KTH Electrical Engineering

Solutions to Exam in Pattern Recognition EN2200

Date: Monday, Oct 19, 2009, 08:00 – 13:00

Place: Q24, V01.

Allowed: Beta (or corresponding), calculator with empty memory. No notes!

Grades: A: at least 22p; B: 19p; C: 16p; D: 13p; E: 10p; out of total 25p.

Language: Optional: Swedish or English.

Results: Friday Nov 6, 2009.

Review: At KTH-S3/ STEX, Osquldas v. 10.

Good Luck!

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

1 In a given pattern-classification application the signal source can be in one of two states, here called $S = 1$ and $S = 2$. The two source states are known to occur with equal probabilities. You can observe a feature vector $\boldsymbol{X} = (X_1, X_2)^T$ with two elements. Depending on the source state $S = i$, the feature vector has a Gaussian conditional distribution, defined by the mean vector μ_i and covariance matrix C_i, with known values

$$\mu_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$
$$\mu_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

(a) Design an optimal classifier that can guess the source state with minimum error probability, and simplify the classifier to show that it is *possible* to make optimal decisions using a *linear* discriminant function of the type $g(x_1, x_2) = ax_1 + bx_2 + c$, together with a threshold mechanism. (4p)
Solution: As both source alternatives are equally probable, we use the *Maximum Likelihood* decision rule. The covariance matrices are equal, $C_1 = C_2 = C$, and we can define a single discriminant function simply as

$$\begin{aligned} g(\boldsymbol{x}) = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|1) - \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|2) = \\ &= (\boldsymbol{x} - \mu_2)^T C^{-1}(\boldsymbol{x} - \mu_2)/2 - (\boldsymbol{x} - \mu_1)^T C^{-1}(\boldsymbol{x} - \mu_1)/2 = \\ &= \mu_1^T C^{-1}\boldsymbol{x} - \mu_2^T C^{-1}\boldsymbol{x} - \mu_1^T C^{-1}\mu_1/2 + \mu_2^T C^{-1}\mu_2/2 = \\ &= (\mu_1 - \mu_2)^T C^{-1}\boldsymbol{x} - (\mu_1 - \mu_2)C^{-1}(\mu_1 + \mu_2)/2 = \\ &= (\mu_1 - \mu_2)^T C^{-1}(\boldsymbol{x} - (\mu_1 + \mu_2)/2) \end{aligned}$$

Then, the optimal classifier decides $S = 1$, whenever $g(\boldsymbol{x}) > 0$ and vice versa. With the given covariance we have

$$C^{-1} = \frac{1}{15}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$$

and

$$g(\boldsymbol{x}) = \frac{1}{15}\begin{pmatrix} 4 & 4 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\left(\boldsymbol{x} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix} 1 & 1 \end{pmatrix}\left(\boldsymbol{x} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \frac{1}{3}(x_1 + x_2 - 2)$$

Thus, the classifier will decide $S = 1$, whenever $x_1 + x_2 > 2$, and vice versa.

(b) What is the the conditional probability that source $S = 1$ was active, given an observed feature vector $(0, 0)$, i.e. $P(S = 1|\boldsymbol{X} = (0, 0)^T)$? (1p)
Solution: Omitting factors that are equal for both feature distributions, we find log-likelihood values

$$L_i = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|i) = -(\boldsymbol{x} - \mu_i)^T C^{-1}(\boldsymbol{x} - \mu_i)/2 + \text{const.}$$

with

$$L_1 = -\frac{1}{30}\begin{pmatrix} -3 & -3 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} -3 \\ -3 \end{pmatrix} = -3$$
$$L_2 = -\frac{1}{30}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = -\frac{1}{3}$$

Thus the conditional probability for $S = 1$ is

$$P(S = 1|\boldsymbol{X} = (0,0)^T) = \frac{e^{-3}}{e^{-3} + e^{-1/3}} \approx 0.065$$

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) When designing an optimal classifier for a source with N_s source states and N_d decision alternatives, using a feature vector with K elements, the optimal performance can always be achieved using some K in the interval $1 \leq K \leq \max(N_s, N_d)$.

Solution: FALSE. Any number of features can be optimal, depending on the application.

(b) Given an observed output sequence $\underline{x} = (x_1, \ldots, x_T)$ from a Hidden Markov Model λ, we can use the results of the Forward algorithm to calculate the conditional probability density

$$P\left((x_{t+1}, \ldots, x_T)|(x_1, \ldots, x_t), \lambda\right)$$

for any $1 \leq t < T$.

Solution: TRUE. The Forward algorithm can calculate a sequence of scale factors defined as $c_t = P\left(x_t|x_1, \ldots, x_{t-1}, \lambda\right)$ for any t. Therefore, using Bayes' rule, we can calculate

$$P\left((x_{t+1}, \ldots, x_T)|(x_1, \ldots, x_t), \lambda\right) = \prod_{u=t+1}^{T} c_u$$

for any $t < T$.

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

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 \\ 0.05 & 0.95 \end{pmatrix};$$

Solution: FALSE. $q \neq A^T q$..

(d) For a scalar random variable X with Gaussian Mixture Model (GMM) probability density function

$$f_X(x) = \sum_{m=1}^{M} w_m g_m(x),$$

the combined density function must be limited as $0 \leq f_X(x) \leq 1$ for any x.
Solution: FALSE. Probability density functions can have any non-negative value, $0 \leq f_X(x)$, but no upper limit.

(e) It is possible to design classifier discriminant functions, that normally use all K elements of the feature vector, such that the classifier can allow one feature element to be *missing* but still make optimal use of the remaining features (although possibly with reduced performance).
Solution: TRUE. The discriminant functions only need to include a pre-designed variant that uses only $K - 1$ features.

3 You can observe some elements of 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

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

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 $P(X_2 = 1\ )$. (2p)
Solution: The Markov chain is stationary; the stationarity is verified by showing that $A^T q = q$. Therefore, the unconditional state probabilities are $P(S_t = i) = P(S_1 = i) = q_i$, for any t.

$$\begin{aligned} P(X_2 = 1) &= \sum_i P(X_2 = 1 \cap S_2 = i) \\ &= P(X_2 = 1 | S_2 = 1) P(S_2 = 1) + P(X_2 = 1 | S_2 = 2) P(S_1 = 2) \\ &= b_{11} \times q_1 + b_{21} \times q_2 \\ &= 0.1 \times 0.2 + 0.7 \times 0.8 \\ &= 0.58 \end{aligned}$$

(b) Calculate $P(S_{15} = 1 | X_{15} = 2 \cap S_{16} = 1 \cap X_{16} = 3 \cap S_{17} = 2)$. (3p)
Solution: Given X_{15} and S_{16}, S_{15} is statistically independent of X_{16} and S_{17}, so we have

$$\begin{aligned} P(S_{15} = 1 | X_{15} = 2 \cap S_{16} = 1 \cap X_{16} = 3 \cap S_{17} = 2) &= P(S_{15} = 1 | X_{15} = 2 \cap S_{16} = 1) \\ &= \frac{P(S_{15} = 1 \cap X_{15} = 2 \cap S_{16} = 1)}{P(X_{15} = 2 \cap S_{16} = 1)} \\ &= \frac{P(S_{15} = 1 \cap X_{15} = 2 \cap S_{16} = 1)}{\sum_i P(S_{15} = i \cap X_{15} = 2 \cap S_{16} = 1)} \end{aligned}$$

$$\begin{aligned} P(S_{15} = i \cap X_{15} = 2 \cap S_{16} = 1) &= P(X_{15} = 2 | S_{15} = i \cap S_{16} = 1) P(S_{15} = i \cap S_{16} = 1) \\ &= P(X_{15} = 2 | S_{15} = i) P(S_{16} = 1 | S_{15} = i) P(S_{15} = i) \\ &= b_{i2} \times a_{i1} \times q_i \end{aligned}$$

$P(S_{15} = i) = q_i$ holds since we have a stationary distribution, the stationarity is verified by showing that $A^T q = q$. The final solution is

$$\begin{aligned}
P(S_{15} = 1 | X_{15} = 2 \cap S_{16} = 1) &= \frac{b_{12} \times a_{11} \times q_1}{b_{12} \times a_{11} \times q_1 + b_{22} \times a_{21} \times q_2} \\
&= \frac{0.4 \times 0.6 \times 0.2}{0.4 \times 0.6 \times 0.2 + 0.2 \times 0.1 \times 0.8} \\
&= 0.75
\end{aligned}$$

4 A signal source is known to generate independent scalar random numbers, each with a Gaussian distribution with mean $\mu = 0$ and unknown variance $\sigma^2 = 1/W$. You have observed a sequence of random numbers $\underline{x} = (x_1, \ldots, x_N)$ generated from this source, and you will now apply Bayesian learning for the unknown inverse-variance parameter W.

We assume a *gamma* prior density function for W, defined as

$$f_W(w) \propto w^{a_0 - 1} e^{-b_0 w}$$

For a nearly non-informative prior distribution we assume hyper-parameter values $a_0 = b_0 =$ some small value, greater than zero.

Show that the posterior density function for W, given the observed sequence, also has the gamma-distribution form,

$$f_{W|\underline{X}}(w|\underline{x}) \propto w^{a_N - 1} e^{-b_N w}$$

and determine the posterior hyper-parameters a_N and b_N, expressed in terms of the observed sequence $\underline{x} = (x_1, \ldots, x_N)$. Discuss briefly how the posterior distribution for W is related to the observed variance of the given sample $\underline{x} = (x_1, \ldots, x_N)$, in the asymptotic limit with $a_0 = b_0 \to 0$. (5p)

Hint: The normalized probability density function for a gamma-distributed random variable W can always be written as

$$f_W(w) = \frac{b^a}{\Gamma(a)} w^{a-1} e^{-bw}$$

where $a > 0$, $b > 0$, and $\Gamma(\;)$ is the Gamma function. The most interesting characteristics of the gamma distribution are

$$\begin{aligned}
E[W] &= \frac{a}{b} \\
\text{var}[W] &= \frac{a}{b^2} \\
\operatorname*{argmax}_w f_W(w) &= \frac{a-1}{b}, \quad \text{for } a \geq 1
\end{aligned}$$

Solution: The conditional probability density for each observed random number x_n is Gaussian with variance $1/w$:

$$f_{X|W}(x_n|w) = \frac{\sqrt{w}}{\sqrt{2\pi}} e^{-x_n^2 w/2}$$

The samples in the sequence are conditionally independent, given the variance. Thus, the probability density for the complete observed sequence is

$$f_{\underline{X}|W}(x_1, \ldots, x_N|w) = \prod_{n=1}^{N} \frac{\sqrt{w}}{\sqrt{2\pi}} e^{-x_n^2 w/2} = \frac{w^{N/2}}{(2\pi)^{N/2}} e^{-\sum_n x_n^2 w/2}$$

Following the general Bayesian-learning approach, we find the posterior parameter probability density as

$$f_{W|\underline{X}}(w|\underline{x}) \propto f_{\underline{X}|W}(x_1,\ldots,x_N|w) f_W(w) \propto w^{N/2} e^{-\sum_n x_n^2 w/2} w^{a_0-1} e^{-b_0 w} = w^{a_N-1} e^{-b_N w}$$

Thus, the posterior parameter density also has the form of a gamma distribution, with

$$\begin{aligned} a_N =& a_0 + \frac{N}{2} \\ b_N =& b_0 + \frac{1}{2} \sum_n x_n^2 \end{aligned}$$

Given the observations, the expected value of the inverse variance is then

$$E[W] = \frac{a_N}{b_N} = \frac{a_0 + \frac{N}{2}}{b_0 + \frac{1}{2}\sum_n x_n^2}$$

In the asymptotic limit with very large N, equivalent to $a_0 = b_0 \to 0$, the expected value is

$$E[W] \to \frac{N}{\sum_n x_n^2}$$

The inverse variance is then simply the inverse of the sample mean square value, which also equals the inverse ML estimate for the variance.

5 In an attempt to design a *speaker-independent word-recognition* classifier, you have trained $N \times M$ different hidden Markov models (HMM); one separate model λ_{nm} for word type $W = n \in \{1, \ldots, N\}$ and speaker $S = m \in \{1, \ldots M\}$. Several training examples were used for each word type and each speaker.

You will now design a classifier to identify sequences of J words, $\underline{W} = (W_1, \ldots, W_J)$, where all words in the sequence are pronounced by the same speaker. The jth recorded test word is represented as usual by a stream of feature vectors, denoted as

$$\underline{\boldsymbol{x}}_j = (\boldsymbol{x}_{j1}, \ldots, \boldsymbol{x}_{jT_j})$$

You already have a procedure (like your Matlab `HMM/logprob`) that calculates the log-probability of any single recorded test word for any of the known models, i.e.,

$$L_{jnm} = \ln P(\underline{\boldsymbol{x}}_j|\lambda_{nm}), \quad \text{for } j = 1, \ldots, J; \quad n = 1, \ldots, N; \quad m = 1, \ldots, M.$$

Construct a decision rule to identify the most probable sequence $\underline{\hat{w}} = \widehat{(w_1, \ldots, w_J)}$ using your calculated log-probabilities L_{jnm}.

Any of the speakers is engaged at random with equal probabilities $1/M$, and each of the possible word types occurs with equal probabilities $1/N$, independently of which speaker is engaged, and independently of the other words in the sequence. The feature distributions are assumed to be conditionally independent across all words in a sequence, given the word types and the speaker. (5p)

Hint: For optimal performance, the classifier should utilize the knowledge that all observed test words were recorded from the *same* speaker, although it is not known who among the possible

speakers actually pronounced the test words.
Solution: As all word sequences are equally probable, we can use the *Maximum-Likelihood* decision rule. The pronunciations of different words in the sequence are conditionally independent, given a particular speaker. The likelihood for any word sequence $\underline{w} = (w_1, \ldots, w_J)$ is therefore

$$Q(w_1, \ldots, w_J) = P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J | w_1, \ldots, w_J) = \sum_{m=1}^{M} P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J \cap S = m | w_1, \ldots, w_J) =$$

$$= \sum_{m=1}^{M} P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J | (w_1, \ldots, w_J) \cap S = m) P(S = m) =$$

$$= \sum_{m=1}^{M} \frac{1}{M} \prod_{j=1}^{J} e^{L_{j,w_j,m}} = \frac{1}{M} \sum_{m=1}^{M} e^{\sum_{j=1}^{J} L_{j,w_j,m}}$$

This likelihood must be evaluated for each of all the possible word sequences, and the highest value is selected. Thus, the final decision rule is

$$\underline{\hat{w}} = \operatorname*{argmax}_{w_1, \ldots, w_J} Q(w_1, \ldots, w_J)$$