Name: Jesse Li **Student ID: 23822462**

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:

Expected value $= 4\int_0^{frac1\sqrt{3}} \frac{2}{\pi(1+x^2)}dx + 3\int_{1/\sqrt{3}}^{1} \frac{2}{\pi(1+x^2)}dx + 2\int_1^{\sqrt{3}} \frac{2}{\pi(1+x^2)}dx$

$= 4(\frac{2tan^{-1}(\frac{1}{\sqrt{3}})}{\pi} - \frac{2tan^{-1}(0)}{\pi}) + 3(\frac{2tan^{-1}(1)}{\pi} - \frac{2tan^{-1}(\frac{1}{\sqrt{3}})}{\pi}) + 2(\frac{2tan^{-1}(\sqrt{3})}{\pi} - \frac{2tan^{-1}(1)}{\pi})$

$= \frac{4}{3} + \frac{1}{2} + \frac{1}{3}$

$= \frac{13}{6}$

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:

The observations are independent, so the likelihood function, L, is given by:

$L = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \exp(-\theta \sum_{i=1}^{n} x_i)$

From lecture we know that maximizing the log-likelihood function is equivalent to maximizing the likelihood function, so we can set the derivative of the log-likelihood function to 0 to find the maximum log-likelihood.

$\frac{d}{d\theta}(\ln(\theta^n \exp(-\theta \sum_{i=1}^{n} x_i))) = 0$

$\frac{d}{d\theta}(n \ln(\theta) + (-\theta \sum_{i=1}^{n} x_i)) = 0$

$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$

$\theta = n / \sum_{i=1}^{n} x_i = 5/(0.9 + 1.7 + 0.4 + 0.3 + 2.6)$

$= 0.84746$

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

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 $x^\top A x$. 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)

$$x^\top A x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 a_{11} + x_2 a_{12} + \cdots + x_n a_{1n} \\ x_1 a_{21} + x_2 a_{22} + \cdots + x_n a_{2n} \\ \vdots \\ x_1 a_{n1} + x_2 a_{n2} + \cdots + x_n a_{nn} \end{bmatrix}$$

$= (x_1 x_1 a_{11} + x_1 x_2 a_{21} + \cdots + x_1 x_n a_{n1}) + (x_1 x_2 a_{12} + x_2 x_2 a_{22} + \cdots + x_2 x_n a_{n2}) + \cdots + (x_1 x_n a_{1n} + x_n x_2 a_{2n} + \cdots + x_n x_n a_{nn})$

$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j a_{ij}$

(b) Proof by contradiction:

We will define a positive definite matrix $A \in \mathbb{R}^{n \times n}$ which has at least one diagonal entry, a_{ii}, that is less than or equal to 0. Consider the vector, x, of all zeros except for $x_i = 1$. From part (a) we know that $x^\top A x = \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k a_{jk}$. Each of the terms in this sum will be zero except the term $x_i x_i a_{ii}$, so $x^\top A x = a_{ii} <= 0$. This contradicts that A is a positive definite matrix, showing that every diagonal entry of a positive matrix must be positive.

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) We wish to show that $x^\top (A + \gamma I)x > 0$.

$x^\top (A + \gamma I)x = x^\top Ax + \gamma x^\top Ix = x^\top Ax + \gamma x^\top x$

We know that $x^\top Ax$ is positive, so we just need to show that $\gamma x^\top x$ is positive as well.

$\gamma x^\top x = \gamma \sum_{i=1}^n x_i^2$

$\gamma \sum_{i=1}^n x_i^2$ is greater than 0 when $x \neq \vec{0}$ and $\gamma > 0$, so $x^\top Ax + \gamma x^\top x$ is greater than 0 and $A + \gamma I$ is positive definite.

(b) The eigenvalues λ of a positive definite matrix A are given by the equation $Ax = \lambda x$. We can multiply both sides of the equation by $x^\top$ to get:

$x^\top Ax = \lambda x^\top x$

We know that $x^\top Ax$ is positive by the definition of a positive definite matrix. Since $x^\top x$ is always positive ($x^\top x = \sum_{i=1}^n x_i^2$), we know that the eigenvalues are all positive.

(c) From the invertible matrix theorem, a matrix is invertible if the number 0 is not an eigenvalue of that matrix. The invertible matrix theorem is a collection of statements that must be either all true or all false for a square matrix. From part (b) we know that all eigenvalues of A are positive, so 0 is not an eigenvalue of A. Therefore, A is invertible.

(d) From the spectral theorem, we know that $A = U\Lambda U^\top$, where U is a matrix with the eigenvectors of A as columns, and Λ is a diagonal matrix with the eigenvalues of A as the diagonal entries. From part (b) we know that all eigenvalues of A are positive, so we can express A as:

$$A = U \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} U^\top = U \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} U^\top$$

$= B^\top B$

From here, we see that vectors $b_1, b_2, ..., b_n$ exist such that $A_{ij} = b_i^\top b_j$. Since A is symmetric, the eigenvectors of A are linearly independent. Every b_i is simply an eigenvector of A multiplied by a scalar (the square root of λ_i), so $b_1, b_2, ..., b_n$ are linearly independent as well.

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 $\frac{\partial \text{Trace}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}}$.

(d) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$. (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) $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial \mathbf{x}} = \frac{\partial(\sum_{i=1}^n x_i a_i)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial(\sum_{i=1}^n x_i a_i)}{\partial x_1} & \frac{\partial(\sum_{i=1}^n x_i a_i)}{\partial x_2} & \cdots & \frac{\partial(\sum_{i=1}^n x_i a_i)}{\partial x_n} \end{bmatrix}^T = \mathbf{a}$

(b) $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\sum_{i=1}^n \sum_{j=1}^n A_{ij}x_i x_j)}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}\left(\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_i x_j + \sum_{i\neq k} A_{ik}x_i x_k + \sum_{j\neq k} A_{kj}x_j x_k + A_{kk}x_k^2\right) =$
$\begin{bmatrix} \sum_{i=1}^n A_{i1}x_i + \sum_{j=1}^n A_{1j}x_j \\ \sum_{i=1}^n A_{i2}x_i + \sum_{j=1}^n A_{2j}x_j \\ \vdots \\ \sum_{i=1}^n A_{in}x_i + \sum_{j=1}^n A_{nj}x_j \end{bmatrix} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$

(c) $\frac{\partial \text{Trace}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}}\left(\sum_{i=1}^n x_{1i}a_{i1} + \sum_{i=1}^n x_{2i}a_{i2} + \cdots + \sum_{i=1}^n x_{ni}a_{in}\right) = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}$

$= \mathbf{A}^T$

(d) Since $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1$ are never negative, proving $(\|\mathbf{x}\|_2)^2 \leq (\|\mathbf{x}\|_1)^2$ will prove $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$.

$(\|\mathbf{x}\|_2)^2 = \sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + ... + x_n^2$

$(\|\mathbf{x}\|_1)^2 = (\sum_{i=1}^n |x_i|)^2 = x_1^2 + x_2^2 + ... + x_n^2 + 2|x_1||x_2| + 2|x_1||x_3| + ...$

From this we can see that $(\|\mathbf{x}\|_2)^2 \leq (\|\mathbf{x}\|_1)^2$, so $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$.

Now we must show that $\|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$. The Cauchy-Schwarz inequality tells us that:

$\left(\sum_{i=1}^n a_i b_i\right)^2 \leq \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$, if $a_i = cb_i$

Let $a_i = 1$ and $b_i = x_i$:

$\left(\sum_{i=1}^n x_i\right)^2 \leq \left(\sum_{i=1}^n 1^2\right)\left(\sum_{i=1}^n x_i^2\right)$

$\left(\sum_{i=1}^n x_i\right)^2 \leq (n)\left(\sum_{i=1}^n x_i^2\right)$

Taking the square root of both sides, we get:

$\sum_{i=1}^n x_i \leq \sqrt{n}\sqrt{\sum_{i=1}^n x_i^2}$,

$\|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$

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

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

where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $\mathbf{y}^{(i)} \in \mathbb{R}$, and $\lambda^{(i)} > 0 \quad \forall \ i \in \{1 \dots n\}$. $\mathbf{X}$, $\mathbf{Y}$ 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) $R[\mathbf{w}] = (\mathbf{X}\mathbf{w} - \mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w} - \mathbf{Y})$

(b) We want to find $\mathbf{w}$ such that:

$0 = \frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left((\mathbf{X}\mathbf{w} - \mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w} - \mathbf{Y}) \right) = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} - \mathbf{Y}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} + \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y} \right)$

$= \frac{\partial}{\partial \mathbf{w}} \text{tr} \left(\mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} - \mathbf{Y}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} + \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y} \right) = \frac{\partial}{\partial \mathbf{w}} \left(tr(\mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}) - tr(2\mathbf{Y}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}) \right)$

$= 2\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}$

Solving for $\mathbf{w}$, using the fact that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is invertible:

$0 = \frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}} = 2\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \rightarrow \mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}$

$\mathbf{w} = (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}$

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

$\frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}((\mathbf{X}\mathbf{w} - \mathbf{Y})^T \mathbf{\Lambda} (\mathbf{X}\mathbf{w} - \mathbf{Y})) + \frac{\partial}{\partial \mathbf{w}}(\gamma \mathbf{w}^T \mathbf{w}) = 2\mathbf{X}^T \mathbf{\Lambda} \mathbf{X}\mathbf{w} - 2\mathbf{X}^T \mathbf{\Lambda} \mathbf{Y} + 2\gamma \mathbf{w}$

Setting $\frac{\partial R[\mathbf{w}]}{\partial \mathbf{w}}$ to 0 and solving for $\mathbf{w}$:

$0 = \mathbf{X}^T \mathbf{\Lambda} \mathbf{X}\mathbf{w} - \mathbf{X}^T \mathbf{\Lambda} \mathbf{Y} + \gamma \mathbf{w}$

$0 = -\mathbf{X}^T \mathbf{\Lambda} \mathbf{Y} + \mathbf{X}^T \mathbf{\Lambda} \mathbf{X}\mathbf{w} + \gamma \mathbf{I}\mathbf{w}$

$0 = -\mathbf{X}^T \mathbf{\Lambda} \mathbf{Y} + (\mathbf{X}^T \mathbf{\Lambda} \mathbf{X} + \gamma \mathbf{I})\mathbf{w}$

$\mathbf{w} = (\mathbf{X}^T \mathbf{\Lambda} \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^T \mathbf{\Lambda} \mathbf{Y}$

(d) Increasing γ increases the penalty the regression model places on large weight values. By adding the L_2 norm to the risk function, our weights will be pulled closer to zero. This can be useful in preventing overfitting.

Problem 7: Classification.

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

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

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 $i \in \{1, 2, \ldots, c+1\}$ 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) If we choose to categorize the datapoint as doubt:

$R(\alpha_{c+1}|x) = \sum_{j=1}^{c} \lambda_r P(\omega_j|x) = \lambda_r$

If we choose to categorize the datapoint as i:

$R(\alpha_i|x) = \left(\sum_{j=1}^{c} \lambda_s P(\omega_j|x)\right) - \lambda_s P(\omega_i|x) = \lambda_s(1 - P(\omega_i|x))$

Therefore, choosing class i will minimize risk when:

$R(\alpha_i|x) \leq R(\alpha_{c+1}|x) \rightarrow \lambda_s(1 - P(\omega_i|x)) \leq \lambda_r \rightarrow P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$

Additionally, in order to minimize risk when we choose not to categorize as doubt, we must choose the class i that has the highest $P(\omega_i|x)$:

$R(\alpha_i|x) \leq R(\alpha_j|x) \rightarrow \lambda_s(1 - P(\omega_i|x)) \leq \lambda_s(1 - P(\omega_j|x)) \rightarrow P(\omega_i|x) \geq P(\omega_j|x)$

(b) When $\lambda_r = 0$, $R(\alpha_{c+1}|x) = 0$ and it will always be optimal to choose to categorize as doubt. Intuitively this makes sense, because if there is no loss associated with categorizing as doubt, then we can minimize our loss by always choosing to classify as doubt.

When $\lambda_r > \lambda_s$:

$\lambda_s(1 - P(\omega_i|x)) \leq \lambda_s < \lambda_r$

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

So if $\lambda_r > \lambda_s$, then it will never be optimal to choose to categorize as doubt. If the penalty for categorizing as doubt is higher than the penalty for making a misclassifaction, it makes sense intuitively that it is always better to try to classify the dataset as not doubt.

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:

(a) Using Bayes' theorem, we get that:

$$P(\omega_1 \mid x) = \frac{P(x \mid \omega_1)P(\omega_1)}{P(x)} = \frac{P(x \mid \omega_1)}{2P(x)}$$

$$P(\omega_2 \mid x) = \frac{P(x \mid \omega_2)P(\omega_2)}{P(x)} = \frac{P(x \mid \omega_2)}{2P(x)}$$

Setting the two equal to each other, we see that the optimal Bayes decision boundary occurs when $P(x \mid \omega_1) = P(x \mid \omega_2)$. This occurs right between the peaks of the two distributions, so the optimal Bayes decision boundary is at $x = \dfrac{\mu_2 + \mu_1}{2}$. The corresponding decision rule is to classify as ω_1 if the unknown point is left of $x = \dfrac{\mu_2 + \mu_1}{2}$, and to classify as ω_2 otherwise.

(b) The probability of misclassification will be the overlapping area between the two Gaussian curves after accounting for the prior. Both $P((\text{misclassified as } \omega_1) \mid \omega_2)$ and $P((\text{misclassified as } \omega_2) \mid \omega_1)$ can be found using z-scores for the optimal Bayes decision boundary, $x = \dfrac{\mu_2 - \mu_1}{2}$. Because of the symmetry of the problem, $P((\text{misclassified as } \omega_1) \mid \omega_2) = P((\text{misclassified as } \omega_2) \mid \omega_1) = P_e$. Using the z-score we are able to obtain the value for a in the integral lower bound.

$$z = \frac{\dfrac{\mu_2 + \mu_1}{2} - \mu_1}{\sigma} = \frac{\mu_2 - \mu_1}{2\sigma}$$