KTH Electrical Engineering

Solutions to Exam in Pattern Recognition 2E1395

Date:	Friday, Oct 20, 2006, 14.00 - 19.00
Place:	V21, 22, 33, 34.
Allowed:	Beta (or corresponding), calculator with empty memory.
Grades:	5: at least 27p; 4: at least 22p; 3: at least 17p (incl. project results).
Language:	Optional: Swedish or English.
Solutions:	To be published on the course web page.
Results:	Friday, Nov 03, 2006.
Review:	At KTH-S3/ STEX, Osquldas v. 10.

Good Luck!

1 About 1 out of 100 people has a special gene that increases the risk of developing heart problems at old age, but the risk can be reduced by medication. Therefore, two laboratory tests, here called A and B, are used to give early indications on this genetic condition. Each test method gives only a binary result, either "+" or "−", with "+" indicating a higher probability for the genetic risk factor. In several studies, the *sensitivity* and *specificity* of the test methods have been determined as shown in the following table. However, the sensitivity (correction: not specificity) of test A is known to be different, depending on the result of test B, as indicated in the table.

Test	Sensitivity	Specificity
A, if B is "+"	0.90	0.98
A, if B is "−"	0.99	0.98
B	0.8	0.9

The sensitivity is the conditional probability that the method indicates a "+" result, given that the patient has the problematic gene. The specificity is the conditional probability that the method indicates a "−" result, given that the patient is healthy.

Test results for one patient are "+−" for tests A and B. What is the probability that this person has the problematic gene? (5p)

Note: For full credit, you must use the formal mathematical language of probability theory. Specify any necessary additional assumptions not already given in the text above.

Solution:

We regard the patient's possible genetic condition as a hidden discrete source state S with two possible outcomes: $S = 0$, if he is healthy, or $S = 1$ meaning he has the problematic gene. The *a priori* state probabilities are given as $P(S = 0) = 0.99$ and $P(S = 1) = 0.01$.

We regard the test results as an outcome, $\mathbf{x} = (1, 0)$, of a feature vector $\mathbf{X}$ with two elements, where element X_i indicates the result of the i-th test, with "+" coded as $X_i = 1$, and "−" as $X_i = 0$.

The given table shows the *sensitivity* value for the i-th test defined as $P(X_i = 1|S = 1)$, and the *specificity* defining $P(X_i = 0|S = 0)$. (We now use the sensitivity and specificity values from the table, although the text above the table states a contradiction to the information in the table.)

Criterion: As the a priori source probabilities are not equal we must use the MAP criterion to decide whether the patient is most probably infected or not.

Discriminant functions: The joint probabilities of states and observations are

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \cap S = 0) =& P(S = 0)P(X_A = 1|X_B = 0 \cap S = 0)P(X_B = 0|S = 0) = \\ =& 0.99 \cdot (1 - 0.98) \cdot 0.9 = 0.0172 \\ P(\mathbf{X} = \mathbf{x} \cap S = 1) =& P(S = 1)P(X_A = 1|X_B = 0 \cap S = 1)P(X_B = 0|S = 1) = \\ =& 0.01 \cdot 0.99 \cdot (1 - 0.8) = 0.00198 \end{aligned}$$

The conditional probability that the patient has the gene is, thus,

$$g_1(\mathbf{x}) = P(S = 1|\mathbf{X} = \mathbf{x}) = \frac{0.00198}{0.0172 + 0.00198} \approx 0.103$$

Decision: Thus, we must conclude that the patient probably does *not* have the problematic gene.

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, it is always sufficient to define N_d discriminant functions, even if $N_d < N_s$.

Solution:
TRUE. The optimal classifier needs one discriminant function for each decision alternative, regardless of the number of source states.

(b) Given an observed output sequence $\mathbf{x} = (x_1, \ldots, x_T)$ from a Hidden Markov Model λ, the result of the Forward algorithm is sufficient to determine the joint probability

$$P\left(S_t = j \cap (x_1, \ldots, x_t)|\lambda\right), \quad \text{for any } t \in \{1 \ldots T\}$$

Solution:
TRUE. Using course-book notation for the Forward algorithm results, the desired probability is $\hat{\alpha}_{jt}/(c_1 \cdot \ldots \cdot c_t)$.

(c) If an optimal MAP classifier is designed using discriminant functions $g_i(\mathbf{x})$, the output value from each discriminant function must be greater than zero, i.e. $g_i(\mathbf{x}) > 0$ for all i.

Solution:
FALSE. Although preliminary discriminant functions are derived as a posteriori state probability values, any monotonous transformation of the discriminant functions gives equally optimal results.

(d) In a Hidden Markov Model used to characterise a sequence of vector-valued observations $\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots)$, where each observed $\mathbf{x}_t$ is a vector with K real-valued elements, the output probability density functions $b_j(\mathbf{x})$ can be represented as Gaussian Mixture Models, *if and only if* each $b_j(\mathbf{x})$ is a sum of exactly K probability density functions.

Solution:
FALSE. The Gaussian Mixture can be a weighted sum of any number of Gaussian density functions, but every Gaussian component must calculate the probability density for a K-dimensional vector as its input argument.

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

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 \\ 0.04 & 0.96 \end{pmatrix};$$

Solution:
TRUE, because $A^T q = q$.

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

The initial state probability vector is

$$q = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}, \text{with elements } P(S_1 = i).$$

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

$$B = \begin{pmatrix} 0.1 & 0.3 & 0.6 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(X_t = k | S_t = i).$$

(a) Calculate $P(S_1 = 1 \cap X_2 = 1)$. (2p)

Solution:

$$\begin{aligned} P(S_1 = 1 \cap X_2 = 1) =& P(S_1 = 1 \cap S_2 = 1 \cap X_2 = 1) + P(S_1 = 1 \cap S_2 = 2 \cap X_2 = 1) = \\ =& q_1 a_{11} b_{11} + q_1 a_{12} b_{21} = \\ =& 0.7 \cdot 0.6 \cdot 0.1 + 0.7 \cdot 0.4 \cdot 0.6 = 0.21 \end{aligned}$$

(b) Calculate $P(S_2 = 1 \cap S_3 = 1 | S_1 = 1 \cap X_1 = 1 \cap X_3 = 1 \cap S_4 = 2 \cap X_5 = 1)$. (3p)

Solution:
S_3 is conditionally independent on X_1 and X_5, given S_1 and S_4, because of the Markov property. Therefore,

$$\begin{aligned} & P(S_2 = 1 \cap S_3 = 1 | S_1 = 1 \cap X_1 = 1 \cap X_3 = 1 \cap S_4 = 2 \cap X_5 = 1) = \\ =& P(S_2 = 1 \cap S_3 = 1 | S_1 = 1 \cap X_3 = 1 \cap S_4 = 2) = \\ =& \frac{P(S_2 = 1 \cap S_3 = 1 \cap X_3 = 1 \cap S_4 = 2 | S_1 = 1)}{P(X_3 = 1 \cap S_4 = 2 | S_1 = 1)} = \\ =& \frac{a_{11} a_{11} b_{11} a_{12}}{P(X_3 = 1 \cap S_4 = 2 | S_1 = 1)} \end{aligned}$$

where

$$P(S_2 = i \cap S_3 = j \cap X_3 = 1 \cap S_4 = 2 | S_1 = 1) = a_{1i} a_{ij} b_{j1} a_{j2}$$

and

$$P(X_3 = 1 \cap S_4 = 2|S_1 = 1) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(S_2 = i \cap S_3 = j \cap X_3 = 1 \cap S_4 = 2|S_1 = 1) =$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} a_{1i}a_{ij}b_{j1}a_{j2} =$$
$$= \begin{pmatrix}0.6 & 0.4\end{pmatrix} \begin{pmatrix}0.6 & 0.4\\ 0.1 & 0.9\end{pmatrix} \begin{pmatrix}0.1 \cdot 0.4\\ 0.6 \cdot 0.9\end{pmatrix} = 0.34$$

Thus,

$$P(S_2 = 1 \cap S_3 = 1|S_1 = 1 \cap X_3 = 1 \cap S_4 = 2) =$$
$$= \frac{a_{11}a_{11}b_{11}a_{12}}{0.34} =$$
$$= \frac{0.0144}{0.34} = 0.0424$$

4 Two signal sources, called $S = 1$ and $S = 2$, both generate random sequences $\mathbf{X} = (X_1, \ldots, X_t, \ldots)$ with scalar random variables X_t. One of these sources is initially chosen at random, with equal probability, and you have observed an output sequence $\mathbf{x} = (x_1, \ldots, x_T)$.

The output sequence is generated as

$$\begin{array}{lll} & X_0 = 0 & \\ \text{If } S = 1, & X_t = U_t a X_{t-1} + W_t, & t = 1, \ldots, T \\ \text{If } S = 2, & X_t = W_t, & t = 1, \ldots, T \end{array}$$

Here, $a > 0$ is a known filter parameter, and U_t is a randomly chosen sign factor, either $U_t = +1$ or $U_t = -1$, with equal probability. All U_t at different t are statistically independent of each other. 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.

Design an optimal classifier that can guess, with minimum error probability, which of the two sources, $S = 1$ or $S = 2$, generated the observed sequence $\mathbf{x} = (x_1, \ldots, x_T)$. (5p)

Hint: Note that

$$P(\mathbf{X}) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdot \ldots \cdot P(X_T|X_1, \ldots, X_{T-1}) =$$
$$= P(X_1)P(X_2|X_1)P(X_3|X_2) \cdot \ldots \cdot P(X_T|X_{T-1})$$

because each element in the sequence can depend at most only on the nearest preceding element.

Solution:

As the two sources are equally probable we use the ML decision rule.

In source $S = 1$, every output sample X_t is conditionally dependent on only the nearest previous sample X_{t-1}. Given $X_{t-1} = x_{t-1}$, the only remaining random component of X_t is W_t, so X_t is then conditionally Gaussian with variance σ^2, and with a mean value determined by U_t as either ax_{t-1} or $-ax_{t-1}$, with equal probability.

Therefore, optimal preliminary discriminant functions can be defined as

$$
\begin{aligned}
g_1(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}|S = 1) = \prod_{t=1}^{T} P(x_t|x_{t-1} \cap S = 1) = \\
&= \prod_{t=1}^{T} \left(0.5 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t - ax_{t-1})^2}{2\sigma^2}} + 0.5 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t - (-ax_{t-1}))^2}{2\sigma^2}} \right) = \\
g_2(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}|S = 2) = \prod_{t=1}^{T} P(x_t|x_{t-1} \cap S = 2) = \\
&= \prod_{t=1}^{T} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t - 0)^2}{2\sigma^2}}
\end{aligned}
$$

As there are only two source alternatives, an optimal classifier can use a single discriminant function defined as

$$
\begin{aligned}
g(\mathbf{x}) &= \ln g_1(\mathbf{x}) - \ln g_2(\mathbf{x}) = \\
&= \sum_{t=1}^{T} \ln \frac{e^{-(x_t - ax_{t-1})^2/2\sigma^2} + e^{-(x_t + ax_{t-1})^2/2\sigma^2}}{2e^{-x_t^2/2\sigma^2}} = \\
&= \sum_{t=1}^{T} \ln \frac{e^{(2ax_t x_{t-1} - a^2 x_{t-1}^2)/2\sigma^2} + e^{(-2ax_t x_{t-1} - a^2 x_{t-1}^2)/2\sigma^2}}{2} = \\
&= \sum_{t=1}^{T} \ln \frac{e^{2ax_t x_{t-1}/2\sigma^2} + e^{-2ax_t x_{t-1}/2\sigma^2}}{2} e^{-a^2 x_{t-1}^2/2\sigma^2} = \\
&= \sum_{t=1}^{T} \ln \cosh(ax_t x_{t-1}/\sigma^2) - a^2 x_{t-1}^2/2\sigma^2 = \\
&= \sum_{t=2}^{T} \ln \cosh(ax_t x_{t-1}/\sigma^2) - a^2 x_{t-1}^2/2\sigma^2
\end{aligned}
$$

Thus, the optimal decision is to guess that $S = 1$ whenever $g(\mathbf{x}) > 0$, and vice versa.

5 You have observed an outcome of the random sequence $\mathbf{X} = (X_1, \ldots, X_T)$ with real-valued scalar elements. The sequence is generated in a filtering process which depends on a hidden binary scalar random variable S, as

$$\begin{aligned} X_{-1} &= X_0 = 0 \\ X_t &= (1-S)aX_{t-1} + SaX_{t-2} + W_t \end{aligned}$$

Before time $t = 0$, the hidden switch variable S is selected randomly as either $S = 0$ or $S = 1$, with equal probability, and then remains constant for all t. Each W_t is a Gaussian random variable with zero mean and unknown variance σ^2. The filter parameter a is not known. The sequence elements W_t are hidden inside the signal source and cannot be observed directly, but we know that all these elements are statistically independent of each other.

Use the Expectation Maximization (EM) approach to find update equations to obtain improved parameter values $\lambda^{new} = (a^{new}, \sigma^{new})$, based on a previous parameter set $\lambda = (a, \sigma)$, and using the observed outcome sequence $\mathbf{x} = (x_1, \ldots, x_T)$. (5p)

Hint: Each step in the EM algorithm maximises the function

$$q(\lambda', \lambda) = E[\ln P(S, \mathbf{x}|\lambda')|\mathbf{x}, \lambda].$$

Solution:
We have only two possible values for the probability of the combination of hidden S and observable $\mathbf{X}$. As each X_t depends statistically on either the previous or pre-previous elements X_{t-1} or X_{t-2}, the probabilities can be calculated as

$$g(0,\lambda) = P(S = 0 \cap (x_1, \ldots, x_T)|\lambda) = 0.5 \prod_{t=1}^{T} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t - a x_{t-1})^2}{2\sigma^2}}$$

$$g(1,\lambda) = P(S = 1 \cap (x_1, \ldots, x_T)|\lambda) = 0.5 \prod_{t=1}^{T} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t - a x_{t-2})^2}{2\sigma^2}}$$

Thus, the two conditional probabilities of the hidden variable S, given the observed sequence, are

$$\gamma_0 = P(S = 0|(x_1, \ldots, x_T), \lambda) = \frac{g(0,\lambda)}{g(0,\lambda) + g(1,\lambda)}$$

$$\gamma_1 = \frac{g(1,\lambda)}{g(0,\lambda) + g(1,\lambda)} = 1 - \gamma_0$$

Using these results, we can define the EM help function as

$$\begin{aligned} q(\lambda', \lambda) =& E\left[\ln g(S, \lambda')\right] = \\ =& \gamma_0 \ln g(0, \lambda') + \gamma_1 \ln g(1, \lambda') = \\ =& \ln 0.5 + \gamma_0 \sum_{t=1}^{T} \left(\ln \frac{1}{\sqrt{2\pi}\sigma'} - \frac{(x_t - a' x_{t-1})^2}{2\sigma'^2} \right) + \gamma_1 \sum_{t=1}^{T} \left(\ln \frac{1}{\sqrt{2\pi}\sigma'} - \frac{(x_t - a' x_{t-2})^2}{2\sigma'^2} \right) = \\ =& \ln 0.5 - T \ln \sqrt{2\pi}\sigma' - \frac{1}{2\sigma'^2} \sum_{t=1}^{T} \gamma_0 (x_t - a' x_{t-1})^2 + \gamma_1 (x_t - a' x_{t-2})^2 \end{aligned}$$

To maximize the function we must simultaneously satisfy the two equations

$$\begin{aligned}
0 = \frac{\partial q}{\partial a'} =& -\frac{1}{2\sigma'^2}\sum_{t=1}^{T} -2\gamma_0(x_t - a'x_{t-1})x_{t-1} - 2\gamma_1(x_t a' x_{t-2})x_{t-2} = \\
=& \frac{1}{\sigma'^2}\sum_{t=1}^{T} \gamma_0 x_t x_{t-1} + \gamma_1 x_t x_{t-2} - \gamma_0 a' x_{t-1}^2 - \gamma_1 a' x_{t-2}^2; \\
0 = \frac{\partial q}{\partial \sigma'} =& -\frac{T}{\sigma'} + \frac{1}{\sigma'^3}\sum_{t=1}^{T} \gamma_0(x_t - a'x_{t-1})^2 + \gamma_1(x_t - a'x_{t-2})^2;
\end{aligned}$$

with solutions

$$\begin{aligned}
a^{new} =& \frac{\gamma_0 \sum_{t=1}^{T} x_t x_{t-1} + \gamma_1 \sum_{t=1}^{T} x_t x_{t-2}}{\gamma_0 \sum_{t=1}^{T} x_{t-1}^2 + \gamma_1 \sum_{t=1}^{T} x_{t-2}^2} = \\
=& \frac{\gamma_0 \sum_{t=2}^{T} x_t x_{t-1} + \gamma_1 \sum_{t=3}^{T} x_t x_{t-2}}{\gamma_0 x_{T-1}^2 + \sum_{t=1}^{T-2} x_t^2} \\
\sigma^{new} =& \sqrt{\frac{1}{T}\sum_{t=1}^{T} \gamma_0(x_t - a^{new}x_{t-1})^2 + \gamma_1(x_t - a^{new}x_{t-2})^2}
\end{aligned}$$