Solution: For very small values of a, the discriminant function can be simplified as

$$\begin{aligned} g_1(\boldsymbol{x}) &\approx \sum_{t=1}^{T}\frac{x_t^2 a\sin\left(2\pi t/T\right)}{\sigma^2} - a\sin\left(2\pi t/T\right) = \\ &= \sum_{t=1}^{T}\frac{x_t^2 a\sin\left(2\pi t/T\right)}{\sigma^2} \end{aligned}$$

If the variance σ^2 is scaled by some factor, the statistical characteristics of the scalar decision variable $Y = g_1(\boldsymbol{X})$ will remain unchanged, because the decision variable depends only on X_t^2/σ^2.

5 (Randomly Amplitude-modulated Noise) A signal source generates a random sequence $\boldsymbol{X} = (X_1, \ldots, X_t, \ldots, X_T)$ with scalar random variables X_t, generated as

$$X_t = \sigma\sqrt{1 + aZ_t}W_t$$

Here, the modulation depth a is an exactly known constant, with $0 < a < 1$, but the amplitude factor σ is only approximately known. The random values W_t are noise samples, each with a Gaussian $N(0, 1)$ distribution, and all W_t samples at different t are statistically independent of each other.

The random sequence $\boldsymbol{Z} = (Z_1, \ldots, Z_t, \ldots, Z_T)$ contains discrete elements Z_t that are either $+1$ or -1, with probability distribution

$$\begin{aligned} P[Z_1 = +1] = P[Z_1 = -1] &= 0.5 \\ P[Z_t = +1 | Z_{t-1} = +1] &= \\ = P[Z_t = -1 | Z_{t-1} = -1] &= r; \quad t = 2, \ldots, T \end{aligned}$$

Here, the conditional probability r is constant and known exactly, with $0 < r < 1$.

(a) Define the signal source formally as a hidden Markov model (HMM) λ, and calculate $E[X_t]$ and $\text{var}[X_t]$, expressed in terms of σ, a, and r. (2p)

Solution: We label the two states simply as $+$ and $-$, representing $Z_t = +1$ and $Z_t = -1$. Then we have the HMM $\lambda = (q, A, B)$, with

$$\begin{aligned} \text{initial probability vector: } q &= \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \\ \text{transition probability matrix: } A &= \begin{pmatrix} r & 1-r \\ 1-r & r \end{pmatrix} \\ \text{output densities: } b_+(x;\sigma,a) &= \frac{1}{\sigma\sqrt{1+a}\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2(1+a)}} \\ b_-(x;\sigma,a) &= \frac{1}{\sigma\sqrt{1-a}\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2(1-a)}} \end{aligned}$$

Because of the symmetry, the HMM is stationary, and $P(Z_t = +1|\lambda) = P(Z_t = -1|\lambda) = 0.5$, at any t. Therefore,

$$\begin{aligned} E[X_t] &= E[X_t|Z_t = +1]\,P(Z_t = +1) + E[X_t|Z_t = -1]\,P(Z_t = -1) = \\ &= \sigma\sqrt{1+a}E[W_t]\,0.5 + \sigma\sqrt{1-a}E[W_t]\,0.5 = 0 \\ \text{var}[X_t] &= E\left[X_t^2|Z_t = +1\right]P(Z_t = +1) + E\left[X_t^2|Z_t = -1\right]P(Z_t = -1) = \\ &= \sigma^2(1+a)0.5 + \sigma^2(1-a)0.5 = \sigma^2 \end{aligned}$$

(b) Apply one step of the Expectation Maximization (EM) algorithm to obtain an improved estimate of the parameter σ, given one observed output sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$. The well-known Forward-Backward algorithm may be used directly, without any theoretical derivation, and all other model parameters are assumed to be known. (3p)

Hint: Each step in the EM algorithm should maximize the function

$$Q(\lambda', \lambda) = E[\ln P(\boldsymbol{Z}, \boldsymbol{x}|\lambda')|\boldsymbol{x}, \lambda]$$

Solution: Using the Forward-Backward algorithm we have obtained, using the previous parameter estimate σ,

$$\gamma_{+,t} = P(Z_t = +1|\boldsymbol{x}, \lambda)$$

and similarly for $\gamma_{-,t}$. We also note that $\gamma_{+,t} + \gamma_{-,t} = 1$.

The Q function depends the free new parameter value σ', as

$$\begin{aligned} Q(\sigma', \lambda) &= \sum_{t=1}^{T} \gamma_{+,t} \ln b_+(x_t; \sigma', a) + \gamma_{-,t} \ln b_-(x_t; \sigma', a) + \ldots = \\ &= \sum_{t=1}^{T} \gamma_{+,t} \left(-\ln \sigma - \frac{x_t^2}{2\sigma'^2(1+a)} \right) + \\ &\quad + \gamma_{-,t} \left(-\ln \sigma' - \frac{x_t^2}{2\sigma'^2(1-a)} \right) + \ldots \end{aligned}$$

Here, the dots $\ldots$ indicate remaining terms that do not depend on σ'. To maximize Q, we apply the necessary condition on the partial derivative:

$$0 = \frac{\partial Q}{\partial \sigma'} = -\frac{T}{\sigma'} + \sum_{t=1}^{T} \left(\frac{\gamma_{+,t} x_t^2}{\sigma'^3(1+a)} + \frac{\gamma_{-,t} x_t^2}{\sigma'^3(1-a)} \right)$$

with solution

$$\sigma'^2 = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\gamma_{+,t} x_t^2}{1+a} + \frac{\gamma_{-,t} x_t^2}{1-a} \right)$$