KTH Electrical Engineering

Solutions to Exam in Pattern Recognition 2E1395

Date:	Tuesday, Jan 17, 2006, 08.00 - 13.00
Place:	Q36.
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.
Results:	Tuesday, Jan 31, 2006.
Review:	At KTH-S3/ STEX, Osquldas v. 10.

Good Luck!

1 About 1 out of 100 000 people are infected by a type of virus that can be silent for many years without any obvious symptoms, but the virus can still cause serious illnes if it is not treated. Therefore, three different clinical laboratory tests, here called A, B, and C, are used to give indications on the infection. Each test method gives only a binary result, either “+” or “−”, with “+” indicating a high risk that the patient has the virus. In several studies, the *sensitivity* and *specificity* of the test methods have been determined as shown in the following table.

Test	Sensitivity	Specificity
A	0.8	0.9
B	0.9	0.8
C	0.8	0.5

The sensitivity is the conditional probability that the method indicates a “+” result, given that the patient is really infected with the virus. 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, B, and C. What is the probability that this person has the virus? (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 infection as a hidden discrete source state S with two possible outcomes: $S = 0$, if he is healthy, or $S = 1$ meaning he is infected. The *a priori* state probabilities are

$$\begin{aligned} P(S = 0) =& 0.99999 \\ P(S = 1) =& 0.00001 \end{aligned}$$

We regard the test results as an outcome, $\mathbf{x}$, of a feature vector $\mathbf{X}$, 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 $P(X_i = 0|S = 0)$.

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: We assume, quite reasonably, that the test results from different labs are statistically independent, given the true state. Therefore, the joint probabilities of states and observations are

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \cap S = 0) =& P(S = 0) \prod_{i=1}^{3} P(X_i = x_i|S = 0) = \\ =& 0.99999 \cdot (1 - 0.9) \cdot 0.8 \cdot (1 - 0.5) = 0.0399996 \\ P(\mathbf{X} = \mathbf{x} \cap S = 1) =& P(S = 1) \prod_{i=1}^{3} P(X_i = x_i|S = 1) = \\ =& 0.00001 \cdot 0.8 \cdot (1 - 0.9) \cdot 0.8 = 0,00000064 \end{aligned}$$

The conditional probability that the patient is infected is, thus,

$$P(S = 1|\mathbf{X} = \mathbf{x}) = \frac{0.00000064}{0.00000064 + 0.0399996} \approx 1.6 \cdot 10^{-5}$$

Decision: Thus, we must conclude that the probablity is still extremely small that the patient is infected. (This would have been the outcome even if all three lab results had been positive. To give more reliable results, the sensitivity and specificity of the lab methods must be improved.)

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) It is possible to design an optimal minimum-risk classifier for a situation with N_s source states and N_d possible decisions, only if $N_s \leq N_d$.

Solution:
FALSE. The minimum-risk classifier is general and can be used for any N_s, either greater or less than N_d.

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

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix};$$

Solution:
FALSE. The state probability distribution converges asymptotically towards $(0.5, 0.5)$ because of the symmetry of the transition probability matrix, but the initial probabilities are not equal for the two states. Thus, it is *not* stationary. The HMM is ergodic, however, because all possible transitions have non-zero probabilities.

(c) Given an observed output sequence $\mathbf{x} = (x_1, \ldots, x_T)$ from a Hidden Markov Model λ, the result of the Viterbi algorithm is sufficient to determine the most probable state at any "time" t, i.e.

$$\hat{i}_t = \underset{i}{\operatorname{argmax}} P\left(S_t = i|(x_1, \ldots, x_T), \lambda\right)$$

Solution:
FALSE. The Viterbi algorithm determines the single most probable state *sequence*, but the most probable state, as defined in the problem, can get its high probability by contributions from many different state sequences and may therefore not be included in the most probable single Viterbi sequence.

(d) An optimal classifier can be designed using discriminant functions $g_i(\mathbf{x})$, if the number of such discriminant functions is equal to the number of possible decisions.

Solution:
TRUE. This has been shown to be a general optimal solution. (In special cases the number of discriminant functions can be reduced.)

(e) In a Hidden Markov Model used to characterise a sequence of continous-valued scalar observations $\mathbf{x} = (x_1, \dots, x_t, \dots)$, the state-conditional output probability density functions $b_j(x)$ can all be Gaussian (normal) density functions and the functions may be *identical* for two or more different states, i.e. $b_j(x) = b_k(x)$ for states $j \neq k$.

Solution:
TRUE. Nothing forbids us to have identical density functions for different states. (This may be useful in practice, for example if the same sound occurs at several positions in a spoken word.)

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

The initial state probability vector is

$$q = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \text{with elements } P(S_1 = i).$$

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 $P(X_2 = 1)$. (2p)

Solution:

$$\begin{aligned} P(X_2 = 1) = & \sum_{j=1}^{2} P(S_2 = j \cap X_2 = 1) = \\ = & \sum_{i=1}^{2} \sum_{j=1}^{2} P(S_1 = i \cap S_2 = j) P(X_2 = 1 | S_2 = j) = \\ = & q_1(a_{11}b_{11} + a_{12}b_{21}) + q_2(a_{21}b_{11} + a_{22}b_{21}) = \\ = & 0.5 \cdot (0.6 \cdot 0.1 + 0.4 \cdot 0.6) + 0.5 \cdot (0.1 \cdot 0.1 + 0.9 \cdot 0.6) = \\ = & 0.425 \end{aligned}$$

(b) Calculate $P(X_2 = 1 | S_1 = 1 \cap X_1 = 1 \cap X_3 = 1 \cap S_4 = 1 \cap X_4 = 1)$. (3p)

Solution:

S_2 is conditionally independent on X_1, given S_1, and on X_4, given S_4, because of the Markov property. Therefore,

$$\begin{aligned}
P(X_2 = 1 | S_1 = 1 \cap X_1 = 1 \cap X_3 = 1 \cap S_4 = 1 \cap X_4 = 1) = \\
=& P(X_2 = 1 | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \\
=& \sum_{i=1}^{2} P(S_2 = i \cap X_2 = 1 | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \\
=& \sum_{i=1}^{2} P(X_2 = 1 | S_2 = i) P(S_2 = i | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \\
=& \sum_{i=1}^{2} b_{i1} P(S_2 = i | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1)
\end{aligned}$$

Here, by Bayes' rule,

$$P(S_2 = i | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \frac{1}{c} P(S_2 = i \cap X_3 = 1 \cap S_4 = 1 | S_1 = 1)$$

where c is just the normalizing factor

$$c = \sum_{k=1}^{2} P(S_2 = k \cap X_3 = 1 \cap S_4 = 1 | S_1 = 1)$$

We define a help vector $\mathbf{g}$ with elements, for i =1 and 2.

$$\begin{aligned}
g_i =& P(S_2 = i \cap X_3 = 1 \cap S_4 = 1 | S_1 = 1) = \\
=& \sum_{j=1}^{2} P(S_2 = i \cap S_3 = j \cap X_3 = 1 \cap S_4 = 1 | S_1 = 1) = \\
=& a_{1i} \sum_{j=1}^{2} a_{ij} b_{j1} a_{j1}
\end{aligned}$$

Denoting element-wise vector multiplication by $\circ$ we can express this as

$$\begin{aligned}
\mathbf{g} &= \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix} \circ A \left(\begin{pmatrix} 0.1 \\ 0.6 \end{pmatrix} \circ \begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix} \right) = \\
&= \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix} \circ \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.06 \\ 0.06 \end{pmatrix} = \\
&= \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix} \circ \begin{pmatrix} 0.06 \\ 0.06 \end{pmatrix} = \\
&= \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix} 0.06
\end{aligned}$$

Thus, after normalization with the factor $c = \sum_k g_k = 0.06$,

$$P(S_2 = * | S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$$

and the desired result is, finally,

$$\begin{aligned}
P(X_2 = 1|S_1 = 1 \cap X_1 = 1 \cap X_3 = 1 \cap S_4 = 1 \cap X_4 = 1) = \\
=&P(X_2 = 1|S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \\
=&\sum_{i=1}^{2} b_{i1} P(S_2 = i|S_1 = 1 \cap X_3 = 1 \cap S_4 = 1) = \\
=&0.1 \cdot 0.6 + 0.6 \cdot 0.4 = \\
=&0.30
\end{aligned}$$

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. You have observed an output sequence $\mathbf{x} = (x_1, \ldots, x_T)$. The output sequence is generated as

$$\begin{aligned}
X_{-1} = X_0 =&0 \\
\text{in source } S = 1\text{:} \quad X_t =&aX_{t-1} + W_t, \quad t = 1, \ldots, T \\
\text{in source } S = 2\text{:} \quad X_t =&aX_{t-2} + W_t, \quad t = 1, \ldots, T
\end{aligned}$$

The filter parameter a is known and equal in both sources. Each sample W_t is a Gaussian random variable with zero mean and known variance σ^2. 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 a given observed sequence. Simplify the decision criterion as far as possible, and show that optimal decisions can be made using only a scalar decision variable calculated as a sum of exactly $2T - 2$ terms, each containing an autocorrelation-style product of the form $x_n x_m$. (5p)

Hint: Note that

$$\begin{aligned}
P(\mathbf{X}) =&P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdot \ldots \cdot P(X_T|X_1, \ldots, X_{T-1}) = \\
=&P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdot \ldots \cdot P(X_T|X_{T-2}, X_{T-1})
\end{aligned}$$

because each element in the sequence depends statistically on only one of the two nearest preceding elements.

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

In both sources, every output sample X_t is conditionally dependent on only the previous sample X_{t-1} or X_{t-2}. Given previous observations $X_{t-k} = x_{t-k}$, the only remaining random component of X_t is W_t, so X_t is then conditionally Gaussian with variance σ^2 and zero mean.

Therefore, optimal preliminary discriminant functions can be defined as

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

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

$$\begin{aligned}
g(\mathbf{x}) =& \ln g_1(\mathbf{x}) - \ln g_2(\mathbf{x}) \\
=& \sum_{t=1}^{T} \frac{-(x_t - a x_{t-1})^2 + (x_t - a x_{t-2})^2}{2\sigma^2} = \\
=& \sum_{t=1}^{T} \frac{+2a x_t x_{t-1} - 2a x_t x_{t-2} - a^2 x_{t-1}^2 + a^2 x_{t-2}^2}{2\sigma^2} = \\
=& \frac{1}{2\sigma^2} \left(\sum_{t=1}^{T} 2a x_t x_{t-1} - \sum_{t=1}^{T} 2a x_t x_{t-2} - a^2 \sum_{t=1}^{T} x_{t-1}^2 + a^2 \sum_{t=1}^{T} x_{t-2}^2 \right) = \\
=& \frac{1}{2\sigma^2} \left(\sum_{t=2}^{T} 2a x_t x_{t-1} - \sum_{t=3}^{T} 2a x_t x_{t-2} - a^2 \sum_{u=1}^{T-1} x_u^2 + a^2 \sum_{u=1}^{T-2} x_u^2 \right) = \\
=& \frac{a}{\sigma^2} \left(\sum_{t=2}^{T} x_t x_{t-1} - \sum_{t=3}^{T} x_t x_{t-2} \right) - \frac{a^2 x_{T-1}^2}{2\sigma^2}
\end{aligned}$$

Here we have simplified the sum by using the known values $x_{-1} = x_0 = 0$. The resulting discriminant function is a sum with $2T - 2$ terms. The optimal decision is to guess that $S = 1$ whenever $g(\mathbf{x}) > 0$, and vice versa. The single discriminant function is clearly only a weighted linear combination of the observed sequence elements.

5 The random sequence $\mathbf{X} = (X_1, \ldots, X_T)$ is generated by an ergodic hidden Markov model (HMM) with N states and known initial probability vector and transition probability matrix. Each element X_t is a scalar continuous-valued non-negative random variable with an *exponential* state-conditional distribution. The state-conditional probability density function for X_t, given that the HMM state is $S_t = j$ at any "time" t, can therefore be expressed as

$$f_{X_t|S_t}(x|j) = b_j(x) = \begin{cases} c_j e^{-c_j x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Thus, each state-conditional density function is completely defined by a parameter c_j. Only crude approximations are known for these parameters. Fortunately, you have observed a very long outcome sequence $\mathbf{x} = (x_1, \ldots, x_T)$. Use the Expectation Maximisation (EM) approach to determine update equations to obtain improved parameter values c_j^{new}, based on the previous parameter set $\mathbf{c} = (c_1, \ldots, c_N)$ and the known HMM initial and transition probabilities, and using the observed training sequence $\mathbf{x} = (x_1, \ldots, x_T)$. (5p)

Hint: Each step in the EM algorithm maximises the function

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

where $\mathbf{S} = (S_1, \ldots, S_T)$ is the hidden state sequence. You may also use known results of the Forward and Backward algorithms without any derivation or proof.

Solution:

The complete source model can be denoted $\lambda = (q, A, \mathbf{c})$, where the Markov initial probability vector q and transition probability matrix A are assumed to be known and fixed parameters. Only the output parameters $\mathbf{c}$ need to be adapted to the data.

For any t we have conditional density functions for X_t as

$$f_{X_t|S_t}(x_t|j; \mathbf{c}) = c_j e^{-c_j x_t}$$

for any state j at time t. Given any particular state sequence $\mathbf{S} = (i_1, \ldots, i_T)$, the observations are independent, because of the fundamental HMM assumption. For the EM algorithm we need to calculate

$$\begin{aligned}\ln P(\mathbf{S} = (i_1, \ldots, i_T) \cap \mathbf{x}|q, A, \mathbf{c}') =& \ln \prod_{t=1}^{T} c'_{i_t} e^{-c'_{i_t} x_t} P(\mathbf{S} = (i_1 \ \ldots, i_T)|q, A) = \\ =& \sum_{t=1}^{T} \ln c'_{i_t} - c'_{i_t} x_t + \underbrace{\ln P(\mathbf{S} = (i_1 \ \ldots, i_T)|q, A)}_{z(i_1, \ldots, i_T)}\end{aligned}$$

where the log-probability $z(i_1, \ldots, i_T)$ depends only on the fixed Markov parameters and does not depend at all on the variable parameters $\mathbf{c}'$.

We will also need the known results of the forward-backward algorithms, calculated using the observed sequence $\mathbf{x}$ and the previous parameter vector $\mathbf{c}$:

$$\gamma_{t,j} = P(S_t = j|\mathbf{x}; \mathbf{c})$$

Using these results, we define the EM help function as the expected value expressed as a weighted sum over all possible state sequences

$$q(\mathbf{c}',\mathbf{c}) = \sum_{i_1}\sum_{i_2}\dots\sum_{i_T}\sum_{t=1}^{T}(\ln c'_{i_t} - c'_{i_t}x_t + z(i_1,\dots,i_T))P((i_1\dots i_T)|\mathbf{x},\mathbf{c}) =$$

$$= \sum_{t=1}^{T}\sum_{i_t}(\ln c'_{i_t} - c'_{i_t}x_t)P(S_t = i_t|\mathbf{x},\mathbf{c})\cdot$$

$$\cdot \underbrace{\sum_{i_1}\dots\sum_{i_{t-1}}\sum_{i_{t+1}}\dots\sum_{i_T}P((i_1\dots i_{t-1},i_{t+1}\dots i_T)|\mathbf{x},i_t,\mathbf{c})}_{=1} + \cdots =$$

$$= \sum_{t=1}^{T}\sum_{i}(\ln c'_i - c'_i x_t)\gamma_{t,i} + \cdots$$

where $\cdots$ represents the remaining terms that do not depend on $\mathbf{c}'$.

To maximize the function we must satisfy the equations, for every $j = 1,\dots,N$:

$$0 = \frac{\partial q}{\partial c'_j} = \sum_{t=1}^{T}\gamma_{t,j}(\frac{1}{c'_j} - x_t)$$

with solution

$$c_j^{new} = c'_j = \frac{\sum_{t=1}^{T}\gamma_{t,j}}{\sum_{t=1}^{T}\gamma_{t,j}x_t}$$

This update formula seems intuitively reasonable, as the expected value of the exponential distribution is $1/c_j$, and the derived new estimate $1/c_j^{new}$ is simply a weighted average of the observed $\mathbf{x}$ sequence. The weight factors $\gamma_{t,j}$ were calculated as the conditional probability that observation x_t was produced by state no. j.