Name: Hye Soo Choi **Student ID: 23274190**

CS 189: Introduction to Machine Learning

Homework 2

Due: February 18, 2016 at 11:59pm

Instructions

- Homework 2 is completely a written assignment; no coding involved.
- We prefer that you typeset your answers using the LaTeX template on bCourses. If there is not enough space for your answer, you may continue your answer on the next page. Make sure to start each question on a new page.
- Neatly handwritten and scanned solutions will also be accepted. Make sure your answers are readable!
- Submit a PDF with your answers to the Homework 2 assignment on Gradescope. You should be able to see CS 189/289A on Gradescope when you log in with your bCourses email address. Please make a Piazza post if you have any problems accessing Gradescope.
- While submitting to Gradescope, you will have to select the pages containing your answer for each question.
- The assignment covers concepts in probability, linear algebra, matrix calculus, and decision theory.
- **Start early. This is a long assignment. Some of the material may not have been covered in lecture; you are responsible for finding resources to understand it.**

Problem 1: Expected Value.

A target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let X be the distance of the hit from the center (in feet), and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot?

Solution: The expected value is calculated as follows:

$$\begin{aligned}
&\int_0^{\frac{1}{\sqrt{3}}} 4 \cdot \frac{2}{\pi(1+x^2)} dx + \int_{\frac{1}{\sqrt{3}}}^1 3 \cdot \frac{2}{\pi(1+x^2)} dx + \int_1^{\sqrt{3}} 2 \cdot \frac{2}{\pi(1+x^2)} dx \\
=&\int_0^{\frac{\pi}{6}} 4 \cdot \frac{2}{\pi(1+\tan^2\theta)} \sec^2\theta d\theta + \int_{\frac{\pi}{6}}^{\frac{\pi}{4}} 3 \cdot \frac{2}{\pi(1+\tan^2\theta)} \sec^2\theta d\theta + \int_{\frac{\pi}{4}}^{\frac{\pi}{3}} 2 \cdot \frac{2}{\pi(1+\tan^2\theta)} \sec^2\theta d\theta \\
=&\int_0^{\frac{\pi}{6}} 4 \cdot 2 d\theta + \int_{\frac{\pi}{6}}^{\frac{\pi}{4}} 3 \cdot 2 d\theta + \int_{\frac{\pi}{4}}^{\frac{\pi}{3}} 2 \cdot 2 d\theta \\
=&8 \cdot \frac{\pi}{6} + 6 \cdot \frac{\pi}{12} + 4 \cdot \frac{\pi}{12} \\
=&\frac{13}{6}\pi
\end{aligned}$$

Problem 2: MLE.

Assume that the random variable X has the exponential distribution

$$f(x;\theta) = \theta e^{-\theta x} \qquad x \geq 0, \theta > 0$$

where θ is the parameter of the distribution. Use the method of maximum likelihood to estimate θ if 5 observations of X are $x_1 = 0.9$, $x_2 = 1.7$, $x_3 = 0.4$, $x_4 = 0.3$, and $x_5 = 2.6$, generated i.i.d. (i.e., independent and identically distributed).

Solution:

$$f(x_1, x_2, x_3, x_4, x_5; \theta) = \prod_{i=1}^{5} f(x_i; \theta) = \theta^5 e^{-\theta \sum_{i=1}^{5} x_i} = \theta^5 e^{-5.9\theta}$$

$$\frac{\partial f(x_1, x_2, x_3, x_4, x_5; \theta)}{\partial \theta} = 5 \cdot \theta^4 e^{-5.9\theta} - 5.9\, \theta^5 e^{-5.9\theta} = \theta^4 e^{-5.9\theta}(5 - 5.9\theta)$$

Under the constraint that $\theta > 0$

$$\frac{\partial f(x_1, x_2, x_3, x_4, x_5; \theta)}{\partial \theta} = 0,$$

which is equivalent to

$$5 - 5.9\theta = 0.$$

Therefore the maximum likelihood estimator of θ is $\hat{\theta} = \frac{50}{59}$.

Definition. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. We say that A is **positive definite** if $\forall x \in \mathbb{R}^n \mid x \neq \vec{0},\ x^\top Ax > 0$. Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n,\ x^\top Ax \geq 0$.

Problem 3: Positive Definiteness.

Let $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n\times n}$ be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

(a) Give an explicit formula for $x^\top Ax$. Write your answer as a sum involving the elements of A and x.

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$).

Solution: (a)

$$\begin{aligned} x^\top Ax &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \\ &= \begin{bmatrix} \sum_{i=1}^n x_i a_{i1} & \sum_{i=1}^n x_i a_{i2} & \cdots & \sum_{i=1}^n x_i a_{in} \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n x_i x_j a_{ij} \end{aligned}$$

(b)

For $\forall i \in [n]$, consider a n-dimensional unit ith coordinate vector $z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}^\top$, where

$$z_k = \begin{cases} 1 & k = i \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$z^\top Az = \sum_{j=1}^n \sum_{i=1}^n z_i z_j a_{ij} = a_{ii} > 0.$$

This is due to the result from previous problem and the assumption that $z_j = 0$ for $j \neq i$.

Problem 4: Short Proofs.

A is symmetric in all parts.

(a) Let A be a positive semidefinite matrix. Show that $A + \gamma I$ is positive definite for any $\gamma > 0$.

(b) Let A be a positive definite matrix. Prove that all eigenvalues of A are greater than zero.

(c) Let A be a positive definite matrix. Prove that A is invertible. (Hint: Use the previous part.)

(d) Let A be a positive definite matrix. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^\top x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^\top B$.)

Solution: (a) For any non-zero vector $x \in \mathbb{R}^n$

$$x^\top (A + \gamma I)x = x^\top Ax + x^\top x > 0,$$

because $x^\top Ax \geq 0$ and $x^\top x > 0$ $(x \neq \vec{0})$. This proves that $A + \gamma I$ is positive definite for any $\gamma > 0$.

(b) If λ is an eigenvalue of A, then there exists a nonzero vector $y \in \mathbb{R}$ such that $Ay = \lambda y$. Then

$$y^\top Ay = y^\top (Ay) = y^\top (\lambda y) = \lambda y^\top y > 0,$$

because A is assumed to be positive definite. Since $y^\top y > 0$, this implies that $\lambda > 0$.

(c) Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be all the eigenvalues of A and $u_1, u_2, \cdots, u_n$ be the corresponding unit eigenvectors of A. It is known from previous problem that all eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ are strictly positive. By Spectral theorem,

$$A = U\Lambda U^\top,$$

where $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$ is orthogonal and $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$. Consider a new matrix $B = UDU^\top$ where $D = diag(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_n})(\lambda_i > 0)$. Then

$$\begin{aligned} AB &= U\Lambda U^\top UDU^\top = U\Lambda DU^\top = UU^\top = I \\ BA &= UDU^\top U\Lambda U^\top = UD\Lambda U^\top = UU^\top = I. \end{aligned}$$

Here we used the fact that $UU^\top = D\Lambda = \Lambda D = I$. This demonstrates that B is the inverse matrix of A and therefore A is invertible.

(d) With Λ, U that we defined in the solution for previous problem, $A = U\Lambda U^\top$. Define $\Lambda^{\frac{1}{2}}$ as $diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n})$; B as $\Lambda^{\frac{1}{2}}U^\top$; and x_i as the ith column vector of B for $i \in [n]$. Then $A = B^\top B = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ and element-wise computation gives the desired result that $A_{ij} = x_i^\top x_j$ for all $i, j \in [n]$. Additionally, because $B = \Lambda^{\frac{1}{2}}U^\top$, $x_i = \lambda_i v_i$ where v_i is the ith column vector of $U^\top$. Since $U^\top U = UU^\top = I$ and therefore $U^\top$ is invertible, its column vector v_i's are linearly independent. This implies that x_i's are also linearly independent because x_i is a constant multiple of v_i for all $i \in [n]$.

Problem 5: Derivatives and Norm Inequalities.

Derive the expression for following questions. Do not write the answers directly.

(a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial\mathbf{x}}$.

(b) Let $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{x}}$.

(c) Let $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times n}$. Derive $\frac{\partial\text{Trace}(\mathbf{XA})}{\partial\mathbf{X}}$.

(d) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$. (Note that $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.) (Hint: The Cauchy-Schwarz inequality may come in handy.)

Solution: (a) Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$ and $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^\top$, then $x^\top a = \sum_{i=1}^n x_i a_i$. Therefore,

$$\frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial\mathbf{x}} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \mathbf{a}.$$

(b)

$$\mathbf{x}^\top A\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n A_{ij}x_i x_j.$$

Therefore,

$$\frac{\partial\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\partial x_i} = \sum_{j=1,j\neq i}^n A_{ij}x_j + 2A_{ii}x_i + \sum_{j=1,j\neq i}^n A_{ji}x_j = \sum_{j=1}^n A_{ij}x_j + \sum_{j=1}^n A_{ji}x_j$$

for all $i \in [n]$. Writing this result in matrix notation,

$$\frac{\partial\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\partial\mathbf{x}} = A\mathbf{x} + A^\top\mathbf{x}.$$

(c)

$$\text{Trace}(\mathbf{XA}) = \sum_{i=1}^n \sum_{k=1}^n X_{ik}A_{ki}.$$

Therefore,

$$\frac{\partial\text{Trace}(\mathbf{XA})}{\partial X_{ij}} = A_{ji}, \quad \text{for all } j, i \in [n].$$

It follows that $\frac{\partial\text{Trace}(\mathbf{XA})}{\partial\mathbf{X}} = A$.

(d)

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2 \le \sum_{i=1}^n x_i^2 + \sum_{i=1}^n \sum_{j=1,j\neq i}^n |x_i||x_j| = \|\mathbf{x}\|_1^2.$$

Taking square roots proves the desired result that $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$. By Cauchy-Schwarz ineqaulity,

$$\|\mathbf{x}\|_1^2 = (\sum_{i=1}^n |x_i|)^2 \le (\sum_{i=1}^n 1)(\sum_{i=1}^n |x_i|^2) = n\|\mathbf{x}\|_2^2.$$

Taking square roots proves the desired result that $\|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$.

Problem 6: Weighted Linear Regression.

Let $\mathbf{X}$ be a $n \times d$ data matrix, $\mathbf{Y}$ be the corresponding $n \times 1$ target/label matrix and $\mathbf{\Lambda}$ be the diagonal $n \times n$ matrix containing a weight for each example. More explicitly, we have

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \cdots \\ (\mathbf{x}^{(n)})^T \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \cdots \\ \mathbf{y}^{(n)} \end{bmatrix} \qquad \mathbf{\Lambda} = \text{diag}(\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(n)})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $\mathbf{y}^{(i)} \in \mathbb{R}$, and $\lambda^{(i)} > 0 \quad \forall \; i \in \{1 \ldots n\}$. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{\Lambda}$ are fixed and known.

In this question, we will try to fit a weighted linear regression model for this data. We want to find the value of weight vector $\mathbf{w}$ which best satisfies the following equation $\mathbf{y}^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$, where ϵ is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk (cost) function is defined as follows:

$$\begin{aligned} R[\mathbf{w}] &= \sum_{i=1}^{n} \lambda^{(i)}(\epsilon^{(i)})^2 \\ &= \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})^2 \end{aligned}$$

(a) Write this risk function $R[\mathbf{w}]$ in matrix notation (i.e., in terms of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Lambda}$ and $\mathbf{w}$).

(b) Find the weight vector $\mathbf{w}$ that minimizes the risk function obtained in the previous part. You can assume that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is full rank. (Hint: You may use the expression you derived in Question 5(b).)

(c) The L_2 regularized risk function, for $\gamma > 0$, is

$$R[\mathbf{w}] = \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})^2 + \gamma\|\mathbf{w}\|_2^2$$

Rewrite this new risk function in matrix notation as in (a) and solve for $\mathbf{w}$ as in (b).

(d) How does γ affect the regression model? How does this fit in with what you already know about L_2 regularization? (Hint: Observe the different expressions for $\mathbf{w}$ obtained in (b) and (c).)

Solution: (a)

$$(\mathbf{X}\mathbf{w} - \mathbf{Y})^\top \Lambda (\mathbf{X}\mathbf{w} - \mathbf{Y}) = \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{Y}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{Y}$$

(b) By the solution for Problem 5,

$$\frac{\partial(\mathbf{X}\mathbf{w} - \mathbf{Y})^\top \Lambda (\mathbf{X}\mathbf{w} - \mathbf{Y})}{\partial \mathbf{w}} = 2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{Y}$$

It follows that the $\mathbf{w}$ that minimizes the risk is $\hat{\mathbf{w}}$ satisfying that

$$\frac{\partial(\mathbf{X}\mathbf{w} - \mathbf{Y})^\top \Lambda (\mathbf{X}\mathbf{w} - \mathbf{Y})}{\partial \mathbf{w}} = 2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{Y} = 0.$$

As a result $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$.

(c) The new risk function in matrix notation is written in matrix notation as follows;

$$R[\mathbf{w}] = (\mathbf{X}\mathbf{w} - \mathbf{Y})^\top \Lambda (\mathbf{X}\mathbf{w} - \mathbf{Y}) + \gamma \mathbf{w}^\top \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{Y}^\top \mathbf{X}\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{Y}^\top \mathbf{Y} + \gamma \mathbf{w}^\top \mathbf{w}.$$

$$\frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}} = 2\mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{X}^\top \mathbf{Y} + 2\gamma \mathbf{w}.$$

Therefore the $\mathbf{w}$ that minimizes the risk is $\hat{\mathbf{w}}$ satisfying that

$$\frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}} = 2\mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{X}^\top \mathbf{Y} + 2\gamma \mathbf{w} = 0.$$

In other words, $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \gamma I)^{-1}\mathbf{X}^\top \mathbf{Y}$.

(d) When γ is negligibly small, then the estimator for this new risk would be very similar to the estimator for square error. If γ is large, then the model strongly penalizes a large L_2 norm of the coefficient vector $\mathbf{w}$, preventing the model from exhibitng high variance. Also, the solution adds a positive constant to the diagonal of $\mathbf{X}^\top \mathbf{X}$ before inversion. As we derived from problem 5(a), this makes the problem nonsingular, even if $\mathbf{X}^\top \mathbf{X}$ is not of full rank.

Problem 7: Classification.

Suppose we have a classification problem with classes labeled $1, \dots, c$ and an additional doubt category labeled as $c+1$. Let the loss function be the following:

$$\ell(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \dots, c\} \\ \lambda_r & \text{if } i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where λ_r is the loss incurred for choosing doubt and λ_s is the loss incurred for making a misclassification. Note that $\lambda_r \geq 0$ and $\lambda_s \geq 0$.

Hint: The risk of classifying a new datapoint as class $i \in \{1, 2, \dots, c+1\}$ is

$$R(\alpha_i|x) = \sum_{j=1}^{c} \ell(f(x) = i, y = j) P(\omega_j|x)$$

(a) Show that the minimum risk is obtained if we follow this policy: (1) choose class i if $P(\omega_i|x) \geq P(\omega_j|x)$ for all j and $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$, and (2) choose doubt otherwise.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Is this consistent with your intuition?

Solution: (a)

$$R(\alpha_i|x) = \sum_{j=1}^{c} \ell(f(x) = i, y = j) P(\omega_j|x) = \begin{cases} \lambda_s \sum_{j=1, j \neq i}^{c+1} P(\omega_j|x) & i \in \{1, 2, \cdots, c\} \\ \lambda_r & i = c+1 \end{cases}$$

Therefore for $i, k \in \{1, 2, \cdots, c\}$,

$$R(\alpha_i|x) \leq R(\alpha_k|x)$$

if and only if (erasing common terms)

$$\lambda_s P(\omega_k|x) \leq \lambda_s P(\omega_i|x),$$

in other words $P(\omega_k|x) \leq P(\omega_i|x)$. Also comparing the risk of choosing a class i and doubt for $i \in \{1, 2, \cdots, c\}$,

$$R(\alpha_i|x) \leq R(\alpha_{c+1}|x)$$

if and only if

$$\lambda_s \sum_{j=1, j \neq i}^{c+1} P(\omega_j|x) \leq \lambda_r.$$

In other words $\lambda_s(1 - P(\omega_i|x)) \leq \lambda_r$, or $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$. To obtain minimum risk, we should choose class i when $P(\omega_k|x) \leq P(\omega_i|x)$ for all k in $\{1, 2, \cdots, c\}$ and $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$; and choose doubt otherwise.

(b)

- If $\lambda_r = 0$, then according to the policy derived in previous problem, we should choose class i if $P(\omega_i|x) = 1$; and choose doubt otherwise.
  This is consistent with intuition because if we don't lose anything from choosing doubt then we always wanna choose doubt unless there is the only one class where x could belong to(with probability 1).

- If $\lambda_s < \lambda_r$, then choose i if $P(\omega_i|x) = \max_{k \in \{1,2,\cdots,c\}} P(\omega_k|x)$; and never choose doubt. Again this is consistent with intuition because if the loss from choosing doubt is strictly greater than the loss from choosing an incorrect class, and therefore to minimize the loss we will avoid choosing doubt and will choose the class with highest conditional probability.

Problem 8: Gaussians.

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with $P(\omega_1) = P(\omega_2) = 1/2$. Here, the classes are ω_1 and ω_2. For this problem, we have $\mu_2 \geq \mu_1$.

(a) Find the optimal Bayes decision boundary (i.e., find x such that $P(\omega_1 \mid x) = P(\omega_2 \mid x)$). What is the corresponding decision rule?

(b) Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$. The Bayes error is the probability of misclassification:

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2) P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1) P(\omega_1).$$

Solution:

$$\begin{aligned} P(\omega_i|x) &= \frac{P(x|\omega_i) \cdot \frac{1}{2}}{P(x|\omega_1) \cdot \frac{1}{2} + P(x|\omega_2) \cdot \frac{1}{2}} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}} \\ &= \frac{e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}} \end{aligned}$$

Therefore,

$$P(\omega_1|x) = P(\omega_2|x)$$

if and only if (canceling common denominator and comparing exponents)

$$-\frac{(x-\mu_1)^2}{2\sigma^2} = -\frac{(x-\mu_2)^2}{2\sigma^2}.$$

Equivalently,

$$(x-\mu_1)^2 = (x-\mu_2)^2.$$

Since this tells that x is equidistant from μ_1 and μ_2, or $x = \frac{\mu_1+\mu_2}{2}$. As a result, the corresponding decision rule is to choose

$$\begin{cases} \omega_1 & \text{if } x < \frac{\mu_1+\mu_2}{2}, \\ \omega_2 & \text{if } x \geq \frac{\mu_1+\mu_2}{2}. \end{cases}$$

the Bayes error associated with this decision rule is

$$\begin{aligned}
P_e &= P((\text{misclassified as } \omega_1) \mid \omega_2)P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1)P(\omega_1)\\
&= \int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} dx \cdot \frac{1}{2} + \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \cdot \frac{1}{2}\\
&= \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \cdot \frac{1}{2} + \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \cdot \frac{1}{2}
\end{aligned}$$

(used the change of variable: $z = \dfrac{-(x-\mu_2)}{2}$ for the first integral, $z = \dfrac{(x-\mu_1)}{2}$ for the second)

$$= \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$