KTH Electrical Engineering

Solutions to Exam in Pattern Recognition EN2200

Date: Thursday, Oct 23, 2008, 08:00 – 13:00

Place: E31.

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: Friday Nov 7, 2008.

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

Good Luck!

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

1 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 for a source with N_s source states using an observed K-dimensional feature vector, the number of alternative decisions must be $N_d \leq N_s$.

Solution: FALSE. Any number of decision categories are possible, depending on the application.

(b) It is possible to classify observed K-dimensional feature vectors $\boldsymbol{x}$ optimally, if some of the K feature elements have a discrete probability distribution while other features have a continuous distribution.

Solution: TRUE. The only requirement is that it is possible to calculate likelihood values $g_i($
$vcx)$ that are proportional to the combined probability mass/density for the observed vector, given the source category.

(c) Using a hidden Markov model λ and an observed sequence $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ we can calculate the probability of state $S_t = j$, given the partial observed sequence, as

$$\hat{\alpha}_{j,t} = P(S_t = j | (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t), \lambda)$$

using the forward algorithm. This result of the forward algorithm remains correct, even if the HMM uses a *scaled* version of the state-conditional probability density functions, i.e. $b_j(x_t) = h f_{X_t|S_t}(x|j)$, where h is a fixed but unknown scale factor.

Solution: TRUE. The scaled forward variables are normalized for each t, to precisely compensate for any scaling of the probability densities, and therefore still yield the correct state probabilities, as defined.

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

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.7 \\ 0.1 \\ 0.2 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 & 0 \\ 0.07 & 0.93 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

Solution: TRUE. Stationary, because $q = A^T q$. (It is non-ergodic because the Markov chain will stay in state 3 forever, if it happens to start in state 3.)

(e) In a Gaussian Mixture Model (GMM) with M components, all the GMM weight factors $w_m; m = 1 \ldots M$ must be limited as $0 \leq w_m \leq 1$.

Solution: TRUE. The sum of all weight factors must be equal to 1, and none can be negative.

2 In one 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 signal 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

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

(a) Design an optimal classifier that can guess the source state with minimum error probability, and show that optimal decisions can be made using a single scalar variable $Y = a|X_1| + b|X_2|$ and a simple threshold mechanism. Determine suitable values for a and b. (3p)

Solution:

Criterion: As both source states are equally probable, we use the ML criterion.

Feature distributions: The two conditional feature distributions are Gaussian, with zero means and

$$C_1^{-1} = \begin{pmatrix} 1/3 & 0 \\ 0 & 1 \end{pmatrix}; \quad C_2^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/3 \end{pmatrix}$$

Discriminant functions: We choose two discriminant functions as

$$g_i(\boldsymbol{x}) = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|i) = -\frac{1}{2}\boldsymbol{x}^T C_i^{-1}\boldsymbol{x} - \ln 2\pi\sqrt{\det C_i}, \quad i \in \{1, 2\}$$

Simplified: As there are only two alternative decisions, we can use a single discriminant function, as

$$\begin{aligned} g(\boldsymbol{x}) =& g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = -x_1^2/6 - x_2^2/2 + x_1^2/2 + x_2^2/6 \propto \\ &\propto x_1^2 - x_2^2 \end{aligned}$$

Decision: The optimal classifier must guess state 1, iff $x_1^2 \geq x_2^2$, or equivalently, iff $|x_1| \geq |x_2|$. Thus, a single decision variable $Y = |X_1| - |X_2|$ is optimal, i.e. with $a = 1$ and $b = -1$.

(b) Sketch the boundary between the decision regions of your classifier. (1p)

Solution: See fig. 1.

(c) Your boss (who did not pass the pattern-recognition course) has asked you to re-design the classifier using only one *single* feature, with minimal reduction in classifier performance. Do you choose to use only feature X_1, only X_2, or some other third feature $X_3 = f(X_1, X_2)$ that is a function of the original features? A brief motivation for your choice is required. (1p)

Solution: We just proved in part (a) that the combined feature $X_3 = |X_1| - |X_2|$ is optimal.

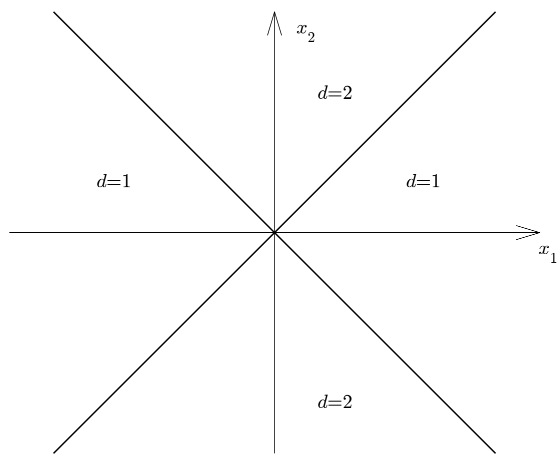

Figure 1: Decision regions for the optimal classifier.

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

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix}, \text{with elements } a_{ij} = P(S_{t+1} = j | S_t = i).$$

The output probability matrix is

$$B = \begin{pmatrix} 0.1 & 0.3 & 0.6 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(X_t = k | S_t = i).$$

(a) Calculate $P(X_2 = 1)$. (2p)
Solution:

$$\begin{aligned}
P(X_2 = 1) = &\sum_{j=1}^{2} P(S_2 = j \cap X_2 = 1) = \\
= &\sum_{j=1}^{2} \underbrace{P(X_2 = 1 | S_2 = j)}_{b_{j1}} P(S_2 = j) = \\
= &\sum_{j=1}^{2} b_{j1} \sum_{i=1}^{2} \underbrace{P(S_1 = i \cap S_2 = j)}_{q_i a_{ij}} = \\
= &b_{11}(q_1 a_{11} + q_2 a_{21}) + b_{21}(q_1 a_{12} + q_2 a_{22}) = \\
= &0.352
\end{aligned}$$

(b) Calculate $P(S_2 = 1 | X_1 = 3 \cap X_2 = 1 \cap S_3 = 2 \cap X_3 = 2 \cap S_4 = 1)$. (3p)
Solution: S_2 is conditionally independent of X_3 and S_4, given S_3, because of the Markov property. Therefore, and using Bayes' rule,

$$\begin{aligned}
P(S_2 = 1 | X_1 = 3 \cap X_2 = 1 \cap S_3 = 2 \cap X_3 = 2 \cap S_4 = 1) = \\
= &P(S_2 = 1 | X_1 = 3 \cap X_2 = 1 \cap S_3 = 2) = \\
= &\frac{P(X_1 = 3 \cap S_2 = 1 \cap X_2 = 1 \cap S_3 = 2)}{\sum_{j=1}^{2} P(X_1 = 3 \cap S_2 = j \cap X_2 = 1 \cap S_3 = 2)}
\end{aligned}$$

Here

$$\begin{aligned}
P(X_1 = 3 \cap S_2 = j \cap X_2 = 1 \cap S_3 = 2) = &\sum_{i=1}^{2} q_i b_{i3} a_{ij} b_{j1} a_{j2} = \\
= &\begin{cases} 0.0101, & j = 1 \\ 0.1021, & j = 2 \end{cases}
\end{aligned}$$

Thus,

$$P(S_2 = 1|X_1 = 3 \cap X_2 = 1 \cap S_3 = 2) = \frac{0.0101}{0.0101 + 0.1021} = 0.0904$$

4 A random generator produces a sequence of scalar random values $(X_1, \ldots, X_t, \ldots)$ as

$$X_t = cZ_t + dW_t$$

Here, c and d are real-valued known constants. The Z_t and W_t values cannot be observed directly. W_t is for every t a Gaussian random variable with mean 0 and variance 1. The random sequence $\underline{Z} = (Z_1, \ldots, Z_t, \ldots)$ contains discrete elements Z_t that can be either $Z_t = +1$ or $Z_t = -1$. All W_t values are statistically independent across different t, but the Z_t values have the following conditional probability-mass distribution:

$$\begin{aligned} P(Z_1 = +1) =& 1 \\ P(Z_t = +1|Z_{t-1} = +1) =& P(Z_t = -1|Z_{t-1} = -1) = s, \quad t \in \{2, 3, \ldots\} \end{aligned}$$

(a) Show that this signal source can be characterized as a hidden Markov model $\lambda = \{q, A, B\}$, by constructing explicit expressions for all components of the HMM. (1p)

Solution: We assign state 1 as $Z_t = +1$ and state 2 as $Z_t = -1$. Then,

$$\begin{aligned} q =& \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ A =& \begin{pmatrix} s & 1-s \\ 1-s & s \end{pmatrix} \\ b_j(x) =& \frac{1}{d\sqrt{2\pi}} e^{-\frac{(x-\mu_j)^2}{2d^2}}, \qquad \mu_j = \begin{cases} c, & j = 1 \\ -c, & j = 2 \end{cases} \end{aligned}$$

(b) Determine the HMM state probabilities for any $t \in \{1, 2, \ldots\}$. (2 p)
Hint: Express the state probability with a stationary term p_s and a deviation d_t from the stationary value, as

$$\begin{aligned} P(Z_t = +1) =& p_s + d_t \\ P(Z_t = -1) =& 1 - p_s - d_t \end{aligned}$$

Find the constant p_s, and determine d_t as a function of t.

Solution: Let us define $p_t = P(Z_t = +1) = p_s + d_t$. At any transition we have

$$\begin{aligned} p_{t+1} = p_s + d_{t+1} =& s(p_s + d_t) + (1-s)(1 - p_s - d_t) \\ p_s + d_{t+1} =& (2s-1)p_s + (2s-1)d_t + 1 - s \end{aligned}$$

For the stationary condition we have $d_{t+1} = d_t = 0$, which yields the equation

$$\begin{aligned} p_s =& (2s-1)p_s + 1 - s \\ p_s =& \frac{1-s}{2-2s} = \frac{1}{2} \end{aligned}$$

as expected, because of the state-transition symmetry. Thus, $d_1 = 1/2$, and for the transient probability component we obtain

$$\begin{aligned} d_{t+1} =& (2s-1)d_t \\ d_t =& d_1(2s-1)^{t-1} = \frac{1}{2}(2s-1)^{t-1} \end{aligned}$$

(c) Determine explicit expressions for

$$\begin{aligned} \mu_t =& E\left[X_t\right] \\ \sigma_t^2 =& \operatorname{var}\left[X_t\right] \end{aligned}$$

for any $t \in \{1, 2, \ldots\}$. (2 p)

Solution:

$$\begin{aligned} \mu_t =& E\left[X_t\right] = E\left[X_t | Z_t = +1\right] P(Z_t = +1) + E\left[X_t | Z_t = -1\right] P(Z_t = -1) = \\ =& c(p_s + d_t) - c(1 - p_s - d_t) = \\ =& 2cd_t + 2cp_s - c = \\ =& c(2s-1)^{t-1}; \\ \sigma_t^2 =& E\left[(X_t - \mu_t)^2\right] = \\ =& E\left[(X_t - \mu_t)^2 | Z_t = +1\right] P(Z_t = +1) + E\left[(X_t - \mu_t)^2 | Z_t = -1\right] P(Z_t = -1) = \\ =& E\left[(X_t - c + c - \mu_t)^2 | Z_t = +1\right] P(Z_t = +1) + E\left[(X_t + c - c + \mu_t)^2 | Z_t = -1\right] P(Z_t = -1) = \\ =& ((d^2 + (c - \mu_t)^2)(p_s + d_t) + ((d^2 + (c + \mu_t)^2)(1 - p_s - d_t) = \\ =& d^2 + c^2 + \mu_t^2 - 2c\mu_t(p_s + d_t)2 + 2c\mu_t = \\ =& d^2 + c^2 + \mu_t^2 - 2c\mu_t(2s-1)^{t-1} = \\ =& d^2 + c^2 - c^2(2s-1)^{2(t-1)} \end{aligned}$$

The mean starts at $+c$ and decreases asymptotically to zero, whereas the variance starts at d^2 and increases asymptotically to $d^2 + c^2$, which seems intuitively reasonable.

5 A sequence of random scalar values $\underline{X} = (X_1, \ldots, X_t, \ldots)$ is generated by the following auto-regressive filter process:

$$\begin{aligned} X_0 =& 0 \\ X_t =& aX_{t-1} + cW_t \end{aligned}$$

Here, a and c are constant parameters, and W_t is for each t a Gaussian random variable with mean 0 and variance 1, and all W_t are statistically independent across different t. The constant c is exactly known, but the value of a is unknown. You have observed an outcome sequence $\underline{x} = (x_1, \ldots, x_t, \ldots, x_T)$ generated by this random source.

You will now apply *Bayesian Learning* to determine to what extent the value of a can be known, after using the observed sequence. In this approach we assume that the parameter a is an outcome of a random variable A, but it remains constant for all t. Before the observation of $\underline{x}$, we express our total uncertainty about the parameter value by modeling A as a Gaussian random variable with mean 0 and a very large variance σ_0^2.

Determine the *posterior* conditional probability density function for A,

$$f_{A|\underline{X}}(a|\underline{x})$$

given the observed $\underline{x} = (x_1, \ldots, x_t, \ldots, x_T)$. Show that this density function has a Gaussian form, and determine its mean μ_T and variance σ_T^2, given the observed sequence. (5p)
Hint: For any given value of a and the observed previous value x_{t-1}, the X_t is a Gaussian random variable with conditional mean ax_{t-1} and conditional variance c^2. Determine the likelihood for the combined event $(\underline{X} = \underline{x} \cap A = a)$ and then identify this likelihood expression as a conditional density for A by regarding it as a function of a.

Solution: The prior density function for A is (disregarding unimportant constants)

$$f_A(a) \propto e^{-\frac{a^2}{2\sigma_0^2}}$$

The conditional density for the complete observed sequence, given any particular outcome of A is (disregarding constants again)

$$f_{\underline{X}|A}(x_1, \ldots, x_t, \ldots, x_T|a) \propto \prod_{t=1}^{T} e^{-\frac{(x_t - ax_{t-1})^2}{2c^2}} = \\ = e^{-\frac{1}{2c^2}\left(\sum_t x_t^2 - 2ax_tx_{t-1} + a^2x_{t-1}^2\right)}$$

The probability density of the parameter, given the observations, is then

$$f_{A|\underline{X}}(a|\underline{x}) \propto f_{\underline{X},A}(x_1, \ldots, x_t, \ldots, x_T, a) = f_{\underline{X}|A}(x_1, \ldots, x_t, \ldots, x_T|a) f_A(a) \propto \\ \propto e^{-\frac{1}{2c^2}\left(a^2(c^2/\sigma_0^2 + \sum_t x_{t-1}^2) - 2a\sum_t x_tx_{t-1} + \cdots\right)}$$

where the $\cdots$ in the exponent represents the remaining expression that is independent of a.

As the exponent is a quadratic expression in a, it is clear that the posterior density for A must be Gaussian. We simply denote its mean and variance after T observations by μ_T and σ_T^2, and express the posterior density as

$$f_{A|\underline{X}}(a|\underline{x}) \propto e^{-\frac{1}{2\sigma_T^2}\left(a^2 - 2\mu_T a + \cdots\right)}$$

and then just identify

$$\frac{1}{\sigma_T^2} = \frac{c^2/\sigma_0^2 + \sum_t x_{t-1}^2}{c^2}$$
$$\frac{\mu_T}{\sigma_T^2} = \frac{\sum_t x_t x_{t-1}}{c^2}$$

which yields, finally,

$$\mu_T = \frac{\sum_t x_t x_{t-1}}{c^2/\sigma_0^2 + \sum_t x_{t-1}^2}$$
$$\sigma_T^2 = \frac{c^2}{c^2/\sigma_0^2 + \sum_t x_{t-1}^2}$$

To account for the fact that we had no prior knowledge about A we can just let $\sigma_0 \to \infty$ in these expressions.