ECEN 649 Pattern Recognition – Spring 2015

Homework Assignment 1

Due on: Feb 9

Note: Before beginning, please consult homework policy on e-learning (in the homework folder).

Problems:

1. This problem demonstrates nicely subtle issues regarding partial information and prediction. A certain show host has placed a case with US$1,000,000 behind one of three identical doors. Behind each of the other two doors he placed a donkey. The host asks the contestant to pick one door but not to open it. The host then opens one of the other two doors to reveal a donkey. He then asks the contestant if he wants to stay with his door or switch to the other unopened door. Assume that the host is honest and that if the contestant initially picked the correct door, the host randomly picks one the two donkey doors to open. Which of the following strategies is rationally justifiable:

   (a) The contestant should never switch to the other door.

   (b) The contestant should always switch to the other door.

   (c) There is not enough information or the choice between (a) and (b) is indifferent.

   To get full credit, you must argue this by correctly computing the probabilities of success.

2. The random experiment consists of throwing two fair dice. Let us define the events:

   $D = \{\text{the sum of the dice equals } 6\}$

   $E = \{\text{the sum of the dice equals } 7\}$

   $F = \{\text{the first die lands } 4\}$

   $G = \{\text{the second die lands } 3\}$

   Show the following, both by arguing and by computing probabilities:

   (a) D is not independent of F and D is not independent of G.

   (b) E is independent of F and E is independent of G.

   (c) E is not independent of (F, G), in fact, E is completely determined by (F, G). (Here is an example where an event is independent of each of two other events but is not independent of the joint occurrence of these events.)

3. Suppose that a typist monkey is typing randomly, but that each time he types the "wrong character," it is discarded from the output. Assume also that the monkey types 24-7 at the rate of one character per second, and that each character can be one of 27 symbols (the alphabet without punctuation plus space). Given that *Hamlet* has about 130,000 characters, what is the average number of days that it would take the monkey to compose the famous play?

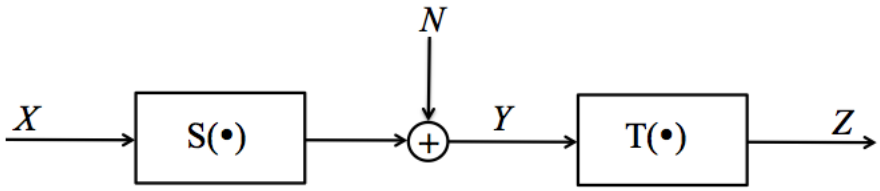
4. Suppose that 3 balls are selected without replacement from an urn containing 4 white balls, 6 red balls, and 2 black balls. Let $X_i = 1$ if the i-th ball selected is white, and let $X_i = 0$ otherwise, for $i = 1, 2, 3$. Give the joint PMF of

   (a) X_1, X_2

   (b) X_1, X_2, X_3

5. Consider 12 independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. Compute $E[X]$, $E[Y]$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $E[X+Y]$, $\mathrm{Var}(X+Y)$, $\mathrm{Cov}(X,Y)$, and $\rho(X,Y)$. (Hint: You may want to compute these in the order given.)

6. Consider the system represented by the block diagram below.



   The functionals are given by $S(X) = aX + b$, and $T(Y) = Y^2$. The additive noise is $N \sim N(0, \sigma_N^2)$. Assuming that the input signal is $X \sim N(\mu_X, \sigma_X^2)$:

   (a) Find the PDF of Y.

   (b) Find the PDF of Z.

   (c) Compute the probability that the output is bounded by a constant $k > 0$, i.e., find $P(Z \leq k)$.

7. (Bi-variate Gaussian Distribution) Suppose (X, Y) are jointly Gaussian.

   (a) Show that the joint pdf is given by:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}$$

   where $E[X] = \mu_x$, $\mathrm{Var}(X) = \sigma_x^2$, $E[Y] = \mu_y$, $\mathrm{Var}(Y) = \sigma_y^2$, and ρ is the correlation coefficient between X and Y.

   (b) Show that the conditional pdf of Y, given $X = x$, is a univariate Gaussian density with parameters:

$$\mu_{Y|X} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) \quad \text{and} \quad \sigma_{Y|X}^2 = \sigma_y^2(1-\rho^2)$$

(c) Conclude that the conditional expectation $E[Y|X]$ (which can be shown to be the "best" predictor of Y given X), is in the Gaussian case a linear function of X. This is the foundation of optimal linear filtering in Signal Processing. Plot the regression line for the case $\sigma_x = \sigma_y$, $\mu_x = 0$, fixed μ_y and a few values of ρ. What do you observe as the correlation ρ changes? What happens for the case $\rho = 0$?

8. Consider the example of a random sequence $X(n)$ of 0-1 binary r.v.'s given in class:

   - Set X(0) = 1
   - From the next 2 points, pick one randomly and set to 1, the other to zero.
   - From the next 3 points, pick one randomly and set to 1, the rest to zero.
   - From the next 4 points, pick one randomly and set to 1, the rest to zero.
   - ...

   Show that $X(n)$:

   (a) converges to 0 in probability

   (b) converges to 0 in the mean-square sense

   (c) does not converge to 0 with probability 1. In fact, show that

$$P\left(\lim_{n\to\infty} X(n) = 0\right) = 0$$

ECEN 649 Pattern Recognition – Spring 2015

Homework Assignment 1

Due on: Feb 9

Note: Before beginning, please consult homework policy on e-learning (in the homework folder).

Problems:

1. This problem demonstrates nicely subtle issues regarding partial information and prediction. A certain show host has placed a case with US$1,000,000 behind one of three identical doors. Behind each of the other two doors he placed a donkey. The host asks the contestant to pick one door but not to open it. The host then opens one of the other two doors to reveal a donkey. He then asks the contestant if he wants to stay with his door or switch to the other unopened door. Assume that the host is honest and that if the contestant initially picked the correct door, the host randomly picks one the two donkey doors to open. Which of the following strategies is rationally justifiable:

   (a) The contestant should never switch to the other door.

   (b) The contestant should always switch to the other door.

   (c) There is not enough information or the choice between (a) and (b) is indifferent.

   To get full credit, you must argue this by correctly computing the probabilities of success.

   Solution: Let us define the events:

   $$A = \{\text{The door first opened by the contestant has the prize}\}$$
   $$B = \{\text{The last unopened door has the prize}\}$$

   There are three possibilities: the contestant should never switch, always switch, or it does not matter, if $P(A) > P(B)$, $P(A) < P(B)$, or $P(A) = P(B)$, respectively. It is obvious that $P(A) = \frac{1}{3}$. As for $P(B)$, we have:

   $$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

   Clearly, $P(B|A) = 0$ and $P(B|A^c) = 1$ (the latter is so because the host is forced to open the remaining donkey door), so that

   $$P(B) = 0 + 1.P(A^c) = 1 - P(A) = 1 - \frac{1}{3} = \frac{2}{3}$$

   Therefore, $P(B) > P(A)$, and the contestant should always switch.

   The solution was easy because of the appropriate definition of the events. The solution can be much more complicated if the events are defined differently — for example, take a look at the solution in the Wikipedia entry on this problem (live link):

   http://en.wikipedia.org/wiki/Monty_Hall_problem

Note also that the answer would be different if the host was dishonest. For example, if the host would only open a door and offer a switch if the contestant picked the correct door first, then the probability of winning by switching would be zero.

Note: This problem is sometimes called the "Monty Hall Problem" (Monty Hall was a popular TV game show host). It is a "paradox" in the sense that many people intuitively expect that, since there are two options available after the host opens a door, one should have $P(A) = P(B) = \frac{1}{2}$, so it should not matter whether the contestant switches or not. On the other hand, other people feel that the host may be trying to mislead the contestant, who thus should never switch doors (this is of course not allowed in the above formulation of the problem). These wrong perceptions do not stand up to the probabilistic analysis of the problem. This problem is completely equivalent to another (and older) paradox called "The Three Prisoners Problem," proposed by Martin Gardner, in which there are three prisoners, one of which is going to be executed and the rest will be pardoned. The prison guard reveals to the first prisoner which of the other two will be freed, which effectively makes the first prisoner not want to switch his fate with the remaining prisoner. For those who are curious about the Monty Hall paradox, I recommend the Wikipedia entry mentioned earlier. In addition, you can actually play the game here: http://math.ucsd.edu/~crypto/Monty/monty.html.

2. The random experiment consists of throwing two fair dice. Let us define the events:

   $D = \{\text{the sum of the dice equals } 6\}$

   $E = \{\text{the sum of the dice equals } 7\}$

   $F = \{\text{the first die lands } 4\}$

   $G = \{\text{the second die lands } 3\}$

   Show the following, both by arguing and by computing probabilities:

   (a) D is not independent of F and D is not independent of G.

   Solution: Intuitively, D is not independent of either F or G because if we are interested in throwing a combined 6 on the sum of the dice, it is necessary *not* to throw a 6 on any individual die. So the occurrence of D depends on the outcomes of each of the dice, and thus D cannot be independent of either F or G. In probabilistic terms, we have:

   $$P(D) = P(\{1,5\} \cup \{2,4\} \cup \{3,3\} \cup \{4,2\} \cup \{5,1\}) = \frac{5}{36}$$

   so that

   $$P(D,F) = P(\{4,2\}) = \frac{1}{36} \neq \frac{5}{216} = \frac{5}{36} \times \frac{1}{6} = P(D)P(F)$$

   and similarly

   $$P(D,G) = P(\{3,3\}) = \frac{1}{36} \neq \frac{5}{216} = \frac{5}{36} \times \frac{1}{6} = P(D)P(G)$$

   (b) E is independent of F and E is independent of G.

   Solution: Here, a curious thing happens. No single outcome on any of the die can invalidate the possibility of throwing a combined 7 on the sum of the dice. So we

cannot conclude as before that E is not independent of either F or G. Let us examine the probabilities:

$$P(E) = P(\{1,6\} \cup \{2,5\} \cup \{3,4\} \cup \{4,3\} \cup \{5,2\} \cup \{6,1\}) = \frac{6}{36} = \frac{1}{6}$$

so that

$$P(E,F) = P(\{4,3\}) = \frac{1}{36} \ = \ \frac{1}{6} \times \frac{1}{6} = P(E)P(F)$$

and similarly

$$P(E,G) = P(\{4,3\}) = \frac{1}{36} \ = \ \frac{1}{6} \times \frac{1}{6} = P(E)P(G)$$

Therefore, E is indeed independent of both F and G (separately). This is not true for any combined sum other than 7.

(c) E is not independent of (F,G), in fact, E is completely determined by (F,G). (Here is an example where an event is independent of each of two other events but is not independent of the joint occurrence of these events.)

Solution: If we consider the outcomes of both dice together, then obviously the sum will depend on that. Furthermore, it will be completely determined in the sense that the conditional probability of the sum given the individual outcomes will be either one or zero. In the present case, we have:

$$P(E|F,G) = \frac{P(E,F,G)}{P(F,G)} = \frac{P(\{4,3\})}{P(\{4,3\})} = 1 \ \neq \ \frac{1}{6} = P(E)$$

3. Suppose that a typist monkey is typing randomly, but that each time he types the "wrong character," it is discarded from the output. Assume also that the monkey types 24-7 at the rate of one character per second, and that each character can be one of 27 symbols (the alphabet without punctuation plus space). Given that *Hamlet* has about 130,000 characters, what is the average number of days that it would take the monkey to compose the famous play?

   Solution: Let T_i be the amount of tries the monkey takes to get the i-th character correct. The variables T_i are independent and identically-distributed. Furthermore, each T_i is a geometric random variable with parameter $p = 1/27$, so that $E[T_i] = 27$. The average total time to complete *Hamlet* is simply $130,000 \times E[T_1] = 130,000 \times 27 = 3510000$ seconds (since each try takes one second). As each day contains 86400 seconds, this corresponds to 40.625 days, or precisely 40 days and 15 hours.

4. Suppose that 3 balls are selected without replacement from an urn containing 4 white balls, 6 red balls, and 2 black balls. Let $X_i = 1$ if the i-th ball selected is white, and let $X_i = 0$ otherwise, for $i = 1, 2, 3$. Give the joint PMF of

   (a) X_1, X_2

Solution: By conditioning on the first draw,

$$P(X_1 = 0, X_2 = 0) = P(X_2 = 0 \mid X_1 = 0)P(X_1 = 0) = \frac{7}{11} \times \frac{8}{12} = \frac{14}{33}$$
$$P(X_1 = 0, X_2 = 1) = P(X_2 = 1 \mid X_1 = 0)P(X_1 = 0) = \frac{4}{11} \times \frac{8}{12} = \frac{8}{33}$$
$$P(X_1 = 1, X_2 = 0) = P(X_2 = 0 \mid X_1 = 1)P(X_1 = 1) = \frac{8}{11} \times \frac{4}{12} = \frac{8}{33}$$
$$P(X_1 = 1, X_2 = 1) = P(X_2 = 1 \mid X_1 = 1)P(X_1 = 1) = \frac{3}{11} \times \frac{4}{12} = \frac{1}{11}$$

Another way of solving this is to imagine that the balls are all numbered from 1 to 12. Then there is clearly a total of 12×11 possible outcomes for the first two draws. Of these, 8×7 consist of two non-white balls, 8×4 consist of one white ball and one non-white ball, and 4×3 consists of two white balls.

(b) X_1, X_2, X_3

Solution: By conditioning on the first two draws, and using the result in a),

$$\begin{aligned} P(X_1 = 0, X_2 = 0, X_3 = 0) &= P(X_3 = 0 \mid X_1 = 0, X_2 = 0)P(X_1 = 0, X_2 = 0) \\ &= \frac{6}{10} \times \frac{14}{33} = \frac{14}{55} \end{aligned}$$
$$\begin{aligned} P(X_1 = 0, X_2 = 0, X_3 = 1) &= P(X_3 = 1 \mid X_1 = 0, X_2 = 0)P(X_1 = 0, X_2 = 0) \\ &= \frac{4}{10} \times \frac{14}{33} = \frac{28}{165} \end{aligned}$$
$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 0) &= P(X_3 = 0 \mid X_1 = 0, X_2 = 1)P(X_1 = 0, X_2 = 1) \\ &= \frac{7}{10} \times \frac{8}{33} = \frac{28}{165} \end{aligned}$$
$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 1) &= P(X_3 = 1 \mid X_1 = 0, X_2 = 1)P(X_1 = 0, X_2 = 1) \\ &= \frac{3}{10} \times \frac{8}{33} = \frac{12}{165} \end{aligned}$$
$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 0) &= P(X_3 = 0 \mid X_1 = 1, X_2 = 0)P(X_1 = 1, X_2 = 0) \\ &= \frac{7}{10} \times \frac{8}{33} = \frac{28}{165} \end{aligned}$$

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1) &= P(X_3 = 1 \mid X_1 = 1, X_2 = 0)P(X_1 = 1, X_2 = 0) \\ &= \frac{3}{10} \times \frac{8}{33} = \frac{12}{165} \end{aligned}$$
$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 0) &= P(X_3 = 0 \mid X_1 = 1, X_2 = 1)P(X_1 = 1, X_2 = 1) \\ &= \frac{8}{10} \times \frac{1}{11} = \frac{4}{55} \end{aligned}$$
$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 1) &= P(X_3 = 1 \mid X_1 = 1, X_2 = 1)P(X_1 = 1, X_2 = 1) \\ &= \frac{2}{10} \times \frac{1}{11} = \frac{1}{55} \end{aligned}$$

As before, this could be solved by realizing there are $12 \times 11 \times 10$ ways of performing the first three draws. Of these, $8 \times 7 \times 6$ consist of three non-white balls, etc.

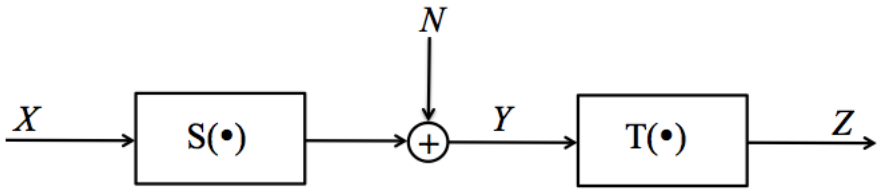
5. Consider 12 independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. Compute $E[X]$, $E[Y]$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $E[X+Y]$, $\mathrm{Var}(X+Y)$, $\mathrm{Cov}(X,Y)$, and $\rho(X,Y)$. (Hint: You may want to compute these in the order given.)

   Solution: Note that X and Y are binomial r.v.s with parameters $(n = 12, p = \frac{1}{6})$, whereas $X+Y$ is a binomial r.v. with parameters $(n = 12, p = \frac{1}{3})$. Therefore,

$$\begin{aligned}
E[X] &= E[Y] = 12 \times \frac{1}{6} = 2 \\
\mathrm{Var}(X) &= \mathrm{Var}(Y) = 12 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{3} \\
E[X+Y] &= 12 \times \frac{1}{3} = 4 \quad (= E[X] + E[Y]) \\
\mathrm{Var}(X+Y) &= 12 \times \frac{1}{3} \times \frac{2}{3} = \frac{8}{3} \quad (\neq \mathrm{Var}(X) + \mathrm{Var}(Y)!) \\
\mathrm{Cov}(X,Y) &= \frac{1}{2}(\mathrm{Var}(X+Y) - \mathrm{Var}(X) - \mathrm{Var}(Y)) = \frac{1}{2}\left(\frac{8}{3} - \frac{5}{3} - \frac{5}{3}\right) = -\frac{1}{3} \\
\rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{-\frac{1}{3}}{\frac{5}{3}} = -0.2\,.
\end{aligned}$$

   Notice that X and Y are negatively correlated (since the more 1's there are, the fewer 2's there must be, and vice-versa).

6. Consider the system represented by the block diagram below.



   The functionals are given by $S(X) = aX + b$, and $T(Y) = Y^2$. The additive noise is $N \sim N(0, \sigma_N^2)$. Assuming that the input signal is $X \sim N(\mu_X, \sigma_X^2)$:

   (a) Find the pdf of Y.

   Solution: First recall that if $X \sim N(\mu_X, \sigma_X^2)$ then $S(X) = aX + b$ is again Gaussian, with $S(X) \sim N(a\mu_X + b, a^2\sigma_X^2)$. In addition, recall that the sum of two independent Gaussian random variables is again Gaussian, the mean and variance of which are simply the sum of the means and variances, respectively, of the original variables.

Since $S(X)$ and N are independent, we have $Y = S(X)+N \sim N(a\mu_X+b, a^2\sigma_X^2+\sigma_N^2)$. The PDF of Y is therefore

$$f_Y(y) = \frac{1}{\sqrt{2\pi(a^2\sigma_X^2+\sigma_N^2)}} \exp\left\{-\frac{(y-a\mu_X-b)^2}{2(a^2\sigma_X^2+\sigma_N^2)}\right\}.$$

(b) Find the pdf of Z.

Solution: First we find the PDF of Z:

$$F_Z(z) = P(Z \le z) = P(Y^2 \le z) = P(-\sqrt{z} \le Y \le \sqrt{z}) = F_Y(\sqrt{z}) - F_Y(-\sqrt{z}).$$

Differentiation gives the pdf of Z:

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{1}{2\sqrt{z}}(f_Y(\sqrt{z}) + f_Y(-\sqrt{z})).$$

Using the result of the previous item, one obtains:

$$f_Z(z) = \frac{1}{2\sqrt{2\pi(a^2\sigma_X^2+\sigma_N^2)z}} \left(\exp\left\{-\frac{(\sqrt{z}-a\mu_X-b)^2}{2(a^2\sigma_X^2+\sigma_N^2)}\right\} + \exp\left\{-\frac{(\sqrt{z}+a\mu_X+b)^2}{2(a^2\sigma_X^2+\sigma_N^2)}\right\}\right).$$

(c) Compute the probability that the output is bounded by a constant $k > 0$, i.e., find $P(Z \le k)$.

Solution: From the previous item:

$$\begin{aligned} P(Z \le k) &= P(-\sqrt{k} \le Y \le \sqrt{k}) = P\left(\frac{-\sqrt{k}-\mu_Y}{\sigma_Y} \le \frac{Y-\mu_Y}{\sigma_Y} \le \frac{\sqrt{k}-\mu_Y}{\sigma_Y}\right) \\ &= \Phi\left(\frac{\sqrt{k}-\mu_Y}{\sigma_Y}\right) - \Phi\left(\frac{-\sqrt{k}-\mu_Y}{\sigma_Y}\right) \\ &= \Phi\left(\frac{\sqrt{k}-a\mu_X-b}{\sqrt{a^2\sigma_X^2+\sigma_N^2}}\right) - \Phi\left(\frac{-\sqrt{k}-a\mu_X-b}{\sqrt{a^2\sigma_X^2+\sigma_N^2}}\right). \end{aligned}$$

7. (Bi-variate Gaussian Distribution) Suppose (X, Y) are jointly Gaussian.

(a) Show that the joint pdf is given by:

$$\begin{aligned} f_{X,Y}(x,y) = &\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\} \end{aligned}$$

where $E[X] = \mu_X$, $\mathrm{Var}(X) = \sigma_X^2$, $E[Y] = \mu_Y$, $\mathrm{Var}(Y) = \sigma_Y^2$, and ρ is the correlation coefficient between X and Y.

Solution: The multivariate Gaussian density is given by:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{1}$$

In the bivariate case, one has

$$d = 2, \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_x^2 & \operatorname{cov}(x,y) \\ \operatorname{cov}(x,y) & \sigma_y^2 \end{bmatrix}$$

$$\det(\Sigma) = \sigma_x^2 \sigma_y^2 - \operatorname{cov}^2(x,y) = \sigma_x^2 \sigma_y^2 (1-\rho^2)$$

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix}$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2(1-\rho^2)} \begin{bmatrix} x-\mu_x & y-\mu_y \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}$$

$$= -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} \right]$$

Substituting these into (1) yields

$$f_{XY}(x,y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} \right] \right\}$$

which is the required expression.

(b) Show that the conditional pdf of Y, given $X = x$, is a univariate Gaussian density with parameters:

$$\mu_{Y|X} = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x-\mu_x) \quad \text{and} \quad \sigma_{Y|X}^2 = \sigma_y^2 (1-\rho^2)$$

Solution: From the definition of conditional pdf:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

We know that the marginal density $f_X(x)$ is a univariate Gaussian with parameters μ_x and σ_x^2:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

whereas the joint density $f_{XY}(x,y)$ was calculated in the previous item. Substituting these into (1) yields, after some algebraic manipulation:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\,\sigma_y \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_y^2(1-\rho^2)} \left(y - \mu_y - \rho \frac{\sigma_y}{\sigma_x}(x-\mu_x) \right)^2 \right]$$

By direct inspection, we can see that this is a univariate Gaussian density with parameters:

$$\mu_{Y|X} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) \quad \text{and} \quad \sigma^2_{Y|X} = \sigma^2_y(1 - \rho^2)$$

as required.

(c) Conclude that the conditional expectation $E[Y|X]$ (which can be shown to be the "best" predictor of Y given X), is in the Gaussian case a linear function of X. This is the foundation of optimal linear filtering in Signal Processing. Plot the regression line for the case $\sigma_x = \sigma_y$, $\mu_x = 0$, fixed μ_y and a few values of ρ. What do you observe as the correlation ρ changes? What happens for the case $\rho = 0$?

Solution: From the previous item, we conclude that the

$$E[Y|X = x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \rho\frac{\sigma_y}{\sigma_x}\,x \;+\; \left(\mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x\right)$$

so that $E[Y|X = x] = ax + b$ is a linear function of x. In the case $\sigma_x = \sigma_y$, $\mu_x = 0$, this reduces to a line with slope $a = \rho$ and intercept $b = \mu_y$, and the prediction will deviate from the mean μ_y by an amount proportional to the value $X = x$, with sensitivity given by the correlation coefficient ρ. If additionally $\rho = 0$, then the regression line is horizontal. In this case, the variable Y is uncorrelated from X (and thus independent, since they are jointly Gaussian) and there is no change in prediction as the value $X = x$ varies; the best predictor reduces to the no-information constant estimator μ_y.

8. Consider the example of a random sequence $X(n)$ of 0-1 binary r.v.'s given in class:

- Set X(0) = 1
- From the next 2 points, pick one randomly and set to 1, the other to zero.
- From the next 3 points, pick one randomly and set to 1, the rest to zero.
- From the next 4 points, pick one randomly and set to 1, the rest to zero.
- ...

Show that $X(n)$:

(a) converges to 0 in probability

Solution: The sequence is composed of blocks, where the first block has length 1, the second block has length 2, and so on. Let $B(n)$ be the ordinal number of the block to which n belongs. Clearly,

$$P(X_n = 1) = \frac{1}{B(n)}$$

Note that $B(n) \to \infty$ as $n \to \infty$. Therefore, for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - 0| > \epsilon) \;=\; \lim_{n\to\infty} P(X_n = 1) \;=\; \lim_{n\to\infty} \frac{1}{B(n)} = 0$$

so that $X_n \to 0$ in probability.

(b) converges to 0 in the mean-square sense

Solution: Each X_n is a Bernoulli random variable with parameter $P(X_n = 1)$. It is easy to see then that

$$E[X_n^2] = 1^2 \times P(X_n = 1) + 0^2 \times P(X_n = 0) \; = \; \frac{1}{B(n)}$$

Therefore, we have that

$$\lim_{n\to\infty} E[|X_n - 0|^2] \; = \; \lim_{n\to\infty} E[X_n^2] \; = \; \lim_{k\to\infty} \frac{1}{B(n)} = 0$$

so that $X_n \to 0$ in the mean-square sense. Of course, we know that this implies that $X_n \to 0$ in probability, so showing (b) automatically shows (a).

(c) does not converge to 0 with probability 1. In fact, show that

$$P\left(\lim_{n\to\infty} X(n) = 0\right) = 0$$

Solution: Within each block, the probability of getting a 1 is one and so is the probability of getting a 0. By the 2nd Borel-Cantelli lemma, we conclude that

$$P([X_n = 1 \;\; i.o.]) = P([X_n = 0 \;\; i.o.]) = 1$$

It follows that

$$P\left(\lim_{n\to\infty} X(n) = 0\right) = 0$$

and thus X_n does not converge to 0 with probability one.

ECEN 649 Pattern Recognition – Spring 2015

Homework 2

Due on: Mar 6

1. Consider 1-dimensional Cauchy class-conditional densities:

$$p(x|Y=i) \;=\; \frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}\,, \quad i=0,1\,,$$

where $-\infty < a_0 < a_1 < \infty$ are location parameters (there are no means for Cauchy distributions), and $b > 0$ is a dispersion parameter. Assume that the classes are equally likely, i.e., $P(Y=0) = P(Y=1) = \frac{1}{2}$.

   (a) Determine the Bayes classifier.

   (b) Determine the Bayes error as a function of the parameters a_0, a_1, and b.

   (c) Plot the Bayes error as a function of $(a_1 - a_0)/b$ and explain what you see. In particular, what are the maximum and minimum (infimum) values of the curve and what do they correspond to?

2. Consider a variation of the pass/fail classification example discussed in class, where the variables T, B, and E are still independent and identically distributed, but now are each distributed uniformly on the interval $[0, 4]$, and the model for Y is

$$Y \;=\; \begin{cases} 1\,, & TBE \le 8 \\ 0\,, & \text{otherwise.} \end{cases}$$

Find the Bayes classifier and the Bayes error when

   (a) T, B are observable.

   (b) only T is observable.

Hint: The probability density function for the product of two independent uniform r.v.'s defined on the interval $[0, L]$ is given by:

$$f(x) = \frac{1}{L^2}\,\ln\frac{L^2}{x}\,, \quad 0 < x < L^2,$$

with $f(x) = 0$ outside the interval $[0, L^2]$. In addition, note that $\int \ln x \;=\; x\ln x - x$.

3. This problem concerns the extension to the multiple-class case of concepts derived in class for the two-class case. Let $Y \in \{0, 1, \ldots, c-1\}$, where c is the number of classes, and let

$$\eta_i(x) \;=\; P(Y = i \mid X = x)\,, \quad i = 0, 1, \ldots, c-1\,,$$

for each $x \in R^d$. We need to remember that these probabilities are not independent, but satisfy $\eta_0(x) + \eta_1(x) + \cdots + \eta_{c-1}(x) = 1$, for each $x \in R^d$, so that one of the functions is redundant. In the two-class case, this is made explicity by using a single $\eta(x)$, but using the redundant set above proves advantageous in the multiple-class case, as seen below. Hint: you should answer the following items in sequence, using the previous anwers in the solution of the following ones.

(a) Given a classifier $\psi : R^d \to \{0, 1, \ldots, c-1\}$, show that its conditional error $P(\psi(X) \neq Y \mid X = x)$ is given by

$$P(\psi(X) \neq Y \mid X = x) \;=\; 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i}\, \eta_i(x) \;=\; 1 - \eta_{\psi(x)}(x)\,.$$

(b) Assuming that X has a density (i.e., X is a continuous feature vector), show that the classification error of ψ is given by

$$\epsilon \;=\; 1 - \sum_{i=0}^{c-1} \int\limits_{\{x|\psi(x)=i\}} \eta_i(x) p(x)\, dx\,.$$

(c) Prove that the Bayes classifier is given by

$$\psi^*(x) = \arg\max_{i=0,1,\ldots,c-1} \eta_i(x)\,, \quad x \in R^d\,.$$

Hint: Start by considering the difference between conditional expected errors $P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)$.

(d) Show that the Bayes error is given by

$$\epsilon^* \;=\; 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right]\,.$$

4. This problem concerns classification with a rejection option. Assume that there are c classes and $c+1$ "actions" $\alpha_0, \alpha_1, \ldots, \alpha_c$. For $i = 0, \ldots, c-1$, action α_i is simply to classify into class i, whereas action α_c is to reject, i.e., abstain from committing to any of the classes, for lack of enough evidence. This can be modeled as a Bayes decision theory problem, where the cost λ_{ij} of taking action α_i when true state of nature is j is given by:

$$\lambda_{ij} \;=\; \begin{cases} 0\,, & i = j, \text{ for } i, j = 0, \ldots, c-1 \\ \lambda_r\,, & i = c \\ \lambda_m\,, & \text{otherwise,} \end{cases}$$

where λ_r is the cost associated with a rejection, and λ_m is the cost of misclassifying a sample. Determine the optimal decision function $\alpha^* : R^d \to \{\alpha_0, \alpha_1, \ldots, \alpha_c\}$ in terms of the posterior probabilities $\eta_i(x)$ — see the previous problem — and the cost parameters. As should be expected, the occurrence of rejections will depend on the relative cost λ_r/λ_m. Explain what happens when this ratio is zero, 0.5, and greater or equal than 1.

5. Consider the general two-class Gaussian model, where

$$p(x|Y = i) \sim N_d(\mu_i, \Sigma_i)\,, \quad i = 0, 1\,.$$

In Discriminant Analysis, it is common to say that each class defines a *population* Π_i, for $i = 0, 1$, and that a sample (e.g., patient, fish, metal) X comes from population Π_i, which is denoted by $X \in \Pi_i$, if $Y = i$.

(a) Given a *linear discriminant* $g(x) = a^t x + b$, where $a \in R^d$ and $b \in R$ are arbitrary parameters (these are not the optimal parameters), compute the classification error of the associated classifier

$$\psi(x) = \begin{cases} 1\,, & g(x) = a^t x + b \geq 0 \\ 0\,, & \text{otherwise.} \end{cases}$$

in terms of Φ (the c.d.f. of a standard normal random variable), and the parameters $a, b, \mu_0, \mu_1, \Sigma_0, \Sigma_1, c_0$ and c_1, where μ_i and Σ_i are the parameters of the Gaussian populations and $c_i = P(X \in \Pi_i)$ are the prior probabilitites, for $i = 0, 1$.

Hint: The classification error is given by

$$\begin{aligned} \epsilon &= P(\psi(X) \neq Y) \\ &= P(\psi(X) = 1 \mid Y = 0)P(Y = 0) + P(\psi(X) = 0 \mid Y = 1)P(Y = 1)\,. \end{aligned}$$

In the language of Discriminant Analysis, this becomes:

$$\begin{aligned} \epsilon &= P(g(X) \geq 0 \mid X \in \Pi_0)P(X \in \Pi_0) + P(g(X) < 0 \mid X \in \Pi_1)P(X \in \Pi_1) \\ &= c_0\epsilon^0 + c_1\epsilon^1\,, \end{aligned}$$

where $c_i = P(X \in \Pi_i)$ and ϵ^i is the error *conditional* to class i, for $i = 0, 1$. The overall error ϵ is thus a convex combination of the conditional errors ϵ^0 and ϵ^1, where the weights are given by the prior probabilities. To compute the conditional error ϵ^i, one would have, in principle, to solve the multidimensional integral of a Gaussian density over a half space; for example, for class 0,

$$\epsilon^0 = \int\limits_{\{x|g(x)\geq 0\}} p(x|Y = 0)\,dx = \int\limits_{\{x|g(x)\geq 0\}} N_d(\mu_0, \Sigma_0)\,dx\,.$$

This integral can be solved using some tricks (see Prob 2.32 in DHS), but there is a much easier, "pattern-recognition" way of computing this. Notice that

$$\epsilon^0 \;=\; P(g(X) \geq 0 \mid X \in \Pi_0) = P(a^t Z + b \geq 0)\,, \text{ where } Z \sim N_d(\mu_0, \Sigma_0)\,.$$

Use the properties of the Gaussian distribution to write this in terms of Φ.

(b) Using the result from the previous item, show that if $\Sigma_0 = \Sigma_1 = \Sigma$ and $c_0 = c_1 = \frac{1}{2}$, then the Bayes error for the problem is given by

$$\epsilon^* \;=\; \Phi\left(-\frac{\delta}{2}\right)\,,$$

where $\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$ is the Mahalanobis distance between the classes. Therefore, in this case, there is a tight relationship (in fact, one-to-one) between the Mahalanobis distance and the Bayes error. What is the maximum and minimum (infimum) Bayes errors and when do they happen?

6. This problem shows that the a-priori probabilities can have a huge impact on the optimal classifier. We showed that in the Gaussian model with equal covariance matrices, the optimal classifier is a hyperplane that passes through the midpoint between μ_0 and μ_1, provided that the classes are equally likely. State the condition on the prior probabilities $P(Y = 0)$ and $P(Y = 1)$ such that the hyperplane not ony does not pass through the midpoint between μ_0 and μ_1, but it does not pass between μ_0 and μ_1 at all.

7. We pointed out in class that $\epsilon_{\text{NN}} = 0 \Leftrightarrow \epsilon^* = 0$ and $\epsilon_{\text{NN}} = \frac{1}{2} \Leftrightarrow \epsilon^* = \frac{1}{2}$. The question is whether it is possible to find a problem where $\epsilon_{\text{NN}} = \epsilon^* = \delta$ with $0 < \delta < \frac{1}{2}$, i.e., an intermediate value not at the extremes 0 and $\frac{1}{2}$. Show that this is so, by considering a one-dimensional problem with class-conditional densities

$$p(x \mid Y = i) \;=\; \begin{cases} 1\,, & 0 \leq x \leq \frac{1}{2} \\ 1\,, & i + 1 \leq x \leq i + \frac{3}{2} \\ 0\,, & \text{otherwise,} \end{cases}$$

for $i = 0, 1$. Assuming that $P(Y = 0) = P(Y = 1) = \frac{1}{2}$, show that $\epsilon_{\text{NN}} = \epsilon^* = \frac{1}{4}$.

Hint: Plot the probability densities and posterior probabilities.

8. Consider Theorem 9.4 in DGL. Explain why the exponential bound

$$P(\epsilon_n - \epsilon^* > \tau) \leq 2e^{-n\tau^2/32}\,,$$

for every $n > n_0(\tau)$ and any $\tau > 0$, implies strong universal consistency of the cubic histogram rule.

Hint: Use the First Borel-Cantelli Lemma.

ECEN 649 Pattern Recognition – Spring 2014

Homework 2

Due on: Mar 3

1. Consider 1-dimensional Cauchy class-conditional densities:

$$p(x|Y=i) \;=\; \frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}\,, \quad i=0,1\,,$$

where $-\infty < a_0 < a_1 < \infty$ are location parameters (there are no means for Cauchy distributions), and $b > 0$ is a dispersion parameter. Assume that the classes are equally likely, i.e., $P(Y=0) = P(Y=1) = \frac{1}{2}$.

(a) Determine the Bayes classifier.

Solution: We need to solve the equation $P(Y=0|X=x^*) = P(Y=1|X=x^*)$. Since the classes are equally-likely, this is equivalent to $p(x^*|Y=0) = p(x^*|Y=1)$. By inspection of the Cauchy class-conditional densities, it is clear that this will happen if and only if

$$(x^*-a_0)^2 = (x^*-a_1)^2 \;\Leftrightarrow\; x^* = \frac{a_0+a_1}{2}$$

(b) Determine the Bayes error as a function of the parameters a_0, a_1, and b.

Solution:

$$\begin{aligned}\epsilon^* &= \int_{-\infty}^{x^*} p(x|Y=1)\,P(Y=1)\,dx + \int_{x^*}^{\infty} p(x|Y=0)\,P(Y=0)\,dx \\ &= 2\,\frac{1}{2}\int_{x^*}^{\infty} p(x|Y=0)\,P(Y=0)\,dx \;\text{ (by symmetry)} \\ &= \int_{\frac{a_0+a_1}{2}}^{\infty} \frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_0}{b}\right)^2}\,dx\end{aligned}$$

By making the substitution $u = \frac{x-a_0}{b}$, we obtain

$$\epsilon^* = \frac{1}{\pi}\int_{\frac{a_1-a_0}{2b}}^{\infty} \frac{1}{1+u^2}\,du \;=\; \frac{1}{\pi}\arctan|u|\,\Big|_{u=\frac{a_1-a_0}{2b}}^{u\to\infty} \;=\; \frac{1}{2} - \frac{1}{\pi}\arctan\left|\frac{a_1-a_0}{2b}\right|$$

(c) Plot the Bayes error as a function of $(a_1-a_0)/b$ and explain what you see. In particular, what are the maximum and minimum (infimum) values of the curve and what do they correspond to?

Solution: We have

$$\epsilon^*(w) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{w}{2}\right)$$

where, by definition, $w = \left|\frac{a_1 - a_0}{b}\right| > 0$. The plot of this function can be seen in Figure 1. We can see that the Bayes error decays monotonically with increased "standard separation" between the classes, i.e. with larger values of $\left|\frac{a_1 - a_0}{b}\right|$. For example, the Bayes error is halved (equal to 0.25) when $w = 2$, that is, $|a_1 - a_0|$ is equal to $2b$ units (b plays here a similar role to the standard deviation of Gaussian densities). From Figure 1, we can see that the maximum value of ϵ^* is 0.5, which occurs for $w = 0$, that is, $a_0 = a_2$. This corresponds to the case where the class-conditional densities are equal, so that there is maximal confusion between the classes — a Bayes error of 0.5 means that the best one can do is equivalent to flipping a coin. Conversely, as the distance $\left|\frac{a_1 - a_0}{2b}\right|$ becomes infinitely large compared to b, then the class-conditional densities are maximally separated and the Bayes error tends to its minimum (infimum, in this case) value of zero.

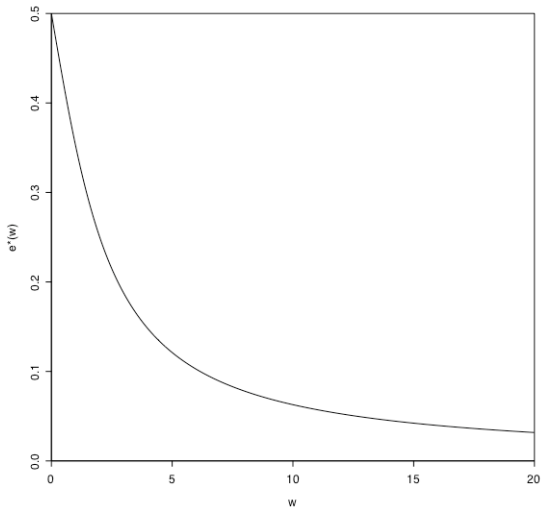


Figure 1: Bayes error as a function of standard separation between classes in the Cauchy case.

2. Consider a variation of the pass/fail classification example discussed in class, where the variables T, B, and E are still independent and identically distributed, but now are each distributed uniformly on the interval $[0, 4]$, and the model for Y is

$$Y = \begin{cases} 1\,, & TBE \leq 8 \\ 0\,, & \text{otherwise.} \end{cases}$$

Find the Bayes classifier and the Bayes error when

(a) T, B are observable.

Solution: If variables T and B are available, we use the fact that

$$P(E \leq a) = \begin{cases} \frac{a}{4}, & 0 \leq a \leq 4 \\ 1, & a > 4 \end{cases}$$

to obtain:

$$\begin{aligned} \eta(T,B) &= P(Y=1|T,B) = P(TBE \leq 8|T,B) = P\left(E \leq 8/TB|T,B\right) \\ &= \begin{cases} \frac{2}{TB}, & TB \geq 2 \\ 1, & TB < 2 \end{cases} \end{aligned}$$

Therefore, the Bayes decision is given by:

$$\psi^*(T,B) = \begin{cases} 1, & \eta(T,B) \geq \frac{1}{2} \\ 0, & \text{otw} \end{cases} = \begin{cases} 1, & TB \leq 4 \\ 0, & \text{otw} \end{cases}$$

There are two ways of computing the Bayes error. The first is via direct integration:

$$\begin{aligned} \epsilon^* &= E\left[\min\{\eta(T,B), 1-\eta(T,B)\}\right] \\ &= \int\limits_{\eta(T,B)\geq\frac{1}{2}} (1-\eta(T,B)) f(T,B)\, dBdT + \int\limits_{\eta(T,B)<\frac{1}{2}} \eta(T,B) f(T,B)\, dBdT \end{aligned}$$

In our case, $f(T,B) = \frac{1}{16} I_{\{0\leq T\leq 4, 0\leq B\leq 4\}}$. Therefore

$$\epsilon^* = \frac{1}{16}\left[\sum_{i=1}^{3} \iint\limits_{R_i} (1-\eta(T,B))\, dBdT + \iint\limits_{R_4} \eta(T,B)\, dBdT\right] \tag{1}$$

where the regions R_i (see Figure 2) are given by:

$$\begin{aligned} R_1 &= \{0 \leq T \leq 4,\ 0 \leq B \leq \min\{4, 2/T\}\} \\ R_2 &= \{1/2 \leq T \leq 1,\ 2/T \leq B \leq 4\} \\ R_3 &= \{1 \leq T \leq 4,\ 2/T \leq B \leq 4/T\} \\ R_4 &= \{1 \leq T \leq 4,\ 4/T \leq B \leq 4\} \end{aligned}$$

Notice that $1-\eta(T,B) = 0$ over R_1, $1-\eta(T,B) = 1-2/TB$ over R_2 and R_3, and $\eta(T,B) = 2/TB$ over R_4. Therefore, equation (1) becomes:

$$\epsilon^* = \frac{1}{16}\left[\sum_{i=2}^{3} \iint\limits_{R_i} 1\, dBdT - \sum_{i=2}^{3} \iint\limits_{R_i} \frac{2}{TB}\, dBdT + \iint\limits_{R_4} \frac{2}{TB}\, dBdT\right] \tag{2}$$

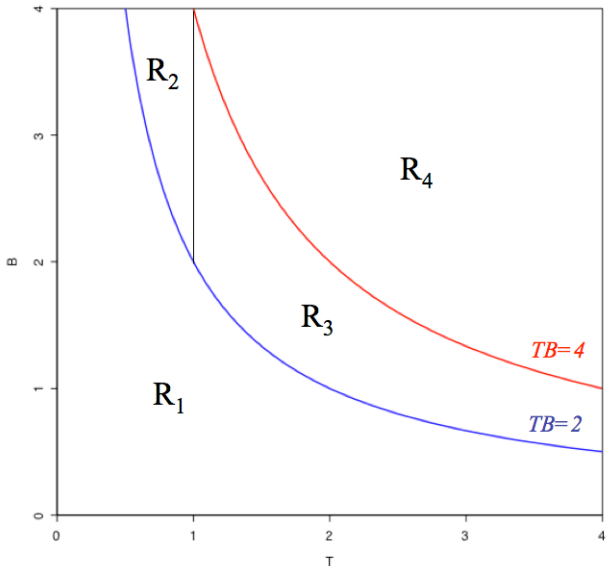


Figure 2: Integration regions in equation 1.

The special form of the regions R_2,R_3, and R_4 allows one to readily compute the double integrals in (2), to get:

$$\iint_{R_2} 1\, dBdT = 2 - 2\ln 2 \qquad \iint_{R_3} 1\, dBdT = 4\ln 2$$

$$\iint_{R_2} \frac{2}{TB}\, dBdT = (\ln 2)^2 \qquad \iint_{R_3} \frac{2}{TB}\, dBdT = \iint_{R_4} \frac{2}{TB}\, dBdT = 4(\ln 2)^2$$

The integrals above are all straightforward to compute, needing only knowledge of the anti-derivative $\int \frac{1}{u}\, du = \ln u$. Substituting these into (2) gives:

$$\epsilon^* = \frac{1}{16}\left[2 + 2\ln 2 - (\ln 2)^2\right] \cong 0.1816\,.$$

The second way to compute ϵ^* is to use the approach that was used in class:

$$\begin{aligned}
\epsilon^* &= E[P(\psi(T,B) \neq Y|T,B)] \\
&= E[P(\psi^*(T,B) = 0, Y = 1|T,B) + P(\psi^*(T,B) = 1, Y = 0|T,B)] \\
&= E[I_{\psi^*(T,B)=0}\,\eta(T,B) + I_{\psi^*(T,B)=1}(1-\eta(T,B))] \\
&= \int_4^{16} \eta(T,B) f(T,B)\, d(T,B) + \int_0^4 (1-\eta(T,B)) f(T,B)\, d(T,B) \qquad (3) \\
&= -\frac{1}{8}\int_4^{16} \frac{\ln 16u}{u}\, du - \frac{1}{16}\int_2^4 \left(1 - \frac{2}{u}\right) \ln 16u\, du \\
&= \frac{1}{16}\left[2 + 2\ln 2 - (\ln 2)^2\right] \cong 0.1816\,.
\end{aligned}$$

This method of solution is overall simpler, as it requires only one-dimensional integration. Note, however, that it requires knowledge of the slightly more complex anti-derivative $\int \frac{\ln u}{u}\, du = \frac{1}{2}(\ln u)^2$.

(b) only T is observable.

Solution: If only T is available, we use the hint to get

$$f_{E\times B}(a) = \begin{cases} \frac{1}{16}\ln\frac{16}{a}, & 0 \le a \le 16 \\ 0, & \text{otw} \end{cases}$$

and from this

$$P(EB \le a) = \int_0^a f_{E\times B}(u)\, du = \begin{cases} \frac{a}{16}\left(1 + \ln\frac{16}{a}\right), & 0 \le a \le 16 \\ 1, & a > 16 \end{cases}$$

to obtain

$$\begin{aligned}
\eta(T) &= P(Y = 1|T) = P(TBE \le 8|T) = P\left(EB \le 8/T|T\right) \\
&= \begin{cases} \frac{1}{2T}(1 + \ln 2T), & T \ge \frac{1}{2} \\ 1, & T < \frac{1}{2} \end{cases}
\end{aligned}$$

Since $\frac{1}{2T}(1 + \ln 2T) = 1/2$ and $T \ge 1/2$ imply $T \cong 2.678$, the Bayes decision is given by:

$$\psi^*(T) = \begin{cases} 1, & \eta(T) \ge \frac{1}{2} \\ 0, & \text{otw} \end{cases} \cong \begin{cases} 1, & T \le 2.678 \\ 0, & \text{otw} \end{cases}$$

The Bayes error is given by:

$$\epsilon^* = \int\limits_{\eta(T)\ge\frac{1}{2}} (1-\eta(T)) f(T)\, dT + \int\limits_{\eta(T)<\frac{1}{2}} \eta(T) f(T)\, dT$$

Since $f(T) = \frac{1}{4} I_{\{0 \leq T \leq 4\}}$, and recalling the result of part a), we can write:

$$\epsilon^* \cong \frac{1}{4} \left[\int_{1/2}^{2.678} \left(1 - \frac{1}{2T} - \frac{\ln 2T}{2T} \right) dT \; + \; \int_{2.678}^{4} \left(\frac{1}{2T} + \frac{\ln 2T}{2T} \right) dT \right] \cong 0.3031$$

The Bayes error for one variable is therefore larger than for two variables, as expected.

Hint: The probability density function for the product of two independent uniform r.v.'s defined on the interval $[0, L]$ is given by:

$$f(x) = \frac{1}{L^2} \, \ln \frac{L^2}{x} \, , \quad 0 < x < L^2,$$

with $f(x) = 0$ outside the interval $[0, L^2]$. In addition, note that $\int \ln x \; = \; x \ln x - x$.

3. This problem concerns the extension to the multiple-class case of concepts derived in class for the two-class case. Let $Y \in \{0, 1, \ldots, c-1\}$, where c is the number of classes, and let

$$\eta_i(x) \; = \; P(Y = i \mid X = x) \, , \quad i = 0, 1, \ldots, c-1 \, ,$$

for each $x \in R^d$. We need to remember that these probabilities are not independent, but satisfy $\eta_0(x) + \eta_1(x) + \cdots + \eta_{c-1}(x) = 1$, for each $x \in R^d$, so that one of the functions is redundant. In the two-class case, this is made explicity by using a single $\eta(x)$, but using the redundant set above proves advantageous in the multiple-class case, as seen below. Hint: you should answer the following items in sequence, using the previous anwers in the solution of the following ones.

(a) Given a classifier $\psi : R^d \to \{0, 1, \ldots, c-1\}$, show that its conditional error $P(\psi(X) \neq Y \mid X = x)$ is given by

$$P(\psi(X) \neq Y \mid X = x) \; = \; 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i} \, \eta_i(x) \; = \; 1 - \eta_{\psi(x)}(x) \, .$$

Solution: We have that

$$\begin{aligned} P(\psi(X) \neq Y | X = x) \; &= \; \sum_{i=0}^{c-1} P(\psi(X) = i, Y \neq i | X = x) \\ &= \; \sum_{i=0}^{c-1} I_{\psi(X)=i} \, P(Y \neq i | X = x) \; = \; \sum_{i=0}^{c-1} I_{\psi(X)=i} \, (1 - \eta_i(x)) \\ &= \; 1 \; - \; \sum_{i=0}^{c-1} I_{\psi(x)=i} \, \eta_i(x) \; = \; 1 - \eta_{\psi(x)}(x) \, . \end{aligned}$$

(b) Assuming that X has a density (i.e., X is a continuous feature vector), show that the classification error of ψ is given by

$$\epsilon \,=\, 1 - \sum_{i=0}^{c-1} \int\limits_{\{x|\psi(x)=i\}} \eta_i(x) p(x)\, dx\,.$$

Solution: Directly from the previous item,

$$\begin{aligned}\epsilon \,=\, \int P(\psi(X) \neq Y | X = x)\, p(x)\, dx \,&=\, 1 \,-\, \sum_{i=0}^{c-1} \int I_{\psi(x)=i}\, \eta_i(x)\, p(x)\, dx \\ &=\, 1 \,-\, \sum_{i=0}^{c-1} \int\limits_{\{x|\psi(x)=i\}} \eta_i(x)\, p(x)\, dx\,.\end{aligned}$$

(c) Prove that the Bayes classifier is given by

$$\psi^*(x) = \arg\max_{i=0,1,\ldots,c-1} \eta_i(x)\,, \quad x \in R^d\,.$$

Hint: Start by considering the difference between conditional expected errors $P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)$.

Solution: Again using the result of item (a),

$$\begin{aligned}&P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x) \\ &\quad =\, \eta_{\psi^*(x)}(x) - \eta_{\psi(x)}(x) \,=\, \left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right] \,-\, \eta_{\psi(x)}(x) \geq 0\,,\end{aligned}$$

by definition of $\psi^*(x)$. Integration over the feature space yields

$$\epsilon - \epsilon^* \,=\, E\left[P(\psi(X) \neq Y \mid X = x) - P(\psi^*(X) \neq Y \mid X = x)\right] \,\geq\, 0\,.$$

(d) Show that the Bayes error is given by

$$\epsilon^* \,=\, 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right]\,.$$

Solution: Using the result of item (a) and the definition of ψ^*,

$$\epsilon^* \,=\, E\left[P(\psi^*(X) \neq Y \mid X = x)\right] \,=\, 1 - E\left[\eta_{\psi^*(x)}(x)\right] \,=\, 1 - E\left[\max_{i=0,1,\ldots,c-1} \eta_i(X)\right]\,.$$

4. This problem concerns classification with a rejection option. Assume that there are c classes and $c+1$ "actions" $\alpha_0, \alpha_1, \ldots, \alpha_c$. For $i = 0, \ldots, c-1$, action α_i is simply to classify into class i, whereas action α_c is to reject, i.e., abstain from committing to any of the classes, for lack of enough evidence. This can be modeled as a Bayes decision theory problem, where the cost λ_{ij} of taking action α_i when true state of nature is j is given by:

$$\lambda_{ij} \;=\; \begin{cases} 0\,, & i = j,\ \text{for } i, j = 0, \ldots, c-1 \\ \lambda_r\,, & i = c \\ \lambda_m\,, & \text{otherwise,} \end{cases}$$

where λ_r is the cost associated with a rejection, and λ_m is the cost of misclassifying a sample. Determine the optimal decision function $\alpha^* : R^d \to \{\alpha_0, \alpha_1, \ldots, \alpha_c\}$ in terms of the posterior probabilities $\eta_i(x)$ — see the previous problem — and the cost parameters. As should be expected, the occurrence of rejections will depend on the relative cost λ_r/λ_m. Explain what happens when this ratio is zero, 0.5, and greater or equal than 1.

Solution: The optimal decision function minimizes the conditional risk

$$R(\alpha(x) = \alpha_i | X = x) \;=\; \sum_{j=0}^{c-1} \lambda_{ij} \eta_j(x)$$

at each point $x \in R^d$. For $i = 0, 1, \ldots, c-1$, this gives

$$R(\alpha(x) = \alpha_i | X = x) \;=\; \lambda_m (1 - \eta_i(x))\,,$$

while for $i = c$ (rejection), one obtains

$$R(\alpha(x) = \alpha_c | X = x) \;=\; \lambda_r\,.$$

It becomes clear then that the optimal decision is

$$\alpha^*(x) \;=\; \begin{cases} \text{classify into class } i\,, & \text{if } i = \arg\max\limits_{j=0,1,\ldots,c-1} \eta_j(x) \text{ and } 1 - \max\limits_{j=0,1,\ldots,c-1} \eta_j(X) \le \frac{\lambda_r}{\lambda_m}\,, \\ \text{reject}\,, & \text{if } 1 - \max\limits_{j=0,1,\ldots,c-1} \eta_j(X) > \frac{\lambda_r}{\lambda_m}\,. \end{cases}$$

Rejection depends on the magnitude of the ratio λ_r/λ_m in comparison to the "margin" $1 - \max \eta_j(X)$. The larger the latter is, the more confidence one has in choosing class $i = \arg\max \eta_j(x)$. If $\lambda_r/\lambda_m = 0$, then one will always reject, i.e., one is always unwilling to classify because the cost of rejection is too small (a degenerate case). If $\lambda_r/\lambda_m = 0.5$, then one will reject classification unless $\max \eta_j(X)$ is at least 0.5 (a reasonable case, if $c \ge 3$). If $\lambda_r/\lambda_m = 1$ then one will never reject, because the cost of rejection is too high (this corresponds to the classical case).

5. Consider the general two-class Gaussian model, where

$$p(x|Y=i) \sim N_d(\mu_i, \Sigma_i)\,, \quad i = 0, 1\,.$$

In Discriminant Analysis, it is common to say that each class defines a *population* Π_i, for $i = 0, 1$, and that a sample (e.g., patient, fish, metal) X comes from population Π_i, which is denoted by $X \in \Pi_i$, if $Y = i$.

(a) Given a *linear discriminant* $g(x) = a^t x + b$, where $a \in R^d$ and $b \in R$ are arbitrary parameters (these are not the optimal parameters), compute the classification error of the associated classifier

$$\psi(x) \;=\; \begin{cases} 1\,, & g(x) = a^t x + b \geq 0 \\ 0\,, & \text{otherwise.} \end{cases}$$

in terms of Φ (the c.d.f. of a standard normal random variable), and the parameters $a, b, \mu_0, \mu_1, \Sigma_0, \Sigma_1, c_0$ and c_1, where μ_i and Σ_i are the parameters of the Gaussian populations and $c_i = P(X \in \Pi_i)$ are the prior probabilitites, for $i = 0, 1$.

Hint: The classification error is given by

$$\begin{aligned} \epsilon \;&=\; P(\psi(X) \neq Y) \\ &=\; P(\psi(X) = 1 \mid Y = 0)P(Y = 0) + P(\psi(X) = 0 \mid Y = 1)P(Y = 1)\,. \end{aligned}$$

In the language of Discriminant Analysis, this becomes:

$$\begin{aligned} \epsilon \;&=\; P(g(X) \geq 0 \mid X \in \Pi_0)P(X \in \Pi_0) + P(g(X) < 0 \mid X \in \Pi_1)P(X \in \Pi_1) \\ &=\; c_0\epsilon^0 + c_1\epsilon^1\,, \end{aligned}$$

where $c_i = P(X \in \Pi_i)$ and ϵ^i is the error *conditional* to class i, for $i = 0, 1$. The overall error ϵ is thus a convex combination of the conditional errors ϵ^0 and ϵ^1, where the weights are given by the prior probabilities. To compute the conditional error ϵ^i, one would have, in principle, to solve the multidimensional integral of a Gaussian density over a half space; for example, for class 0,

$$\epsilon^0 \;=\; \int\limits_{\{x|g(x)\geq 0\}} p(x|Y=0)\,dx \;=\; \int\limits_{\{x|g(x)\geq 0\}} N_d(\mu_0, \Sigma_0)\,dx\,.$$

This integral can be solved using some tricks (see Prob 2.32 in DHS), but there is a much easier, "pattern-recognition" way of computing this. Notice that

$$\epsilon^0 \;=\; P(g(X) \geq 0 \mid X \in \Pi_0) = P(a^t Z + b \geq 0)\,, \text{ where } Z \sim N_d(\mu_0, \Sigma_0)\,.$$

Use the properties of the Gaussian distribution to write this in terms of Φ.

Solution: Using the properties of multivariate Gaussian distributions (see Lecture 2), we know that, if $Z \sim N_d(\mu, \Sigma)$, $a \in R^d$, and $b \in R$, then $a^T Z + b$ is a univariate Gaussian random variable with mean $a^T \mu + b$ and variance $a^T \Sigma a$. Therefore, following the hint,

$$\begin{aligned}\epsilon^0 &= P(g(X) \geq 0 \mid X \in \Pi_0) = P(a^t Z + b \geq 0) \\ &= 1 - F_{a^t Z+b}(0) = 1 - \Phi\left(-\frac{a^T \mu_0 + b}{\sqrt{a^T \Sigma_0 a}}\right) = \Phi\left(\frac{a^T \mu_0 + b}{\sqrt{a^T \Sigma_0 a}}\right),\end{aligned}$$

since $Z \sim N_d(\mu_0, \Sigma_0)$. Similarly, we have

$$\epsilon^1 = P(g(X) < 0 \mid X \in \Pi_1) = P(a^t Z + b < 0) = F_{a^t Z+b}(0) = \Phi\left(-\frac{a^T \mu_1 + b}{\sqrt{a^T \Sigma_1 a}}\right),$$

since $Z \sim N_d(\mu_1, \Sigma_1)$. Using the hint, one obtains

$$\epsilon = c_0 \epsilon^0 + c_1 \epsilon^1 = c_0\, \Phi\left(\frac{a^T \mu_0 + b}{\sqrt{a^T \Sigma_0 a}}\right) + c_1\, \Phi\left(-\frac{a^T \mu_1 + b}{\sqrt{a^T \Sigma_1 a}}\right). \tag{4}$$

(b) Using the result from the previous item, show that if $\Sigma_0 = \Sigma_1 = \Sigma$ and $c_0 = c_1 = \frac{1}{2}$, then the Bayes error for the problem is given by

$$\epsilon^* = \Phi\left(-\frac{\delta}{2}\right),$$

where $\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$ is the Mahalanobis distance between the classes. Therefore, in this case, there is a tight relationship (in fact, one-to-one) between the Mahalanobis distance and the Bayes error. What is the maximum and minimum (infimum) Bayes errors and when do they happen?

Solution: By plugging in the optimal values of a and b, with the assumption that $c_0 = c_1 = \frac{1}{2}$ (see Lecture 3),

$$\begin{aligned}a &= \Sigma^{-1}(\mu_1 - \mu_0) \\ b &= -\frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 + \mu_0)\end{aligned}$$

in eq. (4) and performing some algebraic simplifications, one obtains the desired result. The maximum value of the Bayes error $\epsilon^* = 0.5$ happens when $\delta = 0$, i.e., the class-conditional densities are equal to each other (maximum confusion), whereas the minimum (infimum, in this case) value $\epsilon^* = 0$ happens as $\delta \to \infty$, i.e., the classes are infinitely separated.

6. This problem shows that the a-priori probabilities can have a huge impact on the optimal classifier. We showed that in the Gaussian model with equal covariance matrices, the optimal classifier is a hyperplane that passes through the midpoint between μ_0 and μ_1, provided that the classes are equally likely. State the condition on the prior probabilities $P(Y = 0)$ and $P(Y = 1)$ such that the hyperplane not ony does not pass through the midpoint between μ_0 and μ_1, but it does not pass between μ_0 and μ_1 at all.

   Solution:

   From equation (65) in DHS, we have that

$$x_0 = \frac{1}{2}(\mu_1 + \mu_0) - t(\mu_1 - \mu_0)$$

   where

$$t = \frac{1}{\Delta^2} \ln \frac{P(Y=1)}{P(Y=0)} \tag{5}$$

   Here, $\Delta = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$ is the Mahalanobis distance between the classes. If $P(Y = 1) = P(Y = 0)$, then $t = 0$ and the decision hyperplane passes through the midpoint between μ_1 and μ_0. On the other hand, if $P(Y = 1) \neq P(Y = 0)$ (without loss of generality, let us assume $P(Y = 1) > P(Y = 0)$, so that $t > 0$), then we can see that x_0 moves along the line defined by μ_1 and μ_0, towards μ_0, according to the bias given by t. The critical point $x_0 = \mu_0$ corresponds to

$$t(\mu_1 - \mu_0) = \frac{1}{2}(\mu_1 - \mu_0) \Rightarrow t = \frac{1}{2}$$

   For $t > 2$, the decision hyperplane will not pass between the means. From (5), we can see that this is equivalent to

$$\ln \frac{P(Y=1)}{P(Y=0)} > \frac{1}{2}\Delta^2 \Rightarrow \frac{P(Y=1)}{P(Y=0)} > e^{\frac{1}{2}\Delta^2}$$

   The skewness of the situation is due to a large difference (large ratio) between the a-priori probabilities.

7. We pointed out in class that $\epsilon_{\text{NN}} = 0 \Leftrightarrow \epsilon^* = 0$ and $\epsilon_{\text{NN}} = \frac{1}{2} \Leftrightarrow \epsilon^* = \frac{1}{2}$. The question is whether it is possible to find a problem where $\epsilon_{\text{NN}} = \epsilon^* = \delta$ with $0 < \delta < \frac{1}{2}$, i.e., an intermediate value not at the extremes 0 and $\frac{1}{2}$. Show that this is so, by considering a one-dimensional problem with class-conditional densities

$$p(x \mid Y = i) = \begin{cases} 1\,, & 0 \leq x \leq \frac{1}{2} \\ 1\,, & i+1 \leq x \leq i+\frac{3}{2} \\ 0\,, & \text{otherwise,} \end{cases}$$

   for $i = 0, 1$. Assuming that $P(Y = 0) = P(Y = 1) = \frac{1}{2}$, show that $\epsilon_{\text{NN}} = \epsilon^* = \frac{1}{4}$.

Hint: Plot the probability densities and posterior probabilities.

Solution:

We need to compute $\eta(x)$. First note that

$$p(x) = p(x|Y=0)P(Y=0) + p(x|Y=1)P(Y=1) = \frac{1}{2}\left[p(x|Y=0) + p(x|Y=1)\right]$$
$$= \begin{cases} 1\,, & 0 \le x \le \frac{1}{2} \\ \frac{1}{2}\,, & 1 \le x \le \frac{3}{2} \text{ and } 2 \le x \le \frac{5}{2} \\ 0\,, & \text{otherwise.} \end{cases}$$

Therefore,

$$\eta(x) = \frac{p(x|Y=1)P(Y=1)}{p(x)} = \frac{1}{2}\frac{p(x|Y=1)}{p(x)} = \begin{cases} \frac{1}{2}, & 0 \le x \le \frac{1}{2} \\ 1, & 2 \le x \le \frac{5}{2} \\ 0, & 1 \le x \le \frac{3}{2} \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

It follows that

$$\min\{\eta(x), 1-\eta(x)\} = \begin{cases} \frac{1}{2}\,, & 0 \le x \le \frac{1}{2} \\ 0\,, & 1 \le x \le \frac{3}{2} \text{ and } 2 \le x \le \frac{5}{2} \\ \text{undefined}\,, & \text{otherwise.} \end{cases}$$

Therefore, the Bayes error is given by

$$\epsilon^* = E[\min\{\eta(x), 1-\eta(x)\}] = \int \min\{\eta(x), 1-\eta(x)\}p(x)\,dx = \int_0^{\frac{1}{2}} \frac{1}{2}\,dx = \frac{1}{4}\,,$$

whereas the asymptotic 1-NN error rate is given by

$$\epsilon_{\text{NN}} = E[2\eta(x)(1-\eta(x))] = \int 2\eta(x)(1-\eta(x))p(x)\,dx = \int_0^{\frac{1}{2}} \frac{1}{2}\,dx = \frac{1}{4}\,.$$

Therefore, $\epsilon_{NN} = \epsilon^*$, even though $\epsilon^* \neq 0$ and $\epsilon^* \neq \frac{1}{2}$. This is made possible by the ingenious way of picking the class-conditional densities $p(x|Y=0)$ and $p(x|Y=0)$ (the ability to come up with such counter-examples often comes in handy in showing facts about pattern recognition).

8. Consider Theorem 9.4 in DGL. Explain why the exponential bound

$$P(\epsilon_n - \epsilon^* > \tau) \le 2e^{-n\tau^2/32}\,,$$

for every $n > n_0(\tau)$ and any $\tau > 0$, implies strong universal consistency of the cubic histogram rule.

Hint: Use the First Borel-Cantelli Lemma.

Solution: For any given $\tau > 0$, define the events

$$A_n = \{\epsilon_n - \epsilon^* > \tau\}, \quad \text{for } n > n_0(\tau)\,.$$

We have that

$$\sum_{n>n_0(\tau)} P(A_n) \;\leq\; 2 \sum_{n>n_0(\tau)} e^{-\frac{n\tau^2}{32}} \;\leq\; 2\int_0^\infty e^{-\frac{x\tau^2}{32}}\,dx \;=\; \frac{64}{\tau^2} \;<\; \infty\,.$$

Therefore, we can use the first Borel-Cantelli lemma to conclude that

$$P\left(\limsup_{n\to\infty} A_n\right) = P\left(\limsup_{n\to\infty}\{\epsilon_n - \epsilon^* > \tau\}\right) = 0\,. \tag{6}$$

Now, note the following relations between events:

$$\left\{\lim_{n\to\infty} \epsilon_n - \epsilon^* = 0\right\} \;=\; \bigcap_{m=1}^{\infty} \liminf_{n\to\infty}\{\epsilon_n - \epsilon^* < 1/m\} \;=\; \left[\bigcup_{m=1}^{\infty} \limsup_{n\to\infty}\{\epsilon_n - \epsilon^* > 1/m\}\right]^c\,. \tag{7}$$

But, by an application of the union bound and (6)

$$P\left(\bigcup_{m=1}^{\infty} \limsup_{n\to\infty}\{\epsilon_n - \epsilon^* > 1/m\}\right) \;\leq\; \sum_{m=1}^{\infty} P\left(\limsup_{n\to\infty}\{\epsilon_n - \epsilon^* > 1/m\}\right) \;=\; 0\,.$$

Therefore, it follows from (7) that

$$P\left(\lim_{n\to\infty} \epsilon_n - \epsilon^* = 0\right) \;=\; 1 - P\left(\bigcup_{m=1}^{\infty} \limsup_{n\to\infty}\{\epsilon_n - \epsilon^* > 1/m\}\right) \;=\; 1\,,$$

that is,

$$P\left(\lim_{n\to\infty} \epsilon_n = \epsilon^*\right) = 1\,,$$

so that we have strong universal consistency.

Note: This result is a reflection of the general fact that, while $P(|X_n - X| > \tau) \to 0$ for all $\tau > 0$ is enough to get convergence of X_n to X in probability (by definition), if the convergence of the probabilities to zero is fast enough to obtain

$$\sum_n P(|X_n - X| > \tau) < \infty\,,$$

then X_n also converges to X with probability one. Almost-sure convergence becomes therefore a question of convergence rate.

ECEN 649 Pattern Recognition – Spring 2015

Homework 3

Due on: April 13

1. Consider a linear discriminant $g(x) = a^t x + b$.

   (a) Use the method of Lagrange multipliers to show that the distance of a point x_0 to the hyperplane $g(x) = 0$ is given by $|g(x_0)|/||a||$.

   (b) Use the previous result to show that the margin in a linear SVM $g(x) = a^t x + b = 0$ is given by $1/||a||$.

2. Consider the following training data consisting of 4 points:

   $$x_1 = (-1, 1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, -1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

   (a) Run manually the perceptron algorithm for these training data, considering the initial parameters to be $a(0) = (1, 0)$ and $a_0(0) = 0$, and a fixed step length $\ell = 1$. Plot the designed perceptron classifier.

   (b) By assuming the same initial parameters, find the condition on the fixed step length ℓ that allows the perceptron algorithm to find a solution after a single iteration.

   (c) By assuming the same initial parameters, and a fixed step length $\ell = 1$, show what happens with the perceptron algorithm if the training data are instead:

   $$x_1 = (-1, -1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, 1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

   Can you fix this by changing ℓ or the initial parameters?

3. Show that the polynomial kernel $K(x, y) = (1 + x^T y)^p$ satisfies Mercer's condition.

4. Consider a network with l and m neurons in two hidden layers (see Figure 30.3 in DGL). This network is specified by:

   $$\zeta(x) = c_0 + \sum_{i=1}^{l} c_i \xi_i(x)$$

   where $\xi_i(x) = \sigma(\phi_i(x))$, for $i = 1, \ldots, l$, and

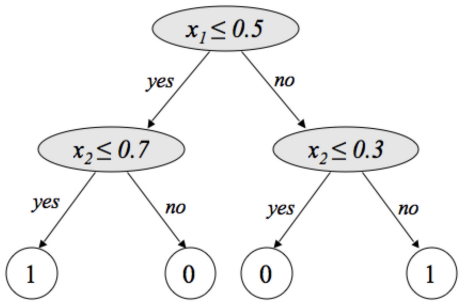
   $$\phi_i(x) = b_{i0} + \sum_{j=1}^{m} b_{ij} \upsilon_j(x)\,, \quad i = 1, \ldots, l$$

   where $\upsilon_j(x) = \sigma(\chi_j(x))$, for $j = 1, \ldots, m$, and

   $$\chi_j(x) = a_{j0} + \sum_{k=1}^{d} a_{jk} x_k\,, \quad j = 1, \ldots, m$$

   Determine the backpropagation algorithm updates for the coefficients c_i, b_{ij}, and a_{jk}. Find the backpropagation equation(s) for this problem.

5. Consider the simple CART classifier in R^2 depicted below, consisting of three splitting nodes and four leaf nodes.



Design an equivalent two-hidden-layer neural network with threshold sigmoids, with three neurons in the first hidden layer and four neurons in the second hidden layer (note the correspondence with the numbers of splitting nodes and leaf nodes).

6. This problem concerns a parallel between discrete classification and Gaussian classification. Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a discrete feature vector, such that $\mathbf{X} \in \{0,1\}^d$, i.e., all features are binary. Assume furthermore that the features are conditionally independent given $Y = 0$ and given $Y = 1$, i.e., the features are independent "inside" each class — compare this to spherical class-conditional Gaussian densities, where the features are also conditionally independent given $Y = 0$ and given $Y = 1$.

   (a) As in the Gaussian case with equal covariance matrices, prove that the Bayes classifier is linear, i.e., show that $\psi^*(x) = I_{g(\mathbf{x})>0}$, for $\mathbf{x} \in \{0,1\}^d$, where the discriminant $g(\mathbf{x})$ is given by

   $$g(\mathbf{x}) \;=\; a^T\mathbf{x} + b\,.$$

   Give the values of a and b in terms of the class-conditional distribution parameters $p_i = P(X_i = 1|Y = 0)$ and $q_i = P(X_i = 1|Y = 1)$, for $i = 1, \ldots, d$ (notice that these are different than the parameters p_i and q_i defined in the lecture), and the prior probability $c = P(Y = 1)$.

   (b) Suppose that sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ is available, where $\mathbf{X}_j = (X_{j1}, \ldots, X_{jd})$, for $j = 1, \ldots, d$. Just as is done for LDA in the Gaussian case, obtain a sample discriminant $g_n(\mathbf{x})$ from $g(\mathbf{x})$ in the previous item, by plugging in maximum-likelihood (ML) estimators $\hat{p}_i$, $\hat{q}_i$, and $\hat{c}$ for the unknown parameters p_i, q_i, and c. The maximum-likelihood estimators in this case are the empirical frequencies (you can use this fact without showing it). As in LDA, show that the designed discrete classifier $\psi_n(\mathbf{x}) = I_{g_n(\mathbf{x})>0}$, for $\mathbf{x} \in \{0,1\}^d$, is linear, by showing that

   $$g_n(\mathbf{x}) \;=\; \mathbf{a}_n^T\mathbf{x} + b_n\,.$$

   Give the values of $\mathbf{a}_n$ and b_n in terms of $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.

ECEN 649 Pattern Recognition – Spring 2015

Homework 3

Due on: April 13

1. Consider a linear discriminant $g(x) = a^t x + b$.

   (a) Use the method of Lagrange multipliers to show that the distance of a point x_0 to the hyperplane $g(x) = 0$ is given by $|g(x_0)|/||a||$.

   Solution: We need to minimize the distance $||x - x_0||$ of a point x on the hyperplane to the given point x_0. The point x is on the hyperplane if and only if $g(x) = a^T x + b = 0$. Therefore, we can find the minimum distance by solving the following optimization problem:

   $$\min ||x - x_0||^2$$
   $$\text{s.t. } a^T x + b = 0\,.$$

   We can do this by using the Lagrange multiplier method that we used in class for SVMs. First, we change this into an unconstrained optimization problem by introducing one Lagrange multiplier λ and coding the constraint into the objective function, which gives the functional

   $$L(x, \lambda) = ||x - x_0||^2 - \lambda(a^T x + b)\,. \tag{1}$$

   We need to minimize this with respect to the unconstrained variable x; therefore, we set $\partial L/\partial x = 0$, and get

   $$2(x - x_0) - \lambda a = 0 \;\Rightarrow\; x - x_0 = \frac{\lambda}{2} a \;\Rightarrow\; x = x_0 + \frac{\lambda}{2} a\,. \tag{2}$$

   At this point, we could proceed as in the case of SVMs: substitute (2) back into (1) to eliminate x and obtain a quadratic functional that depends only on λ, solve this dual problem to find λ, and substitute into (2). It turns out that in this case it is easier to employ the following equivalent approach: use the constraint $a^T x + b = 0$ in combination with (2) to solve for λ:

   $$a^T \left(x_0 + \frac{\lambda}{2} a \right) + b = 0 \;\Rightarrow\; a^T x_0 + b + \frac{\lambda}{2} ||a||^2 = 0 \;\Rightarrow\; \lambda^* = -\frac{2 g(x_0)}{||a||^2}$$

   and then substitute into (2) to get the optimal minimum distance :

   $$(x - x_0)^* = \frac{\lambda^*}{2} a = -\frac{g(x_0)}{||a||^2}\, a \;\Rightarrow\; ||x - x_0||^* = \frac{|g(x_0)|}{||a||}\,.$$

   (b) Use the previous result to show that the margin in a linear SVM $g(x) = a^t x + b = 0$ is given by $1/||a||$.

   Solution: If $g(x) = a^T x + b = 0$ is the solution of the linear SVM, then a point x_0 on the margin hyperplane satisfies

   $$g(x_0) = a^T x_0 + b = 1\,.$$

   Using the result in part (a), we obtain that the distance of x_0 to the SVM hyperplane, i.e., the margin, is given by $g(x_0)/||a|| = 1/||a||$.

2. Consider the following training data consisting of 4 points:

$$x_1 = (-1, 1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, -1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

(a) Run manually the perceptron algorithm for these training data, considering the initial parameters to be $a(0) = (1, 0)$ and $a_0(0) = 0$, and a fixed step length $\ell = 1$. Plot the designed perceptron classifier.

Solution: First we need to transform the vectors by adding a constant unit coordinate

$$x_1' = (1, -1, 1),\ x_2' = (1, 1, 1),\ x_3' = (1, -1, -1),\ x_4' = (1, 1, -1)$$

and then negating the vectors from class 0, which gives the vectors z_i for use in the perceptron algorithm:

$$z_1 = (1, -1, 1),\ z_2 = (1, 1, 1),\ z_3 = (-1, 1, 1),\ z_4 = (-1, -1, 1)$$

The transformed initial vector is $b(0) = (0, 1, 0)$. We start by determining the set $\mathcal{Y}_1$ of misclassified points, i.e., the z_i such that $b(0)^T z_i < 0$. This yields $\mathcal{Y}_1 = \{z_1, z_4\}$. The perceptron update is

$$b(1) = b(0) + \ell \sum_{z_i \in \mathcal{Y}_1} z_i = (0, 1, 0) + 1 \times [(1, -1, 1) + (-1, -1, 1)] = (0, -1, 2) \qquad (3)$$

We now observe that $b(1)^T z_i \geq 0$ for all z_i, that is, $\mathcal{Y}_2 = \emptyset$. Therefore, $b(1)$ is in the solution region and and the perceptron algorithm stops. From $b(1)$ we get the solution

$$a = (-1, 2),\ a_0 = 0$$

The corresponding perceptron classifier is depicted in Figure 1.

(b) By assuming the same initial parameters, find the condition on the fixed step length ℓ that allows the perceptron algorithm to find a solution after a single iteration.

Solution: The perceptron algorithm stopped after just one iteration. For this to happen, we need $b(1)$ to be in the solution region. From (3), we can see that for a general step length ℓ, we have

$$b(1) = (0, 1, 0) + \ell \times [(1, -1, 1) + (-1, -1, 1)] = (0, 1 - 2\ell, 2\ell)$$

We need to check that $b(1)^T z_i \geq 0$ for all z_i. For z_2 and z_3, this is true for any ℓ. For z_1 and z_4, we get a single condition:

$$b(1)^T z_1 = b(1)^T z_4 = -(1 - 2\ell) + 2\ell = 4\ell + 1 \geq 0 \ \Rightarrow\ \ell \geq \frac{1}{4}$$

Therefore, a short step length ($\ell < 0.25$) will lead to slower convergence (Note: this is not necessarily true in general; is some cases, too large a step length can lead to overshooting and slow convergence).

(c) By assuming the same initial parameters, and a fixed step length $\ell = 1$, show what happens with the perceptron algorithm if the training data are instead:

$$x_1 = (-1, -1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, 1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

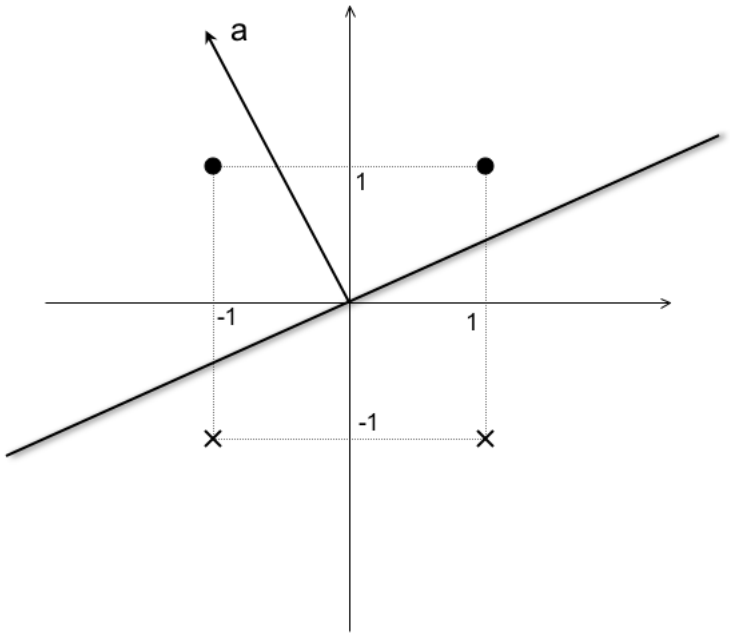


Figure 1: Perceptron Classifier.

Can you fix this by changing ℓ or the initial parameters?

Solution: In this case, the data for the perceptron algorithm becomes:

$$z_1 = (1, -1, -1),\ z_2 = (1, 1, 1),\ z_3 = (-1, 1, -1),\ z_4 = (-1, -1, 1)$$

With initial vector $b(0) = (0, 1, 0)$, we have $\mathcal{Y}_1 = \{z_1, z_4\}$ as before. The perceptron update is

$$b(1) = b(0) + \ell \sum_{z_i \in \mathcal{Y}_1} z_i = (0, 1, 0) + 1 \times [(1, -1, -1) + (-1, -1, 1)] = (0, -1, 0)$$

We observe now that $b(1)^T z_i < 0$ for z_2 and z_3, while z_1 and z_4 are correctly classified, that is, $\mathcal{Y}_2 = \{z_2, z_3\}$. The next update is:

$$b(2) = b(1) + \ell \sum_{z_i \in \mathcal{Y}_2} z_i = (0, -1, 0) + 1 \times [(1, 1, 1) + (-1, 1, -1)] = (0, 1, 0)$$

Therefore, $b(2) = b(0)$ and we are in a cycle! The perceptron algorithm will cycle indefinitely between these two vectors, alternately misclassifying z_1, z_4 and z_2, z_3.

Now, it is known (e.g. see Thm 5.1 in DHS) that the perceptron algorithm with fixed step length must converge to a solution if the problem is linearly separable. Therefore, the logical implication is that the present problem must be non-linearly separable. In fact, this problem is the XOR problem, the minimal non-linearly separable problem in two-dimensional space. Obviously, the perceptron algorithm cannot reach a solution for this data, regardless of any changes to the initial vector or step length.

3. Show that the polynomial kernel $K(x, y) = (1 + x^T y)^p$ satisfies Mercer's condition.

   Solution: Since $x^T y = y^T x$, K is symmetric: $K(x, y) = (1 + x^T y)^p = (1 + y^T x)^p = K(y, x)$. We need to show that K is positive semi-definite, i.e., for any square-integrable function g

$$\int g^2(x)\, dx < \infty$$

   we need to show that

$$\int K(x, y) g(x) g(y)\, dx\, dy \geq 0 \tag{4}$$

   Using the binomial theorem, we can expand $K(x, y) = (1 + x^T y)^p$ as

$$K(x, y) = \sum_{k=0}^{p} \binom{p}{k} (x^T y)^k$$

   Therefore, to satisfy (4), it suffices to show that

$$\int (x^T y)^k g(x) g(y)\, dx\, dy \geq 0, \quad \text{for } k = 0, \ldots, p \tag{5}$$

   Now, using the multinomial theorem we can expand $(x^T y)^k = (x_1 y_1 + \ldots + x_d y_d)^k$ as

$$(x^T y)^k = \sum_{\substack{k_1, \ldots, k_d \geq 0 \\ k_1 + \ldots + k_d \leq k}} \frac{k!}{k_1! \ldots k_d!} x_1^{k_1} y_1^{k_1} \ldots x_d^{k_d} y_d^{k_d}$$

   Therefore, in order to satisfy (5), and thus (4), it suffices to show that

$$\int x_1^{k_1} y_1^{k_1} \ldots x_d^{k_d} y_d^{k_d} g(x) g(y)\, dx\, dy \geq 0, \quad \text{for } k_i \geq 0$$

   But this follows from factorization of the integral:

$$\begin{aligned}
&\int x_1^{k_1} y_1^{k_1} \ldots x_d^{k_d} y_d^{k_d} g(x) g(y)\, dx\, dy \\
&= \int x_1^{k_1} \ldots x_d^{k_d} g(x)\, dx \int y_1^{k_1} \ldots y_d^{k_d} g(y)\, dy \\
&= \left(\int x_1^{k_1} \ldots x_d^{k_d} g(x)\, dx \right)^2 \geq 0
\end{aligned}$$

4. Consider a network with l and m neurons in two hidden layers (see Figure 30.3 in DGL). This network is specified by:

$$\zeta(x) = c_0 + \sum_{i=1}^{l} c_i \xi_i(x)$$

   where $\xi_i(x) = \sigma(\phi_i(x))$, for $i = 1, \ldots, l$, and

$$\phi_i(x) = b_{i0} + \sum_{j=1}^{m} b_{ij} \upsilon_j(x)\,, \quad i = 1, \ldots, l$$

where $\upsilon_j(x) = \sigma(\chi_j(x))$, for $j = 1, \ldots, m$, and

$$\chi_j(x) = a_{j0} + \sum_{k=1}^{d} a_{jk} x_k \,, \quad j = 1, \ldots, m$$

Determine the backpropagation algorithm updates for the coefficients c_i, b_{ij}, and a_{jk}. Find the backpropagation equation(s) for this problem.

Solution: By including bias units (neurons with constant unit output) where necessary, we can write the output of the network as:

$$\zeta(x) \;=\; \sum_{i=0}^{l} c_i \sigma \left[\sum_{j=0}^{m} b_{ij} \sigma \left(\sum_{k=0}^{d} a_{jk} x_k \right) \right]$$

with criterion function

$$J(w) = \frac{1}{2}[t - \zeta(x)]^2$$

where w is the weight vector of coefficients c_i, b_{ij}, a_{jk}, and t is the target

$$t = \begin{cases} 1, & y = 1 \\ -1, & y = 0 \end{cases}$$

For the weights c_i, the problem is essentially equal to what was done in class for the one-hidden-layer:

$$\frac{\partial J}{\partial c_i} \;=\; \frac{\partial J}{\partial \zeta} \frac{\partial \zeta}{\partial c_i} = -[t - \zeta(x)]\, \xi_i(x)$$

so that $\Delta c_i = \ell\, \delta^o \xi_i(x)$, where

$$\delta^o \;\equiv\; -\frac{\partial J}{\partial \zeta} \;=\; t - \zeta(x) \tag{6}$$

For the weights b_{ij}, we have

$$\frac{\partial J}{\partial b_{ij}} \;=\; \frac{\partial J}{\partial \phi_i} \frac{\partial \phi_i}{\partial b_{ij}} = \frac{\partial J}{\partial \phi_i}\, \upsilon_j(x)$$

where

$$\frac{\partial J}{\partial \phi_i} = \frac{\partial J}{\partial \zeta} \frac{\partial \zeta}{\partial \xi_i} \frac{\partial \xi_i}{\partial \phi_i} = -\delta^o\, c_i\, \sigma'(\phi_i(x))$$

so that $\Delta b_{ij} = \ell\, \delta_i^s \upsilon_j(x)$, where

$$\delta_i^s \;\equiv\; -\frac{\partial J}{\partial \phi_i} \;=\; \sigma'(\phi_i(x))\, c_i\, \delta^o \tag{7}$$

For the weights a_{jk}, we have

$$\frac{\partial J}{\partial a_{jk}} \;=\; \frac{\partial J}{\partial \chi_j} \frac{\partial \chi_j}{\partial a_{jk}} = \frac{\partial J}{\partial \chi_j}\, x_k$$

where

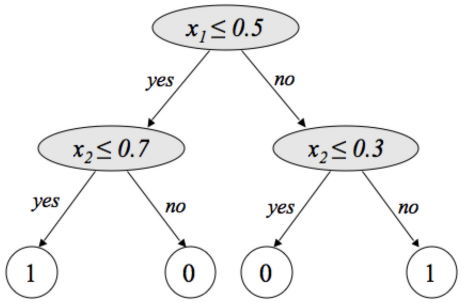
$$\frac{\partial J}{\partial \chi_j} = \sum_{i=1}^{l} \frac{\partial J}{\partial \phi_i} \frac{\partial \phi_i}{\partial \upsilon_j} \frac{\partial \upsilon_j}{\partial \chi_j} = -\sum_{i=1}^{l} \delta_i^s\, b_{ij}\, \sigma'(\chi_j(x))$$

so that $\Delta a_{jk} = \ell\, \delta_j^f x_k$, where

$$\delta_j^f \equiv -\frac{\partial J}{\partial \chi_j} = \sigma'(\chi_j(x)) \sum_{i=1}^{l} b_{ij}\, \delta_i^s \tag{8}$$

Equations (6), (7), and (8) are the backpropagation equations for this problem. Given the neuron outputs $\zeta(x)$, $\phi_i(x)$ and $\chi_j(x)$, they allow one to compute δ^o, δ_i^s, and δ_j^f, in this order (from the output backwards), and this in turns allows one to find the weight updates Δc_i, Δb_{ij}, and Δa_{jk}.

5. Consider the simple CART classifier in R^2 depicted below, consisting of three splitting nodes and four leaf nodes.



Design an equivalent two-hidden-layer neural network with threshold sigmoids, with three neurons in the first hidden layer and four neurons in the second hidden layer (note the correspondence with the numbers of splitting nodes and leaf nodes).

Solution: The CART classifier corresponds to an arrangement classifier, which can be implemented by a neural network with threshold nonlinearities in the first hidden layer. Each hyperplane split correponds to one perceptron in the first hidden layer. In addition, we can implement the desired labels for each leaf region by means of a second hidden layer, where each leaf region corresponds to one perceptron in the second hidden layer, as we will show below.

Perceptron i in the first hidden layer implements a hyperplane decision of the form

$$a_{i1}x_1 + a_{i2}x_2 + a_{i0} \geq 0$$

It is easy to check that the top split node "$x_1 \leq 0.5$" can be implemented by perceptron 1 with coefficients:

$$a_{11} = -1,\ a_{12} = 0,\ a_{10} = \boxed{0.5}$$

Similarly, the split node "$x_2 \leq 0.7$" can be implemented by perceptron 2 with coefficients:

$$a_{21} = 0,\ a_{22} = -1,\ a_{20} = \boxed{0.7}$$

and the split node "$x_2 \leq 0.3$" can be implemented by perceptron 3 with coefficients:

$$a_{31} = 0,\ a_{32} = -1,\ a_{30} = \boxed{0.3}$$

Let us number the leaf nodes L_1 to L_4 from left to right in the tree diagram. Let y_i be the output of perceptron i in the first hidden layer. It is easy to see that we have

$$\begin{aligned} x \in L_1 &\Leftrightarrow (y_1, y_2, y_3) = (1, 1, \times) \\ x \in L_2 &\Leftrightarrow (y_1, y_2, y_3) = (1, 0, \times) \\ x \in L_3 &\Leftrightarrow (y_1, y_2, y_3) = (0, \times, 1) \\ x \in L_4 &\Leftrightarrow (y_1, y_2, y_3) = (0, \times, 0) \end{aligned}$$

where "$\times$" indicates "don't care." (the corresponding output can be 0 or 1.)

We assign leaf node L_i to perceptron i in the second hidden layer. We denote the output of this perceptron by z_i. Our strategy will be to have the second hidden layer implement an indicator function for the leaf nodes, that is, we want

$$x \in L_i \Rightarrow z_i = 1 \text{ and } z_j = 0, \text{ for } j \neq i \tag{9}$$

Therefore, we want the second hidden layer to implement the following Boolean function:

y_1	y_2	y_3	z_1	z_2	z_3	z_4
1	1	0	1	0	0	0
1	1	1	1	0	0	0
1	0	0	0	1	0	0
1	0	1	0	1	0	0
0	0	1	0	0	1	0
0	1	1	0	0	1	0
0	0	0	0	0	0	1
0	1	0	0	0	0	1

The values (y_1, y_2, y_3) can be seen as the vertices of a cube in R^3. Perceptron i in the second hidden layer implements a hyperplane decision of the form

$$b_{i1}y_1 + b_{i2}y_2 + b_{i3}y_3 + b_{i0} \geq 0$$

We need the hyperplane to be oriented in such a way that the two vertices for which $z_i = 1$ in the above table be on the positive side of the hyperplane and all other vertices be on the negative side.

By visualizing the hypercube in 3-D space, it is not difficult to come up with the necessary hyperplanes. For perceptron 1, we have

$$y_1 + y_2 \geq 1.5 \Rightarrow b_{11} = 1,\ b_{12} = 1,\ b_{13} = 0,\ b_{10} = -1.5$$

Similarly, for perceptron 2,

$$y_1 - y_2 \geq 0.5 \Rightarrow b_{21} = 1,\ b_{22} = -1,\ b_{23} = 0,\ b_{20} = -0.5$$

For perceptron 3,

$$y_3 - y_1 \geq 0.5 \Rightarrow b_{31} = -1,\ b_{32} = 0,\ b_{33} = 1,\ b_{30} = -0.5$$

For perceptron 4,

$$y_3 + y_1 \leq 0.5 \;\Rightarrow\; b_{41} = -1,\; b_{42} = 0,\; b_{43} = -1,\; b_{40} = 0.5$$

The output weights c_i must be determined in such a way that

$$x \in L_i \;\Rightarrow\; \begin{cases} c_0 + \sum_{j=1}^{4} c_j z_j \geq 0, & \text{if } \psi(L_i) = 1 \\ c_0 + \sum_{j=1}^{4} c_j z_j < 0, & \text{if } \psi(L_i) = 0 \end{cases}$$

where $\psi(L_i)$ denotes the decision over L_i. From (9), it is clear that this can be accomplished by letting $c_0 = 0$ and assigning $c_i = 1$ or $c_i = -1$ according to whether $\psi(L_i) = 1$ or $\psi(L_i) = 0$, respectively. Therefore, we have

$$c_0 = 0,\; c_1 = 1,\; c_2 = -1,\; c_3 = -1,\; c_4 = 1$$

We have now specified all the weights of the desired neural network.

6. This problem concerns a parallel between discrete classification and Gaussian classification. Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a discrete feature vector, such that $\mathbf{X} \in \{0,1\}^d$, i.e., all features are binary. Assume furthermore that the features are conditionally independent given $Y = 0$ and given $Y = 1$, i.e., the features are independent "inside" each class — compare this to spherical class-conditional Gaussian densities, where the features are also conditionally independent given $Y = 0$ and given $Y = 1$.

   (a) As in the Gaussian case with equal covariance matrices, prove that the Bayes classifier is linear, i.e., show that $\psi^*(x) = I_{g(\mathbf{x})>0}$, for $\mathbf{x} \in \{0,1\}^d$, where the discriminant $g(\mathbf{x})$ is given by

   $$g(\mathbf{x}) \;=\; a^T\mathbf{x} + b\,.$$

   Give the values of a and b in terms of the class-conditional distribution parameters $p_i = P(X_i = 1|Y = 0)$ and $q_i = P(X_i = 1|Y = 1)$, for $i = 1, \ldots, d$ (notice that these are different than the parameters p_i and q_i defined in the lecture), and the prior probability $c = P(Y = 1)$.

   Solution: In the lecture on Bayes classification, we saw that the optimal discriminant can be written as:

   $$g(\mathbf{x}) \;=\; \ln \frac{P(\mathbf{X} = \mathbf{x} \mid Y = 1)}{P(\mathbf{X} = \mathbf{x} \mid Y = 0)} \;+\; \ln \frac{P(Y = 1)}{P(Y = 0)} \tag{10}$$

   In the present case, $P(Y = 1) = c = 1 - P(Y = 0)$, and

   $$\begin{aligned} \frac{P(\mathbf{X} = \mathbf{x} \mid Y = 1)}{P(\mathbf{X} = \mathbf{x} \mid Y = 0)} &= \frac{P(X_1 = x_1, \ldots, X_d = x_d \mid Y = 1)}{P(X_1 = x_1, \ldots, X_d = x_d \mid Y = 0)} \\ &= \frac{P(X_1 = x_1 \mid Y = 1) \cdots P(X_d = x_d \mid Y = 1)}{P(X_1 = x_1 \mid Y = 0) \cdots P(X_d = x_d \mid Y = 0)} \\ &= \frac{\prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}}{\prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}, \end{aligned} \tag{11}$$

where the conditional independence of the feature vector was used to obtain the second equality. Substituting in (10) results in

$$\begin{aligned} g(\mathbf{x}) &= \sum_{i=1}^{d}\left[(x_i \ln\left(\frac{p_i}{q_i}\frac{1-q_i}{1-p_i}\right)\right] + \sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i} + \ln\frac{c}{1-c} \\ &= \mathbf{a}^t\mathbf{x} + b\,, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \mathbf{a} &= \left[\ln\left(\frac{p_1}{q_1}\frac{1-q_1}{1-p_1}\right),\ldots,\ln\left(\frac{p_d}{q_d}\frac{1-q_d}{1-p_d}\right)\right]^T \\ b &= \sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i} + \ln\frac{c}{1-c}\,. \end{aligned} \tag{13}$$

(b) Suppose that sample data $S_n = \{(\mathbf{X}_1, Y_1),\ldots,(\mathbf{X}_n, Y_n)\}$ is available, where $\mathbf{X}_j = (X_{j1},\ldots,X_{jd})$, for $j = 1,\ldots,d$. Just as is done for LDA in the Gaussian case, obtain a sample discriminant $g_n(\mathbf{x})$ from $g(\mathbf{x})$ in the previous item, by plugging in maximum-likelihood (ML) estimators $\hat{p}_i$, $\hat{q}_i$, and $\hat{c}$ for the unknown parameters p_i, q_i, and c. The maximum-likelihood estimators in this case are the empirical frequencies (you can use this fact without showing it). As in LDA, show that the designed discrete classifier $\psi_n(\mathbf{x}) = I_{g_n(\mathbf{x})>0}$, for $\mathbf{x}\in\{0,1\}^d$, is linear, by showing that

$$g_n(\mathbf{x}) = \mathbf{a}_n^T\mathbf{x} + b_n\,.$$

Give the values of $\mathbf{a}_n$ and b_n in terms of $S_n = \{(\mathbf{X}_1, Y_1),\ldots,(\mathbf{X}_n, Y_n)\}$.

Solution: The MLEs of the parameters c, p_i, and q_i, for $i = 1,\ldots,d$ are given by the empirical frequencies:

$$\hat{c} = \frac{1}{n}\sum_{j=1}^{n} I_{Y_j=1}\,; \qquad \hat{p}_i = \frac{\sum_{j=1}^{n} I_{X_{ji}=1}I_{Y_j=0}}{\sum_{j=1}^{n} I_{Y_j=0}}\,; \qquad \hat{q}_i = \frac{\sum_{j=1}^{n} I_{X_{ji}=1}I_{Y_j=1}}{\sum_{j=1}^{n} I_{Y_j=1}} \tag{14}$$

for $i = 1,\ldots,d$. As is done in the case of LDA, these estimators are substituted in the expression for optimal discriminant found in the previous item, leading to the sample discriminant

$$g_n(\mathbf{x}) = \mathbf{a}_n^T\mathbf{x} + b_n\,, \tag{15}$$

where

$$\begin{aligned} \mathbf{a}_n &= \left[\ln\left(\frac{\hat{p}_1}{\hat{q}_1}\frac{1-\hat{q}_1}{1-\hat{p}_1}\right),\ldots,\ln\left(\frac{\hat{p}_d}{\hat{q}_d}\frac{1-\hat{q}_d}{1-\hat{p}_d}\right)\right]^T \\ b_n &= \sum_{i=1}^{d}\ln\frac{1-\hat{p}_i}{1-\hat{q}_i} + \ln\frac{\hat{c}}{1-\hat{c}}\,. \end{aligned} \tag{16}$$

Equations (14) and (16) give the values of $\mathbf{a}_n$ and b_n in terms of $S_n = \{(\mathbf{X}_1, Y_1),\ldots,(\mathbf{X}_n, Y_n)\}$.

ECEN 649 Pattern Recognition – Spring 2015

Homework 4

Due on: May 4

1. Let $X \in R^k$ be a feature set of size k, and let $X_0 \in R$ be an additional feature. Define the augmented feature set $X' = (X, X_0) \in R^{k+1}$. We know that the Bayes error satisfies the monotonicity property $\epsilon^*(X', Y) \leq \epsilon^*(X, Y)$. We seek conditions for the undesirable case where there is no improvement, that is, we actually have $\epsilon^*(X', Y) = \epsilon^*(X, Y)$.

   (a) Show that a sufficient condition for $\epsilon^*(X', Y) = \epsilon^*(X, Y)$ is that X_0 be independent of (X, Y).

   (b) Show that a necessary and sufficient condition for $\epsilon^*(X', Y) = \epsilon^*(X, Y)$ is that

   $$P\left[\, I_{\eta'(X')>1/2} \neq I_{\eta(X)>1/2}\,,\ \eta'(X') \neq 1/2 \,\right] = 0$$

   where $\eta'(X') = P(Y = 1|X')$ and $\eta(X) = P(Y = 1|X)$.

   Hint for part (b): The solution can be done in two steps.

   - Consider the classifier $\psi : R^{k+1} \to \{0, 1\}$ given by $\psi(x') = I_{\eta(x)>1/2}$. Given $x' = (x, x_0)$, this classifier uses only x to make a decision. Show that the error of this classifier in R^{k+1} is equal to the Bayes error $\epsilon^*(X, Y)$ in the smaller space R^k.
   - Now apply the equality from Theorem 2.2 in DGL in R^{k+1}. Note that this equality applies to *any* classifier, not just the plug-in classifier; thus given any classifier $\psi : R^{k+1} \to \{0, 1\}$, one has

   $$\begin{aligned}\epsilon[\psi] - \epsilon^*(X', Y) &= 2 \int_{R^{k+1}} |\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\, dP(x') \\ &= 2\, E\left[|\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\right]\end{aligned}$$

   where ψ^* is the Bayes classifier in R^{k+1}.

2. This problem concerns additional properties of the hold-out estimator.

   (a) Show that

   $$\mathrm{Var}(\hat{\epsilon}_{n,m}) \;=\; \frac{E[\epsilon_n](1 - E[\epsilon_n])}{m} \;+\; \frac{m-1}{m}\,\mathrm{Var}(\epsilon_n)\,. \tag{1}$$

   From this, show that $\mathrm{Var}(\hat{\epsilon}_{n,m}) \to \mathrm{Var}(\epsilon_n)$, as the number of testing samples $m \to \infty$.

   (b) Using (1), show that

   $$\mathrm{Var}(\epsilon_n) \;\leq\; \mathrm{Var}(\hat{\epsilon}_{n,m}) \;\leq\; E[\epsilon_n](1 - E[\epsilon_n])$$

   In particular, this shows that when $E[\epsilon_n]$ is small, so is $\mathrm{Var}(\hat{\epsilon}_{n,m})$.

   Hint: For any random variable X such that $0 \leq X \leq 1$ with probability 1, one has $\mathrm{Var}(X) \leq E[X](1 - E[X])$. (Why?)

(c) Show that the tail probabilities given the training data S_n satisfy:

$$P(|\hat{\epsilon}_{n,m} - \epsilon_n| \geq \tau \mid S_n) \leq e^{-2m\tau^2}, \text{ for all } \tau > 0\,.$$

Hint: Use Hoeffding's Inequality (DGL Theorem 8.1).

(d) By using the Strong Law of Large Numbers, show that, given the training data S_n, $\hat{\epsilon}_{n,m} \to \epsilon_n$ with probability 1.

(e) Repeat item (d), but this time using the result from item (c).

3. For a linear discriminant $g(x) = a^T x + b$ and Gaussian kernels $f_i^{\diamond}(x) \sim \mathcal{N}_d(0, \sigma_i^2 I)$, show that the bolstered resubstitution estimator can be computed by the formula

$$\hat{\epsilon}_n^{\diamond} = \frac{1}{n} \sum_{i=1}^{n} (\Phi_{\sigma_i}(W(x_i)) I_{y_i=0} + \Phi_{\sigma_i}(-W(x_i)) I_{y_i=1}) \tag{2}$$

where $\Phi_{\sigma_i}(u) = P(Z \leq u)$ for a Gaussian r.v. $Z \sim \mathcal{N}(0, \sigma_i^2)$, and

$$W(x) = \frac{a^t x + b}{||a||} \tag{3}$$

Hint: You should be able to use what you know from previous homeworks.

4. For the discrete histogram rule, show that the resubstitution estimator is guaranteed to be optimistically biased: $E[\hat{\epsilon}_r] - E[\epsilon_n] < 0$. In fact, show that

$$E[\hat{\epsilon}_r] \leq \epsilon^* \leq E[\epsilon_n]$$

so that the resubstitution error estimate is on average smaller even than the Bayes error.

Hint: You can show that $E[\hat{\epsilon}_r] \leq \epsilon^*$ by writing $\hat{\epsilon}_r$ as a function of the variables U_i and V_i described in the lecture on the discrete histogram rule, using Jensen's inequality, and then finding $E[U_i]$ and $E[V_i]$ from the fact that U_i and V_i are both binomial random variables (with what parameters?).

ECEN 649 Pattern Recognition – Spring 2015

Homework 4

Due on: May 4

1. Let $X \in R^k$ be a feature set of size k, and let $X_0 \in R$ be an additional feature. Define the augmented feature set $X' = (X, X_0) \in R^{k+1}$. We know that the Bayes error satisfies the monotonicity property $\epsilon^*(X', Y) \leq \epsilon^*(X, Y)$. We seek conditions for the undesirable case where there is no improvement, that is, we actually have $\epsilon^*(X', Y) = \epsilon^*(X, Y)$.

   (a) Show that a sufficient condition for $\epsilon^*(X', Y) = \epsilon^*(X, Y)$ is that X_0 be independent of (X, Y).

   Solution: Using the independence of (X, Y) from X_0, we have

   $$\eta'(X') = P(Y = 1|X') = P(Y = 1|X, X_0) = P(Y = 1|X) = \eta(X)$$

   (This can be shown easily by using the definition of conditional probability and the facts that $P(Y, X, X_0) = P(Y, X)P(X_0)$ and $P(X, X_0) = P(X)P(X_0)$).
   It follows that

   $$\epsilon^*(X', Y) = E[\min\{\eta'(X'), 1 - \eta'(X')\}] = E[\min\{\eta(X), 1 - \eta(X)\}] = \epsilon^*(X, Y)$$

   as required (The second equality can be checked by writing out the integrals).

   (b) Show that a necessary and sufficient condition for $\epsilon^*(X', Y) = \epsilon^*(X, Y)$ is that

   $$P\left[\, I_{\eta'(X')>1/2} \neq I_{\eta(X)>1/2}\,,\ \eta'(X') \neq 1/2 \,\right] = 0$$

   where $\eta'(X') = P(Y = 1|X')$ and $\eta(X) = P(Y = 1|X)$.

Hint for part (b): The solution can be done in two steps.

- Consider the classifier $\psi : R^{k+1} \to \{0, 1\}$ given by $\psi(x') = I_{\eta(x)>1/2}$. Given $x' = (x, x_0)$, this classifier uses only x to make a decision. Show that the error of this classifier in R^{k+1} is equal to the Bayes error $\epsilon^*(X, Y)$ in the smaller space R^k.
- Now apply the equality from Theorem 2.2 in DGL in R^{k+1}. Note that this equality applies to *any* classifier, not just the plug-in classifier; thus given any classifier $\psi : R^{k+1} \to \{0, 1\}$, one has

$$\begin{aligned}\epsilon[\psi] - \epsilon^*(X', Y) &= 2 \int_{R^{k+1}} |\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\, dP(x') \\ &= 2\, E\left[|\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\right]\end{aligned}$$

  where ψ^* is the Bayes classifier in R^{k+1}.

  Solution: The condition given in the assignment can be shown to be only sufficient and not necessary. It is shown below that the modified condition

$$P\left[I_{\eta'(X')>1/2} \neq I_{\eta(X)>1/2}\,,\ \eta'(x') \neq \frac{1}{2}\right] = 0 \tag{1}$$

is both necessary and sufficient for $\epsilon^*(X', Y) = \epsilon^*(X, Y)$. This condition says that the set of points where the optimal classifier using X' disagrees with the optimal classifier using X, *and* one is not on a "boundary decision" point in the larger space, has probability zero.

First consider the classifier $\psi : R^{k+1} \to \{0, 1\}$ given by $\psi(x') = I_{\eta(x)>1/2}$. This classifier uses only x to make a decision given $x' = (x, x_0)$. The error of this classifier in R^{k+1} is given by:

$$\begin{aligned}
\epsilon[\psi] &= \int_{\psi(x')=0} P(Y=1, X'=x')dx' \; + \int_{\psi(x')=1} P(Y=0, X'=x')dx' \\
&= \int_{\eta(x)\le 1/2} \int_R P(Y=1, X=x, X_0=x_0)\, dx_0\, dx \\
&\quad + \int_{\eta(x)>1/2} \int_R P(Y=0, X=x, X_0=x_0)\, dx_0\, dx \\
&= \int_{\eta(x)\le 1/2} P(Y=1, X=x)dx \; + \int_{\eta(x)>1/2} P(Y=0, X=x)dx \\
&= \epsilon^*(X, Y)
\end{aligned}$$

that is, the error of ψ in R^{k+1} is equal to the Bayes error $\epsilon^*(X, Y)$ in the smaller space R^k.

Consider now the equality from Theorem 2.2 in DGL, with $g(x') = \psi(x') = I_{\eta(x)>1/2}$ defined above, and $g^*(x') = \psi^*(x') = I_{\eta'(x')>1/2}$, the optimal classifier in R^{k+1}. Since $\epsilon[\psi] = \epsilon^*(X, Y)$ and $\epsilon[\psi^*] = \epsilon^*(X', Y)$, we have

$$\begin{aligned}
\epsilon^*(X, Y) - \epsilon^*(X', Y) &= 2 \int_{R^{k+1}} |\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\, dP(x') \\
&= 2\, E\left[|\eta'(x') - 1/2|\, I\{\psi(x') \neq \psi^*(x')\}\right]
\end{aligned} \tag{2}$$

The integrand is nonnegative, hence the expectation will be zero (and the Bayes errors will be equal) if and only if the set of points where the integrand is nonzero (positive) has probability zero; in other words, the Bayes errors will be equal if and only if condition (1) is satisfied.

2. This problem concerns additional properties of the hold-out estimator.

   (a) Show that

$$\mathrm{Var}(\hat{\epsilon}_{n,m}) \;=\; \frac{E[\epsilon_n](1 - E[\epsilon_n])}{m} \;+\; \frac{m-1}{m}\,\mathrm{Var}(\epsilon_n)\,. \tag{3}$$

   From this, show that $\mathrm{Var}(\hat{\epsilon}_{n,m}) \to \mathrm{Var}(\epsilon_n)$, as the number of testing samples $m \to \infty$.

   Solution: From the conditional variance formula, we know that

$$\mathrm{Var}(\hat{\epsilon}_{n,m}) \;=\; E[V_{\mathrm{int}}] + \mathrm{Var}(E[\hat{\epsilon}_{n,m}|S_n])$$

But $E[\hat{\epsilon}_{n,m}|S_n] = \epsilon_n$ and $V_{\text{int}} = \epsilon_n(1-\epsilon_n)/m$, so that

$$\begin{aligned}\text{Var}(\hat{\epsilon}_{n,m}) &= E\left[\frac{\epsilon_n(1-\epsilon_n)}{m}\right] + \text{Var}(\epsilon_n) = \frac{1}{m}\left(E[\epsilon_n] - E[\epsilon_n^2]\right) + \text{Var}(\epsilon_n) \\ &= \frac{1}{m}\left(E[\epsilon_n] - \text{Var}(\epsilon_n) - E[\epsilon_n]^2\right) + \text{Var}(\epsilon_n) = \frac{E[\epsilon_n](1-E[\epsilon_n])}{m} + \frac{m-1}{m}\text{Var}(\epsilon_n).\end{aligned} \tag{4}$$

As $m \to \infty$, the coefficients in this convex combination tend to 0 and 1, respectively, showing that $\text{Var}(\hat{\epsilon}_{n,m}) \to \text{Var}(\epsilon_n)$.

(b) Using (3), show that

$$\text{Var}(\epsilon_n) \leq \text{Var}(\hat{\epsilon}_{n,m}) \leq E[\epsilon_n](1-E[\epsilon_n])$$

In particular, this shows that when $E[\epsilon_n]$ is small, so is $\text{Var}(\hat{\epsilon}_{n,m})$.

Hint: For any random variable X such that $0 \leq X \leq 1$ with probability 1, one has $\text{Var}(X) \leq E[X](1-E[X])$. (Why?)

Solution: Using the hint, we obtain

$$\text{Var}(\epsilon_n) \leq E[\epsilon_n](1-E[\epsilon_n]). \tag{5}$$

Applying inequality (5) in (4) yields

$$\text{Var}(\hat{\epsilon}_{n,m}) \leq \frac{E[\epsilon_n](1-E[\epsilon_n])}{m} + \frac{m-1}{m}E[\epsilon_n](1-E[\epsilon_n]) = E[\epsilon_n](1-E[\epsilon_n]).$$

But applying inequality (5) in (4) also yields

$$\text{Var}(\hat{\epsilon}_{n,m}) \geq \frac{\text{Var}(\epsilon_n)}{m} + \frac{m-1}{m}\text{Var}(\epsilon_n) = \text{Var}(\epsilon_n).$$

(c) Show that the tail probabilities given the training data S_n satisfy:

$$P(|\hat{\epsilon}_{n,m} - \epsilon_n| \geq \tau \mid S_n) \leq e^{-2m\tau^2}, \text{ for all } \tau > 0\,.$$

Hint: Use Hoeffding's Inequality (DGL Theorem 8.1).

Solution: Let $Z_i = |Y_i - \psi_n(X_i)|$, where (X_i, Y_i) is a test sample, for $i = 1, \ldots, m$. Then the Z_i are independent, bounded random variables, such that Z_i falls into the interval $[0,1]$ with probability one. In addition, given S_n, we have that $\sum_{i=1}^m Z_i = m\hat{\epsilon}_{n,m}$, and $E[\sum_{i=1}^m Z_i \,|\, S_n] = mE[Z_1] = m\epsilon_n$. Therefore, Hoeffding's inequality gives:

$$P(|m\hat{\epsilon}_{n,m} - m\epsilon_n| > \theta \,|\, S_n) = P(|\hat{\epsilon}_{n,m} - \epsilon_n| > \theta/m \,|\, S_n) \leq e^{-2\theta^2/m}, \text{ for all } \theta > 0$$

Now let $\tau = \theta/m$, that is, $\theta = m\tau$. This leads to the required inequality:

$$P(|\hat{\epsilon}_{n,m} - \epsilon_n| > \tau \,|\, S_n) \leq e^{-2m\tau^2}, \text{ for all } \tau > 0$$

(d) By using the Strong Law of Large Numbers, show that, given the training data S_n, $\hat{\epsilon}_{n,m} \to \epsilon_n$ with probability 1.

Solution: The variables Z_i are independent and identically distributed. They are also bounded, so that all moments are finite. We can thus apply the Strong Law of Large Numbers theorem, which asserts that the sample mean converges to the true mean with probability 1, that is:

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \to E[Z_1] \text{ as } m \to \infty, \text{ with probability 1}$$

which is to say that, given S_n,

$$\hat{\epsilon}_{n,m} \to \epsilon_n \text{ as } m \to \infty, \text{ with probability 1}$$

Therefore, as the number of test samples increases to infinity, we have very strong convergence of the hold-out error estimator to the true classification error given S_n.

(e) Repeat item (d), but this time using the result from item (c).

Solution: The same conclusion can be reached by using Hoeffding's inequality (which notably does not require identically-distributed random variables). From part (a), we saw that this implies convergence of $P(|\hat{\epsilon}_{n,m} - \epsilon_n| > \tau \,|\, S_n)$ to zero for any $\tau > 0$, as $m \to \infty$ (at an exponential rate); that is, given S_n, $\hat{\epsilon}_{n,m}$ coverges to ϵ_n in probability as $m \to \infty$. But, as we showed in the solutions to HW2 (see also DGL Thm A.23), the exponential rate of convergence and the First Borel-Cantelli Lemma transform this convergence in probability to convergence with probability 1.

3. For a linear discriminant $g(x) = a^T x + b$ and Gaussian kernels $f_i^{\diamond}(x) \sim \mathcal{N}_d(0, \sigma_i^2 I)$, show that the bolstered resubstitution estimator can be computed by the formula

$$\hat{\epsilon}_n^{\diamond} = \frac{1}{n}\sum_{i=1}^{n} (\Phi_{\sigma_i}(W(x_i))I_{y_i=0} + \Phi_{\sigma_i}(-W(x_i))I_{y_i=1}) \tag{6}$$

where $\Phi_{\sigma_i}(u) = P(Z \leq u)$ for a Gaussian r.v. $Z \sim \mathcal{N}(0, \sigma_i^2)$, and

$$W(x) = \frac{a^t x + b}{||a||} \tag{7}$$

Hint: You should be able to use what you know from previous homeworks.

Solution: From the class slides, we know that

$$\hat{\epsilon}_n^{\diamond} = \frac{1}{n}\sum_{i=1}^{n} \left(\int_{A_1} f_i^{\diamond}(x - x_i)\, dx\, I_{y_i=0} + \int_{A_0} f_i^{\diamond}(x - x_i)\, dx\, I_{y_i=1} \right) \tag{8}$$

where $A_j = \{x \in R^d \mid \psi_n(x) = j\}$, for $j = 0, 1$. Suppose that $x_i \in A_0$. By exploiting the symmetry of the problem, we may assume, without loss of generality, the geometry depicted

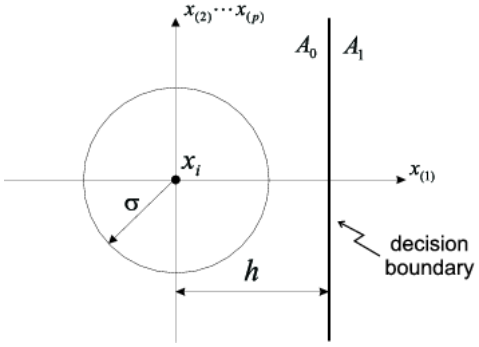


Figure 1: Diagram for calculation of $\int_{A_0} f_i^{\diamond}(x - x_i)\, dx$.

in Fig. 3. We have that

$$\begin{aligned}
\int_{A_0} f_i^{\diamond}(x - x_i)\, dx \;&=\; \int_{A_0} f_i^{\diamond}(x)\, dx \\
&= \int_{A_0} \frac{1}{(2\pi)^{p/2}\sigma_i^p}\, \exp\!\left(-\frac{||x||^2}{2\sigma_i^2}\right) dx \\
&= \int_{-\infty}^{h} \frac{1}{(2\pi)^{1/2}\sigma_i}\, \exp\left(-\frac{x_{(1)}^2}{2\sigma_i^2}\right) dx_{(1)} \\
&\times \underbrace{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma_i}\, \exp\left(-\frac{x_{(2)}^2}{2\sigma_i^2}\right) dx_{(2)} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma_i}\, \exp\left(-\frac{x_{(p)}^2}{2\sigma_i^2}\right) dx_{(p)}}_{1} \\
&= \Phi_{\sigma_i}(h)
\end{aligned} \tag{9}$$

Clearly, $\int_{A_1} f_i^{\diamond}(x - x_i)\, dx = 1 - \Phi_{\sigma_i}(h) = \Phi_{\sigma_i}(-h)$. If $x \in A_1$ instead, then the signs of h are interchanged. Now, the normalized W statistic at point x_i is given by $W_a(x_i) = (-1)^{(j+1)}\, h$ for $x_i \in A_j$, $j = 0, 1$. It follows that $\int_{A_j} f_i^{\diamond}(x - x_i)\, dx = \Phi_{\sigma_i}((-1)^{(j+1)} W_a(x_i))$, for $x_i \in R^p$, $j = 0, 1$. By replacing this into (8), one obtains (6).

4. For the discrete histogram rule, show that the resubstitution estimator is guaranteed to be optimistically biased: $E[\hat{\epsilon}_r] - E[\epsilon_n] < 0$. In fact, show that

$$E[\hat{\epsilon}_r] \;\leq\; \epsilon^* \;\leq\; E[\epsilon_n]$$

so that the resubstitution error estimate is on average smaller even than the Bayes error.

Hint: You can show that $E[\hat{\epsilon}_r] \leq \epsilon^*$ by writing $\hat{\epsilon}_r$ as a function of the variables U_i and V_i described in the lecture on the discrete histogram rule, using Jensen's inequality, and then finding $E[U_i]$ and $E[V_i]$ from the fact that U_i and V_i are both binomial random variables (with what parameters?).

Solution: The resubstitution error estimator for the discrete histogram rule is given by:

$$\hat{\epsilon}_r \;=\; \frac{1}{n} \sum_{i=1}^{b} \min\{U_i, V_i\}$$

where

$$\begin{aligned} U_i &= \#\{X_j = i \mid Y_j = 0\}, \quad i = 1, \ldots, b, \\ V_i &= \#\{X_j = i \mid Y_j = 1\}, \quad i = 1, \ldots, b \end{aligned}$$

are the observed bin counts. We have that

$$\begin{aligned} E[\hat{\epsilon}_r] &= E\left[\frac{1}{n}\sum_{i=1}^{b} \min\{U_i, V_i\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{b} E[\min\{U_i, V_i\}] \\ &\leq \frac{1}{n}\sum_{i=1}^{b} \min\{E[U_i], E[V_i]\} \\ &= \frac{1}{n}\sum_{i=1}^{b} \min\{nc_0p_i, nc_1q_i\} \\ &= \sum_{i=1}^{b} \min\{c_0p_i, c_1q_i\} = \epsilon^* \end{aligned}$$

where we used Jensen's Inequality and the fact that U_i and V_i are distributed as binomial random variables with parameters (n, c_0p_i) and (n, c_1q_i), respectively. On the other hand, it is also true that $\epsilon^* \leq \epsilon_n \Rightarrow \epsilon^* \leq E[\epsilon_n]$, establishing the other inequality.

ECEN 649 Pattern Recognition – Spring 2015

Midterm Exam – Solutions

Problem 1. (40 points)

Please answer each of the following questions accurately and concisely, using only the space provided.

(a) What is the difference between a classifier and a classification rule?

A classifier is _____
A classification rule _____

Solution: A classifier is a partition of the feature space. A classification rule is a method that produces a classifier given training data.

(b) Give an example of a consistent classification rule that is not universally consistent.

The _____ classification rule is consistent under the _____ distribution.

Solution: The LDA rule is consistent if the class-conditional densities are Gaussian with identical covariance matrices. This rule is not universally consistent.

(c) The feature-label distribution in a given problem is such that $\eta(X) = 0$ or $\eta(X) = 1$, with probability 1. What is the Bayes error?

$\varepsilon^* =$ _____

Solution: We have $\varepsilon^* = E\left[\min\{\eta(X), 1 - \eta(X)\}\right] = 0$.

(d) Apply the Cover-Hart theorem to prove that the 1-NN classification rule is consistent under the previous feature-label distribution.

Proof: _____

Solution: Cover-Hart Theorem: $\varepsilon^* \leq \varepsilon_{\text{NN}} \leq 2\varepsilon^* \Rightarrow 0 \leq \varepsilon_{\text{NN}} \leq 0$, i.e., $\varepsilon_{\text{NN}} = 0$.

Problem 2. (40 points)

In a univariate classification problem, we have the following model

$$Y \;=\; T[\,\cos(\pi X) + N\,]\,, \quad 0 \leq X \leq 1\,,$$

where X is uniformly distributed on the interval $[0, 1]$, $N \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise term, and $T[\cdot]$ is the standard 0-1 step function.

(a) Find the Bayes classifier.

(b) Find the Bayes error.

Hint: You may assume that $\int_0^{0.5} \Phi(\cos \pi u)\, du = 0.36$, where Φ is the cumulative distribution function for a standard Gaussian distribution.

Solution:

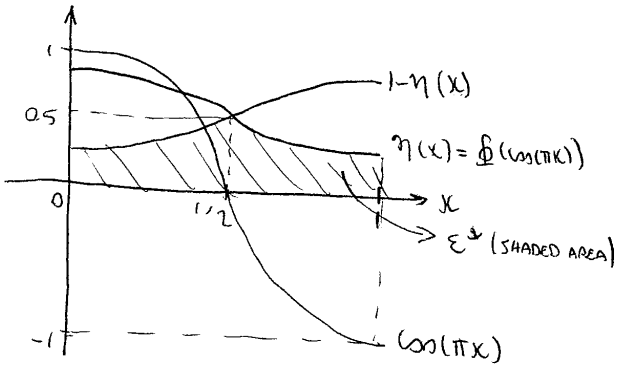
$$\eta(x) = P(Y=1 \mid X=x) = P(\cos(\pi X) + N \geq 0 \mid X = x)$$

$$= P(N \geq -\cos(\pi x)) = 1 - \Phi(-\cos(\pi x)) = \Phi(\cos(\pi x))$$

$$\therefore \eta(x) \geq 1/2 \iff \Phi(\cos(\pi x)) \geq 1/2$$

$$\iff \cos(\pi x) \geq 0 \iff x \in [0, 1/2] \quad (0 \leq x \leq 1)$$

$$\Rightarrow \psi^*(x) = \begin{cases} 1, & \eta(x) \geq 1/2 \\ 0, & \eta(x) < 1/2 \end{cases} = \begin{cases} 1, & 0 \leq x \leq 1/2 \\ 0, & 1/2 < x \leq 1 \end{cases}$$



$$\varepsilon^* = E[\min\{\eta(x), 1-\eta(x)\}]$$

$$= 2\int_0^{0.5} (1-\eta(x))\, dx$$

$$= 2\left(\frac{1}{2} - \int_0^{0.5} \Phi(\cos(\pi x))\, dx\right)$$

$$= 2\left(\frac{1}{2} - 0.36\right) = 0.28 \;//$$

Problem 3. (20 points)

Consider that an experimenter wants to use A 2-D cubic histogram classification rule, with square cells of size h_n, and achieve consistency as the sample size n increases, for any possible distribution of the data. If the experimenter lets h_n decrease as $h_n = \frac{1}{\sqrt{n}}$, would they be guaranteed to achieve consistency and why? If not, how would they need to modify the rate of descrease of h_n to achieve consistency?

Solution: To be guaranteed universal consistency in this case, one must have $h_n \to 0$ and $nh_n^2 \to 0$. With $h_n = 1/\sqrt{n}$, one has $nh_n^2 = n/n \not\to \infty$. This is not sufficient to guarantee consistency, as h_n goes to zero too fast. To be guranteed consistency, one can make, for example, $h_n = 1/\sqrt[4]{n}$, in which case $nh_n^2 = n/\sqrt{n} = \sqrt{n} \to \infty$.

ECEN 649 Pattern Recognition – Spring 2015

Final Exam – Solutions

Problem 1. (30 points)

To get full credit for this problem, you need to give answers that are both complete and *short* (at most 5 lines of text); do not write equations.

(a) What is a sufficient condition to obtain piecewise linear classifiers from a multiple-hidden layer neural network?

Solution: A sufficient condition is that the first hidden layer contain only threshold sigmoids.

(b) How many points does the minimal nonlinearly-separable problem in d dimensions have?

Solution: Since the VC dimension of a hyperplane is $d + 1$, the minimum number of point is $d + 2$.

(c) Describe the basic difference between filter and wrapper feature selection.

Solution: Filter feature selection does not use the classification rule to evaluate candidate feature sets, whereas wrapper feature selection does.

(d) Explain the basic difference between randomized and non-randomized error estimators.

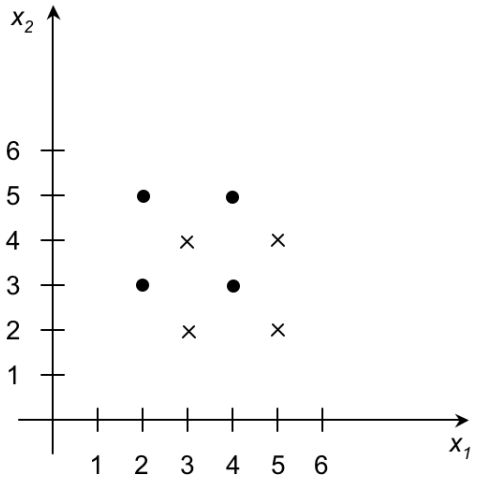
Solution: Non-randomized error estimators are fixed once the training data is specified, while randomized error estimators have internal random factors that produce "internal variance," after the training data is specified.

(e) Explain the basic difference between model selection by minimization of the training error and model selection by structural risk minimization.

Solution: In model selection by minimization of the training error, one uses minimal apparent error as the criterion to select a classification rule, whereas in structural risk minimization, one picks the classification rule that achieves a compromise between minimal apparent error and minimal complexity, by means of penalty term based on the ratio of the VC dimension to the number of samples, in order to avoid overfitting.

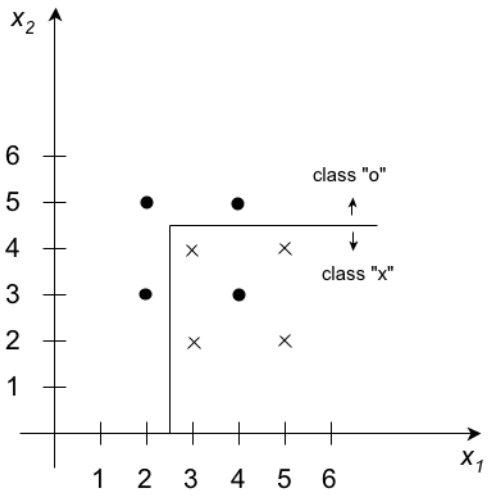
Problem 2. (20 points)

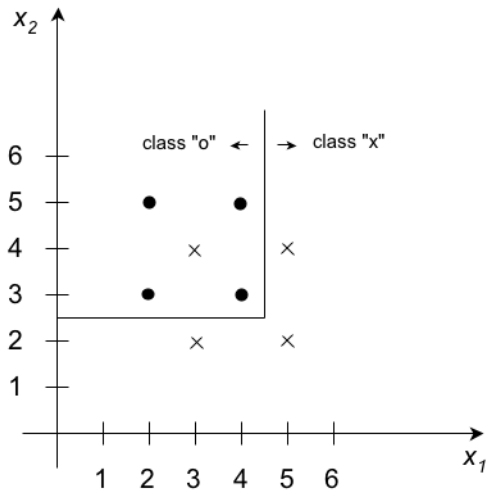
Consider the training data set given in the figure below.

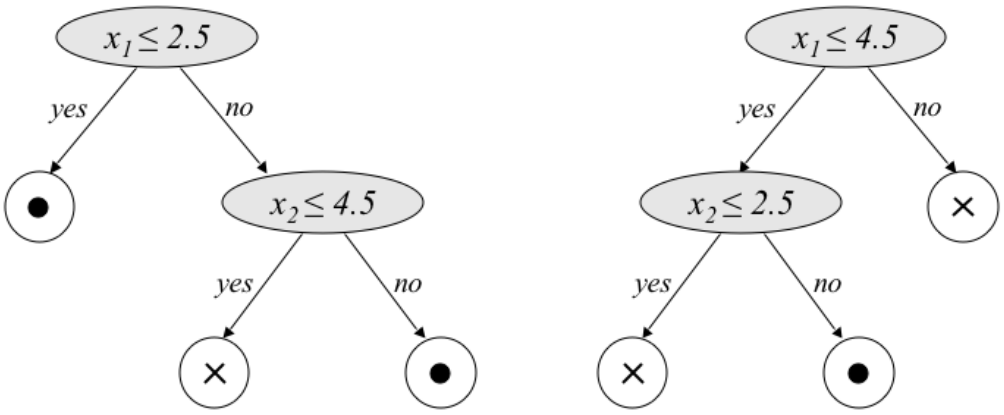


(a) There are exactly two CART classifiers with two splitting nodes (root node plus another node) that produce an apparent error of 0.125. Specify these classifiers in the form of a decision boundary superimposed on the training data and also as a binary decision tree showing the splitting and leaf nodes.

Solution: The decision boundaries and corresponding decision trees for the two CART classifiers are depicted below.







(b) Specify a neural network that can implement either of the two classifiers above, by specifying the number of hidden layers, the number of nodes in each layer, and the nonlinearities. You do NOT need to specify the weights, provided you decribe in detail what each node, including the output node, implements.

Solution: There are two hidden layers with threshold nonlinearities (perceptrons). Each splitting node is implemented by a perceptron in the first hidden layer. On the other hand, each leaf node is implemented by an appropriate boolean function on the outputs of the first hidden layer, which can be done by one perceptron in the second hidden layer, plus appropriate weights for the output node to produce the correct labels. Therefore, there are 2 and 3 nodes in the first and second hidden layers, respectively, plus the output node.

Problem 3. (20 points)

This problem concerns error estimation.

(a) You are given that an error estimator $\hat{\epsilon}_n$ is related to the classification error ϵ_n through the simple model

$$\hat{\epsilon}_n \;=\; \epsilon_n + Z\,,$$

where the conditional distribution of the random variable Z given the training data S_n is Gaussian, $Z \sim N(0, 1/n^2)$. Is $\hat{\epsilon}_n$ randomized or nonrandomized? Find the internal variance and variance of $\hat{\epsilon}_n$. What happens as the sample size grows without bound?

Solution: Since $\hat{\epsilon}_n$ depends on Z, which is random given S_n, the error estimator is randomized. Its internal variance is given by

$$V_{\text{int}} \;=\; \text{Var}(\hat{\epsilon}_n \mid S_n) \;=\; \text{Var}(\epsilon_n + Z \mid S_n) \;=\; \text{Var}(Z \mid S_n) \;=\; \frac{1}{n^2}\,.$$

The variance can be obtained from the conditional variance formula:

$$\text{Var}(\hat{\epsilon}_n) \;=\; E[V_{\text{int}}] + \text{Var}(E[\hat{\epsilon}_n \mid S_n]) \;=\; \frac{1}{n^2} + \text{Var}(\epsilon_n + 0) \;=\; \frac{1}{n^2} + \text{Var}(\epsilon_n)\,.$$

As the sample size grows without bound, the internal variance tends to zero, and the variance of the error estimator becomes equal to just the variance of the true classification error (typically, this will converge to zero as well).

(b) Suppose that a training sample S_n is given, and the true error of the designed classifier is $\epsilon_n = 0.2$. Find the probability that the holdout estimator $\hat{\epsilon}_{n,m}$ will be exactly equal to ϵ_n for this training sample, if $m = 5$ testing samples are available.

Solution: From the lecture slides, we know that

$$P\left(\hat{\epsilon}_{n,m} = \frac{k}{m} \,\middle|\, S_n\right) \;=\; \binom{m}{k} \epsilon_n^k (1-\epsilon_n)^{m-k}, \quad k = 0, \ldots, m\,.$$

Therefore, with $m = 5$, we have

$$P\left(\hat{\epsilon}_{n,m} = \epsilon_n = 0.2 = \frac{1}{5} \,\middle|\, S_n\right) \;=\; \binom{5}{1} 0.2^1\, 0.8^4 \;=\; \frac{2^{12}}{10^4} \;=\; 0.4096\,.$$

Problem 4. (30 points)

This problem concerns dimensionality reduction.

(a) Under equally-likely classes, $p(X|Y=1) \sim N_d(\mu, \sigma^2 I)$ and $p(X|Y=0) \sim N_d(-\mu, \sigma^2 I)$, where the vector μ is known, the variance σ^2 is an unknown parameter, and $d > 2$. Find a two-dimensional sufficient statistic for this problem. Hint: Find $\eta(X)$.

Solution: With equally-likely classes, one has

$$\eta(X) \;=\; P(Y=1|X) \;=\; \frac{p(X|Y=1)}{p(X|Y=1)+p(X|Y=0)} \;=\; \frac{\exp\left(-\frac{1}{2}\frac{||X-\mu||^2}{\sigma^2}\right)}{\exp\left(-\frac{1}{2}\frac{||X-\mu||^2}{\sigma^2}\right) + \exp\left(-\frac{1}{2}\frac{||X+\mu||^2}{\sigma^2}\right)}\,.$$

This is a function X only through $||X-\mu||$ and $||X+\mu||$, therefore $(||X-\mu||\,, ||X+\mu||)$ is a 2-D sufficient statitics, despite the fact that σ^2 is unknown. On the other hand, if μ is unknown, then no dimensionality reduction is possible.

(b) Assume that $X \sim N_4(\mu, \Sigma)$, where

$$\mu \;=\; \begin{bmatrix} 1 \\ 2 \\ -1 \\ 3 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}\,.$$

Obtain the two principal components Z_1 and Z_2.

Solution: Since Σ is diagonal, its eigenvectors coincide with the axis vectors in R^4, and its eigenvalues can be read directly from the diagonal. The two PCs correspond to the directions of the largest eigenvalues, with an additional step of mean removal, so that

$$Z_1 = X_3 + 1 \quad \text{and} \quad Z_2 = X_2 - 2$$