ENGR 691/692 Section 66 (Fall 06): Machine Learning **Assigned: August 30**
Homework 1: Bayesian Decision Theory (solutions) **Due: September 13**

Problem 1: (22 pts) Let the conditional densities for a two-category one-dimensional problem be given by the following Cauchy distribution:

$$p(x|\omega_i) = \frac{1}{\pi b} \cdot \frac{1}{1 + \left(\frac{x-a_i}{b}\right)^2}, \quad i = 1, 2.$$

1. (6 pts) By explicit integration, check that the distribution are indeed normalized.
2. (9 pts) Assuming $P(\omega_1) = P(\omega_2)$, show that $P(\omega_1|x) = P(\omega_2|x)$ if $x = \frac{a_1+a_2}{2}$, that is, the minimum error decision boundary is a point midway between the peaks of the two distributions, regardless of b.
3. (7 pts) Show that the minimum probability of error is given by

$$P(\text{error}) = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} \left| \frac{a_1 - a_2}{2b} \right| .$$

Answer:

1.
$$u = \int_{-\infty}^{\infty} p(x|\omega_i) dx = \frac{1}{\pi b} \int_{\infty}^{\infty} \frac{1}{1 + \left(\frac{x-a_i}{b}\right)^2} dx .$$

   We substitute $y = \frac{x-a_i}{b}$ into the above and get

$$\begin{aligned} k &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+y^2} dy \\ &= \frac{1}{\pi} \tan^{-1}(y) \,\Big|_{-\infty}^{\infty} \\ &= \frac{1}{\pi} \left(\frac{\pi}{2} + \frac{\pi}{2} \right) \\ &= 1 . \end{aligned}$$

2. By setting $p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2)$, we have

$$\frac{1}{\pi b} \cdot \frac{1}{1 + \left(\frac{x-a_1}{b}\right)^2} \cdot \frac{1}{2} = \frac{1}{\pi b} \cdot \frac{1}{1 + \left(\frac{x-a_2}{b}\right)^2} \cdot \frac{1}{2} ,$$

   or, equivalently,

$$x - a_1 = \pm (x - a_2) .$$

   For $a_1 \neq a_2$, this implies that $x = \frac{a_1+a_2}{2}$.

3. Without loss of generality, we assume $a_2 > a_1$. The probability of error is defined as

$$\begin{aligned} P(\text{error}) &= \int_{-\infty}^{\infty} P(\text{error}, x) dx \\ &= \int_{\infty}^{\infty} P(\text{error}|x) p(x) dx . \end{aligned}$$

   Note that the decision boundary is at $\frac{a_1+a_2}{2}$, hence

$$\begin{aligned} P(\text{error}|x) &= \begin{cases} P(\omega_2|x) & \text{if} \quad x \leq \frac{a_1+a_2}{2} \\ P(\omega_1|x) & \text{if} \quad x > \frac{a_1+a_2}{2} \end{cases} \\ &= \begin{cases} \frac{p(x|\omega_2)P(\omega_2)}{p(x)} & \text{if} \quad x \leq \frac{a_1+a_2}{2} \\ \frac{p(x|\omega_1)P(\omega_1)}{p(x)} & \text{if} \quad x > \frac{a_1+a_2}{2} \end{cases} . \end{aligned}$$

Therefore, the probability of error is

$$\begin{aligned} P(\text{error}) &= \int_{-\infty}^{\frac{a_1+a_2}{2}} p(x|\omega_2)P(\omega_2)dx + \int_{\frac{a_1+a_2}{2}}^{\infty} p(x|\omega_1)P(\omega_1)dx \\ &= \frac{1}{2\pi b}\int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{1+(\frac{x-a_2}{b})^2}dx + \frac{1}{2\pi b}\int_{\frac{a_1+a_2}{2}}^{\infty} \frac{1}{1+(\frac{x-a_1}{b})^2}dx \ . \end{aligned}$$

We substitute $y = \frac{x-a_2}{b}$ and $z = \frac{x-a_1}{b}$ into the above and get

$$\begin{aligned} P(\text{error}) &= \frac{1}{2\pi}\left[\int_{-\infty}^{\frac{a_1-a_2}{2b}} \frac{1}{1+y^2}dy + \int_{\frac{a_2-a_1}{2b}}^{\infty} \frac{1}{1+z^2}dz\right] \\ &= \frac{1}{2\pi}\left[\tan^{-1}(y)\Big|_{-\infty}^{\frac{a_1-a_2}{2b}} + \tan^{-1}(z)\Big|_{\frac{a_2-a_1}{2b}}^{\infty}\right] \\ &= \frac{1}{2\pi}\left(\tan^{-1}\frac{a_1-a_2}{2b} + \frac{\pi}{2} + \frac{\pi}{2} - \tan^{-1}\frac{a_2-a_1}{2b}\right) \\ &= \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{a_2-a_1}{2b} \ . \end{aligned}$$

Similarly, if $a_1 > a_2$, we have $P(\text{error}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{a_1-a_2}{2b}$. Therefore, we have shown that

$$P(\text{error}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_1-a_2}{2b}\right| \ .$$

Problem 2: (21 pts) Let $\omega_{max}(\mathbf{x})$ be the state of nature for which $P(\omega_{max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$ for all i, $i = 1, \ldots, c$.

1. (7 pts) Show that $P(\omega_{max}|\mathbf{x}) \geq \frac{1}{c}$.
2. (7 pts) Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(\text{error}) = 1 - \int P(\omega_{max}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \ .$$

3. (7 pts) Show that $P(\text{error}) \leq \frac{c-1}{c}$.

Answer:

1. Since $P(\omega_{max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$, we have

$$\sum_{i=1}^{c} P(\omega_{max}|\mathbf{x}) \geq \sum_{i=1}^{c} P(\omega_i|\mathbf{x}) = 1 \ .$$

   Hence

$$cP(\omega_{max}|\mathbf{x}) \geq 1 \ ,$$

   which implies that $P(\omega_{max}|\mathbf{x}) \geq \frac{1}{c}$.

2. By definition,

$$\begin{aligned} P(\text{error}) &= \int_\Omega P(\text{error}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int_\Omega \left[1 - P(\omega_{max}|\mathbf{x})\right]p(\mathbf{x})d\mathbf{x} \\ &= 1 - \int_\Omega P(\omega_{max}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \ . \end{aligned}$$

3. From 1 and 2, it is clear that

$$P(\text{error}) = 1 - \int_{\Omega} P(\omega_{max}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \leq 1 - \int_{\Omega} \frac{1}{c}p(\mathbf{x})d\mathbf{x} = 1 - \frac{1}{c} = \frac{c-1}{c} \; .$$

Problem 3: (22 pts) In many machine learning applications, one has the option either to assign the pattern to one of c classes, or to reject it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \qquad\qquad i, j = 1, \ldots, c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise}, \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action, rejection, and λ_s is the loss incurred for making any substitution error.

1. (10 pts) Please derive the decision rule with the minimum risk.
2. (6 pts) What happens if $\lambda_r = 0$?
3. (6 pts) What happens if $\lambda_r > \lambda_s$?

Answer:

1. For $i = 1, \ldots, c$,

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \\ &= \lambda_s \sum_{j=1, j\neq i}^{c} P(\omega_j|\mathbf{x}) \\ &= \lambda_s \left[1 - P(\omega_i|\mathbf{x})\right] \; . \end{aligned}$$

   For $i = c+1$,

$$R(\alpha_{c+1}|\mathbf{x}) = \lambda_r \; .$$

   Therefore, the minimum risk is achieved if we decide ω_i if $R(\alpha_i|\mathbf{x}) \leq R(\alpha_{c+1}|\mathbf{x})$, i.e., $P(\omega_i|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$, and reject otherwise.

2. If $\lambda_r = 0$, we always reject.
3. If $\lambda_r > \lambda_s$, we will never reject.

Problem 4: (12 pts + 10 extra points) Let the components of the vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ be binary-valued (0 or 1), and let $P(\omega_j)$ be the prior probability for the state of nature ω_j and $j = 1, \ldots, c$. We define

$$p_{ij} = P(x_i = 1|\omega_j), i = 1, \ldots, d, j = 1, \ldots, c,$$

with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j.

1. (12 pts) Show that the minimum probability of error is achieved by the following decision rule:
   Decide ω_k if $g_k(\mathbf{x}) \geq g_j(\mathbf{x})$ for all j and k, where

$$g_j(\mathbf{x}) = \sum_{i=1}^{d} x_i ln \frac{p_{ij}}{1 - p_{ij}} + \sum_{i=1}^{d} ln(1 - p_{ij}) + lnP(\omega_j) \; .$$

2. **(10 extra pts)** If the components of $\mathbf{x}$ are ternary valued (1, 0, or -1), show that a minimum probability of error decision rule can be derived that involves discriminant functions $g_j(\mathbf{x})$ that are quadratic function of the components x_i.

Answer:

1. Consider the following discriminant function

$$g_j(\mathbf{x}) = \ln\left[p(\mathbf{x}|\omega_j)P(\omega_j)\right] = \ln p(\mathbf{x}|\omega_j) + \ln P(\omega_j) \ .$$

The components of $\mathbf{x}$ are statistically independent for all $\mathbf{x}$ in ω_j, then we can write the density as a product:

$$\begin{aligned} p(\mathbf{x}|\omega_j) &= \prod_{i=1}^{d} p(x_i|\omega_j) \\ &= \prod_{i=1}^{d} p_{ij}^{x_i}(1-p_{ij})^{1-x_i} \ . \end{aligned}$$

Thus we have the discriminant function

$$\begin{aligned} g_j(\mathbf{x}) &= \sum_{i=1}^{d} \left[x_i \ln p_{ij} + (1-x_i)\ln(1-p_{ij})\right] + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} x_i \ln \frac{p_{ij}}{1-p_{ij}} + \sum_{i=1}^{d} \ln(1-p_{ij}) + \ln P(\omega_j) \ . \end{aligned}$$

2. Consider the following discriminant function

$$g_j(\mathbf{x}) = \ln\left[p(\mathbf{x}|\omega_j)P(\omega_j)\right] = \ln p(\mathbf{x}|\omega_j) + \ln P(\omega_j) \ .$$

The components of $\mathbf{x}$ are statistically independent for all $\mathbf{x}$ in ω_j, therefore,

$$p(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} p(x_i|\omega_j) \ .$$

Let

$$\begin{aligned} p_{ij} &= P(x_i = 1|\omega_j) \ , \\ q_{ij} &= P(x_i = 0|\omega_j) \ , \\ r_{ij} &= P(x_i = -1|\omega_j) \ . \end{aligned}$$

It is not hard to check that

$$p(x_i|\omega_j) = \prod_{i=1}^{d} p_{ij}^{\frac{1}{2}x_i + \frac{1}{2}x_i^2} q_{ij}^{1-x_i^2} r_{ij}^{-\frac{1}{2}x_i + \frac{1}{2}x_i^2} \ .$$

Thus the discriminant functions can be written as

$$\begin{aligned} g_j(\mathbf{x}) &= \sum_{i=1}^{d} \left[\left(\frac{1}{2}x_i + \frac{1}{2}x_i^2\right)\ln p_{ij} + (1-x_i^2)\ln q_{ij} + \left(-\frac{1}{2}x_i + \frac{1}{2}x_i^2\right)\ln r_{ij}\right] + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} x_i^2 \ln \frac{\sqrt{p_{ij} r_{ij}}}{q_{ij}} + \frac{1}{2}\sum_{i=1}^{d} x_i \ln \frac{p_{ij}}{r_{ij}} + \sum_{i=1}^{d} \ln q_{ij} + \ln P(\omega_j) \end{aligned}$$

which are quadratic functions of the components x_i.

Question 5: (23 pts) Suppose we have three categories with prior probabilities $P(\omega_1) = 0.5$, $P(\omega_2) = P(\omega_3) = 0.25$ and the class conditional probability distributions

$$\begin{aligned} p(x|\omega_1) &\sim N(0,1) \\ p(x|\omega_2) &\sim N(0.5,1) \\ p(x|\omega_3) &\sim N(1,1) \end{aligned}$$

where $N(\mu, \sigma^2)$ represents the normal distribution with density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

We sample the following sequence of four points: $x = 0.6,\ 0.1,\ 0.9,\ 1.1$.

1. (9 pts) Calculate explicitly the probability that the sequence actually came from ω_1, ω_3, ω_3, ω_2.
2. (6 pts) Repeat for the sequence ω_1, ω_2, ω_2, ω_3.
3. (8 pts) Find the sequence of states having the maximum probability.

Answer: It is straightforward to compute that

$$\begin{array}{lll} p(0.6|\omega_1) = 0.333225 & p(0.6|\omega_2) = 0.396953 & p(0.6|\omega_3) = 0.368270 \\ p(0.1|\omega_1) = 0.396953 & p(0.1|\omega_2) = 0.368270 & p(0.1|\omega_3) = 0.266085 \\ p(0.9|\omega_1) = 0.266085 & p(0.9|\omega_2) = 0.368270 & p(0.9|\omega_3) = 0.396953 \\ p(1.1|\omega_1) = 0.217852 & p(1.1|\omega_2) = 0.333225 & p(1.1|\omega_3) = 0.396953 . \end{array}$$

We denote $\mathbf{X} = (x_1, x_2, x_3, x_4)$ and $\boldsymbol{\omega} = (\omega(1), \omega(2), \omega(3), \omega(4))$. Clearly, there are 3^4 possible values of $\boldsymbol{\omega}$, such as

$$\begin{array}{ccc} (\omega_1, \omega_1, \omega_1, \omega_1) & (\omega_1, \omega_1, \omega_1, \omega_2) & (\omega_1, \omega_1, \omega_1, \omega_3) \\ (\omega_1, \omega_1, \omega_2, \omega_1) & (\omega_1, \omega_1, \omega_2, \omega_2) & (\omega_1, \omega_1, \omega_2, \omega_3) \\ (\omega_1, \omega_3, \omega_1, \omega_1) & (\omega_1, \omega_1, \omega_3, \omega_2) & (\omega_1, \omega_1, \omega_3, \omega_3) \\ \vdots & \vdots & \vdots \\ (\omega_3, \omega_3, \omega_3, \omega_1) & (\omega_3, \omega_3, \omega_3, \omega_2) & (\omega_3, \omega_3, \omega_3, \omega_3) \end{array}$$

For each possible value of $\boldsymbol{\omega}$, we calculate $P(\omega)$ and $P(\mathbf{x}|\boldsymbol{\omega})$ using the following, which assume the independences of x_i and $\omega(i)$:

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= \prod_{i=1}^{4} p(x_i|w(i)) \\ P(\boldsymbol{\omega}) &= \prod_{i=1}^{4} P(\omega(i)) . \end{aligned}$$

For example, if $\boldsymbol{\omega} = (\omega_1, \omega_3, \omega_3, \omega_2)$ and $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$, then we have

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= p((0.6, 0.1, 0.9, 1.1)|(\omega_1, \omega_3, \omega_3, \omega_2)) \\ &= p(0.6|\omega_1)p(0.1|\omega_3)p(0.9|\omega_3)p(1.1|\omega_2) \\ &= 0.333225 \times 0.266085 \times 0.396953 \times 0.333225 \\ &= 0.01173 \end{aligned}$$

and

$$\begin{aligned} P(\boldsymbol{\omega}) &= P(\omega_1)P(\omega_2)P(\omega_3)P(\omega_4) \\ &= \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \\ &= 0.0078125 . \end{aligned}$$

1. Given $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$ and $\boldsymbol{\omega} = (\omega_1, \omega_3, \omega_3, \omega_2)$, we have

$$\begin{aligned} p(\mathbf{X}) &= p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1) \\ &= \sum_{\boldsymbol{\omega}} p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1|\boldsymbol{\omega})P(\boldsymbol{\omega}) \end{aligned}$$

$$
\begin{aligned}
&= p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1|\omega_1, \omega_1, \omega_1, \omega_1)P(\omega_1, \omega_1, \omega_1, \omega_1) \\
&\quad + p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1|\omega_1, \omega_1, \omega_1, \omega_2)P(\omega_1, \omega_1, \omega_1, \omega_2) \\
&\quad \vdots \\
&\quad + p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1|\omega_3, \omega_3, \omega_3, \omega_3)P(\omega_3, \omega_3, \omega_3, \omega_3) \\
&= p(0.6|\omega_1)p(0.1|\omega_1)p(0.9|\omega_1)p(1.1|\omega_1)P(\omega_1)P(\omega_1)P(\omega_1)P(\omega_1) \\
&\quad + p(0.6|\omega_1)p(0.1|\omega_1)p(0.9|\omega_1)p(1.1|\omega_2)P(\omega_1)P(\omega_1)P(\omega_1)P(\omega_2) \\
&\quad \vdots \\
&\quad + p(0.6|\omega_3)p(0.1|\omega_3)p(0.9|\omega_3)p(1.1|\omega_3)P(\omega_3)P(\omega_3)P(\omega_3)P(\omega_3) \\
&= 0.012083\,.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P(\boldsymbol{\omega}|\mathbf{X}) &= P(\omega_1, \omega_3, \omega_3, \omega_2|0.6, 0.9, 0.1, 1.1) \\
&= \frac{p(0.6, 0.9, 0.1, 1.1|\omega_1, \omega_3, \omega_3, \omega_2)P(\omega_1, \omega_3, \omega_3, \omega_2)}{p(\mathbf{X})} \\
&= \frac{0.01173 \times 0.0078125}{0.012083} \\
&= 0.007584\,.
\end{aligned}
$$

2. Following the steps in part 1, we have

$$
\begin{aligned}
P(\omega_1, \omega_2, \omega_2, \omega_3|0.6, 0.1, 0.9, 1.1) &= \frac{p(0.6, 0.1, 0.9, 1.1|\omega_1, \omega_2, \omega_2, \omega_3)P(\omega_1, \omega_2, \omega_2, \omega_3)}{p(\mathbf{X})} \\
&= \frac{0.01794 \times 0.0078125}{0.012083} \\
&= 0.01160\,.
\end{aligned}
$$

3. The sequence $\boldsymbol{\omega} = (\omega_1, \omega_1, \omega_1, \omega_1)$ has the maximum probability to observe $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$. This maximum probability is 0.03966.