CORRECTIONS FOR THE CLASSIFICATION EXAM - THIRD YEAR - SPECIALITY SIGNAL AND IMAGE PROCESSING

Monday, November 28, 2016

Lecture notes and slides authorized

Exercice 1

We consider a classification problem with two classes ω_1 and ω_2 whose densities are

$$f(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-m_i)^2\right] \quad i=1,2 \tag{1}$$

with $x \in \mathbb{R}$, $\sigma > 0$ and $m_1 > m_2$.

1. (3 pts) Derive the Bayesian classification rule associated with this problem when we use the $0-1$ cost function and when the two classes have the prior probabilities $P(\omega_1) = P_1$ and $P(\omega_2) = P_2$. Interpret this result using the centroid distance rule when $P_1 = P_2$ and $P_1 > P_2$. Express the probability of error of this rule as a function of m_1, m_2, σ^2 and the cumulative distribution function of the $\mathcal{N}(0,1)$ Gaussian distribution denoted as F.

   Response: The Bayesian classifier accepts the class ω_1 (denoted as $d^*(x) = \omega_1$ if

   $$f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2)$$

   or equivalently if

   $$\ln[f(x|\omega_1)] + \ln[P(\omega_1)] \geq \ln[f(x|\omega_2)] + \ln[P(\omega_2)].$$

   Straightforward computations lead to

   $$d^*(x) = \omega_1 \Leftrightarrow \frac{m_1 - m_2}{\sigma^2}x \geq \frac{m_1^2 - m_2^2}{2\sigma^2} + \ln\left(\frac{P_2}{P_1}\right).$$

   Since $m_1 > m_2$, we obtain

   $$d^*(x) = \omega_1 \Leftrightarrow x \geq \frac{m_1 + m_2}{2} + \frac{\sigma^2}{m_1 - m_2}\ln\left(\frac{P_2}{P_1}\right).$$

   When the two classes are equiprobable, we have

   $$d^*(x) = \omega_1 \Leftrightarrow x \geq \frac{m_1 + m_2}{2}$$

   which is the centroid distance rule, i.e., the class ω_1 is accepted if x is closer to its centroid m_1 than to the other class centroid m_2. When $P_1 > P_2$, the class ω_1 is more likely than the class ω_2. In this case, the threshold

   $$S = \frac{m_1 + m_2}{2} + \frac{\sigma^2}{m_1 - m_2}\ln\left(\frac{P_2}{P_1}\right)$$

   is smaller than the centroid $\frac{m_1+m_2}{2}$ (since $\ln(P_2/P_1) < 0$ and $m_1 - m_2 > 0$), which corresponds to accepting the class ω_1 more often than in the equiprobable case. This property is in agreement with $P_1 > P_2$.

The error probability of the Bayesian classifier is defined as

$$P_e = P[d^*(X) = \omega_1 | X \in \omega_2] P(X \in \omega_2) + P[d^*(X) = \omega_2 | X \in \omega_1] P(X \in \omega_1)$$

or equivalently

$$P_e = P[X > S | X \in \omega_2] P_2 + P[X > S | X \in \omega_1] P_1.$$

In order to use the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution, we have to express the two probabilities as follows

$$P_e = P\left[\frac{X - m_2}{\sigma} > \frac{S - m_2}{\sigma} | \frac{X - m_2}{\sigma} \sim \mathcal{N}(0,1)\right] P_2 + P\left[\frac{X - m_1}{\sigma} < \frac{S - m_1}{\sigma} | \frac{X - m_1}{\sigma} \sim \mathcal{N}(0,1)\right] P_1.$$

Finally, we obtain

$$P_e = P_2 \left[1 - F\left(\frac{S - m_2}{\sigma}\right)\right] + P_1 F\left(\frac{S - m_1}{\sigma}\right)$$

with

$$\frac{S - m_2}{\sigma} = \frac{m_1 - m_2}{\sigma} + \frac{\sigma}{m_1 - m_2} \ln\left(\frac{P_2}{P_1}\right)$$

and

$$\frac{S - m_1}{\sigma} = \frac{m_2 - m_1}{\sigma} + \frac{\sigma}{m_1 - m_2} \ln\left(\frac{P_2}{P_1}\right).$$

2. (2 pts) Show that the Bayesian decision rule can be written as

$$d^*(x) = \omega_1 \Leftrightarrow g\left[a(x)\right] = \frac{1}{1 + \exp\left[a(x)\right]} \leq \frac{1}{2}$$

where

$$a(x) = \ln\left[\frac{f(x|\omega_1)P(\omega_1)}{f(x|\omega_2)P(\omega_2)}\right].$$

For the example of the previous question, derive the function $a(x)$ and prove that is is affine, i.e., $a(x) = a_1 x + a_2$, where a_1 and a_2 are two functions of $m_1, m_2, \sigma^2, P_1, P_2$ that you will determine.

Response: the Bayesian decision rule

$$d^*(x) = \omega_1 \Leftrightarrow f(x|\omega_1)P(\omega_1) \geq f(x|\omega_2)P(\omega_2)$$

is equivalent to

$$d^*(x) = \omega_1 \Leftrightarrow a(x) = \ln\left[\frac{f(x|\omega_1)P(\omega_1)}{f(x|\omega_2)P(\omega_2)}\right] \geq 0$$

or, by using the fact that the function g is a decreasing function

$$d^*(x) = \omega_1 \Leftrightarrow g\left[a(x)\right] \leq g(0) = \frac{1}{2}.$$

After replacing the expressions of the densities $f(x|\omega_1)$ and $f(x|\omega_1)$ in the expression of $a(x)$, we obtain

$$a(x) = \ln\left[\frac{\exp\left(-\frac{1}{2\sigma^2}(x - m_1)^2\right) P(\omega_1)}{\exp\left(-\frac{1}{2\sigma^2}(x - m_2)^2\right) P(\omega_2)}\right]$$

i.e.,

$$a(x) = \frac{m_1 - m_2}{\sigma^2} x + \frac{m_2^2 - m_1^2}{2\sigma^2} + \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right]$$

that is indeed an affine function of x with

$$a_1 = \frac{m_1 - m_2}{\sigma^2} \quad \text{and} \quad a_2 = \frac{m_2^2 - m_1^2}{2\sigma^2} + \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right].$$

3. (4 pts) Based on the results of the previous question, we can define a so-called logistic regression classifier defined as

$$d_{\mathrm{LR}}(x) = \omega_1 \Leftrightarrow g_{\boldsymbol{a}}(x) = \frac{1}{1+\exp(-a_1 x - a_2)} \leq \frac{1}{2}.$$

where $\boldsymbol{a} = (a_1, a_2)^T$. In a practical application, the parameter vector $\boldsymbol{a}$ can be determined using training data from the two classes ω_1 and ω_2 denoted as $\chi = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $y_i = 0$ if x_i belongs to class ω_1 and $y_i = 1$ else.

- A first idea is to determine the vector $\boldsymbol{a}$ that minimizes the cost function

$$C_1(\chi, \boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left[g_{\boldsymbol{a}}(x_i) - y_i\right]^2.$$

  Why do you think that this cost function is not appropriate for estimating the vector $\boldsymbol{a}$?

- Another idea is to minimize the cost function

$$C_2(\chi, \boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \{-y_i \ln[g_{\boldsymbol{a}}(x_i)] - (1-y_i) \ln[1 - g_{\boldsymbol{a}}(x_i)]\}$$

  with respect to $\boldsymbol{a}$. By considering samples from the class ω_1 (such that $y_i = 0$), analyze the value of the ith term of the cost function when $g_{\boldsymbol{a}}(x_i)$ is close to 1 or close to 0 and explain why this cost function is appropriate. Calculate the gradient of this cost function and show that the steepest descent rule can be expressed as

$$a_1^{n+1} = a_1^n - \frac{\mu}{n} \sum_{i=1}^{n} [g_{\boldsymbol{a}}(x_i) - y_i] x_i, \text{ and } a_2^{n+1} = a_2^n - \frac{\mu}{n} \sum_{i=1}^{n} [g_{\boldsymbol{a}}(x_i) - y_i].$$

Response: We can guess that the first cost function $C_1(\chi, \boldsymbol{a})$ is non-convex and thus not appropriate for its minimization. Let's analyze the second cost function $C_2(\chi, \boldsymbol{a})$. When $y_i = 1$, the ith term of this cost function reduces to $-\ln[g_{\boldsymbol{a}}(x_i)]$, which equals 0 when $g_{\boldsymbol{a}}(x_i)$ is close to 1 and tends to $+\infty$ when $g_{\boldsymbol{a}}(x_i)$ tends to 0. When $y_i = 0$, the ith term of the cost function reduces to $-\ln[1 - g_{\boldsymbol{a}}(x_i)]$, which equals 0 when $g_{\boldsymbol{a}}(x_i) = 0$ and tends to $+\infty$ when $g_{\boldsymbol{a}}(x_i)$ tends to 1. As a consequence, minimizing the cost function $C_2(\chi, \boldsymbol{a})$ will provide a classifier trying to minimize the classification errors, which is precisely what we want.

The gradient of the cost function $C_2(x, \boldsymbol{a})$ is defined as

$$\frac{\partial C_2(\chi, \boldsymbol{a})}{\partial \boldsymbol{a}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ -y_i \frac{1}{g_{\boldsymbol{a}}(x_i)} \frac{\partial g_{\boldsymbol{a}}(x_i)}{\partial \boldsymbol{a}} + (1-y_i) \frac{1}{1 - g_{\boldsymbol{a}}(x_i)} \frac{\partial g_{\boldsymbol{a}}(x_i)}{\partial \boldsymbol{a}} \right\}$$

Straightforward computations lead to

$$a_1^{n+1} = a_1^n - \frac{\mu}{n} \sum_{i=1}^{n} [g_{\boldsymbol{a}}(x_i) - y_i] x_i, \text{ and } a_2^{n+1} = a_2^n - \frac{\mu}{n} \sum_{i=1}^{n} [g_{\boldsymbol{a}}(x_i) - y_i].$$

Questions related to the working paper

Remark: please make sure to justify all your responses very carefully.

1. (1 pt) Explain why higher-order statistics (HOS) are resistant to additive colored Gaussian noise
   Response: the cumulants of orders higher than 2 of a Gaussian sequence are equal to zero. Thus, if the noise $g(n)$ and the signal of interest $x(n)$ are independent, the cumulants of the signal plus noise (received signal $y(n) = x(n) + g(n)$) are equal to the cumulants of the signal plus the cumulants of the noise, i.e., $C_{k,y} = C_{k,x} + C_{k,g}$. When the noise $g(n)$ is Gaussian, its cumulants of order higher than 2 are zero, i.e., $C_{k,g} = 0$ for $k > 2$, which proves that the cumulants of order $k \geq 3$ of the received signal are equal to the cumulants of the noiseless signal of interest. In other words, $C_{k,y} = C_{k,x}$, for $k \geq 3$, showing a kind of non-sensivity to an additive Gaussian noise $g(n)$. This is what the authors mean by "resistant to additive Gaussian noise".

2. (1 pt) Express the 4th order cumulant C_{40} of the signal $y(n)$ as a function of $E[y^4(n)$ and $E[y^2(n)]$.
   Response: Using (4), we obtain

$$C_{40} = E[y^4(n)] - 3E^2[y^2(n)].$$

3. (1 pt). What is a BPSK constellation? Demonstrate that $C_{40} = -2$ for this constellation.
   Response: A BPSK constellation corresponds to the two equiprobable symbols $s_1 = 1$ and $s_2 = -1$. For this constellation, we have $y^2(n) = y^4(n) = 1$, hence $C_{40} = 1 - 3 = -2$.

4. (1 pt). What is a PAM(4) constellation? Demonstrate that $C_{40} = -1.36$ for this constellation.
   Response: A PAM(4) constellation corresponds to the four equiprobable symbols $s_1 = a, s_2 = -a, s_3 = 3a$ and $s_4 = -3a$. For this constellation, we have $y^2(n) = a^2$ with probability $1/2$ and $y^2(n) = 9a^2$ with probability $1/2$. Thus, $E[y^2(n)] = 5a^2$. Similarly, $y^4(n) = a^4$ with probability $1/2$ and $y^4(n) = 81a^4$ with probability $1/2$. Thus, $E[y^4(n)] = 41a^4$ hence $C_{40} = 41a^4 - 3(25a^4) = -34a^4$. It is mentioned in the paper that $C_{21} = E[y^2(n)] = 5a^2 = 1$, which leads to $a = 1/\sqrt{5}$, leading to $C_{40} = -34/25 = -1.36$.

5. (1 pt) Explain why C_{42} is unaffected by a (deterministic) phase rotation.
   Response: We have

$$C_{42} = \text{cum}[y(n), y(n), y^*(n), y^*(n)] = E[|y(n)|^4] - 2E^2[|y(n)|^2] - E[y^2(n)]E[(y^*(n))^2].$$

   When $y(n)$ is multiplied by $e^{j\phi}$, the two first terms $E[|y(n)|^4]$ and $E^2[|y(n)|^2]$ are unchanged since $|y(n)e^{j\phi}| = |y(n)|$. When $y(n)$ is multiplied by $e^{j\phi}$, the last term equals

$$E[y^2(n)e^{2j\phi}]E[(y^*(n))^2 e^{-2j\phi}] = E[y^2(n)]E[(y^*(n))^2]$$

   which does not depend on ϕ. As a consequence, C_{42} is unaffected by a deterministic phase rotation.

6. (1 pt) ? Demonstrate Eq. (15).
   Response: for equiprobable hypotheses H_0 and H_1, the Bayesian classifier accepts H_0 if

$$\frac{1}{\sigma_0} \exp\left\{-\frac{(S-\mu_0)^2}{2\sigma_0^2}\right\} > \frac{1}{\sigma_1} \exp\left\{-\frac{(S-\mu_1)^2}{2\sigma_1^2}\right\}$$

   or equivalently if

$$\ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right) + \frac{(S-\mu_1)^2}{2\sigma_1^2} - \frac{(S-\mu_0)^2}{2\sigma_0^2} > 0.$$

   This inequality can be re-written

$$\frac{(S-\mu_1)^2}{2\sigma_1^2} - \frac{(S-\mu_0)^2}{2\sigma_0^2} + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2} < \ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right) + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}$$

or

$$\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left[\frac{(S-\mu_1)^2}{2\sigma_1^2}-\frac{(S-\mu_0)^2}{2\sigma_0^2}+\frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}\right]<\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left[\ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right)+\frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}\right].$$

Using straightforward computations, we can show that this inequality can be written

$$(S-\mu)^2<a^2$$

with

$$a^2=\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left[\ln\left(\frac{\sigma_1^2}{\sigma_0^2}\right)+\frac{(\mu_1-\mu_0)^2}{\sigma_1^2-\sigma_0^2}\right] \quad \text{and} \quad \mu=\left(\frac{\mu_0}{\sigma_0^2}-\frac{\mu_1}{\sigma_1^2}\right)\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}$$

which proves (15).

7. (1 pt) Explain where the decision rule (18) comes from.
Response: Suppose that we want to use C_{40} for the classification of PSK(8), QAM(4,4), PAM(4) and BPSK constellations. We have $C_{40}=0$ for PSK(8), $C_{40}=-0.68$ for QAM(4,4), $C_{40}=-1.36$ for PAM(4) and $C_{40}=-2$ for BPSK, which leads to the following rule

$$\text{BPSK if } C_{40}<\frac{-2-1.36}{2}=-1.68 \tag{2}$$

$$\text{PAM(4) if } -1.68<C_{40}<\frac{-1.36-0.68}{2}=-1.02 \tag{3}$$

$$\text{QAM(4,4) if } -1.02<C_{40}<\frac{-0.68}{2}=-0.34 \tag{4}$$

$$\text{PSK(8) if } C_{40}>-0.34 \tag{5}$$

This rule is equivalent to (18).

8. (1 pt) In Example 3, explain why the pdf $f(g)=(1-\epsilon)f_N(g)+\epsilon f_I(g)$ corresponds to the presence of outliers in the data. What is the outlier probability for this pdf?
Response: This pdf corresponds to a percentage of $1-\epsilon$ noise samples distributed according to a zero mean Gaussian distribution with variance σ_N^2 and a percentage of ϵ noise samples distributed according to a zero mean Gaussian distribution with variance $\sigma_I^2=100\sigma_N^2$. The samples associated with the $\mathcal{N}(0,\sigma_I^2)$ distribution are the outliers. There is a probability of ϵ to have an outlier in the data.

9. (1 pt) In Example 7, explain why the presence of frequency offset generates symbol points that are smeared along arcs.
Response: The presence of frequency offset is modeled by the term $\exp(j2\pi nf_0T)$. For $n=1$, the first symbol is rotated by a factor $\exp(j2\pi f_0T)$. For $n=2$, the second symbol is rotated by a factor $\exp(j4\pi f_0T)$ etc... As consequence, the received symbols belong to arcs defined by $s_n\exp(j2\pi nf_0T)$.

10. (1pt) In Example 13, where does the statistics q_{LLR} comes from?.
Response: If we consult one of the references such as [19], we can see that q_{LLR} is an approximation of the likelihood ratio test statistics for distinguishing BPSK from MPSK(M) with $M\geq 4$.

11. (1pt) What kind of methods do the authors recommend when the observed data are drawn from an unknown symbol set?
Response: The authors mention in their conclusion that hierarchical agglomerative clustering algorithms (as those based on dendograms that have been studied in this course) could be used for these cases.