Name: **Student ID:**

CS 189: Introduction to Machine Learning

Homework 2 Solutions

Due: February 25, 2016 at 11:59pm

Instructions

- Homework 2 is completely a written assignment; no coding involved.
- We prefer that you typeset your answers using this LaTeX template: hw2.tex. 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.**

Homework Parties

- February 22, 11:30-1:30 PM in 212 Cory
- February 23, 3:30-5:00 PM in 430 Soda

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

$$\begin{aligned} &\int_0^{1/\sqrt{3}} 4\frac{2}{\pi(1+x^2)}dx + \int_{1/\sqrt{3}}^{1} 3\frac{2}{\pi(1+x^2)}dx + \int_1^{\sqrt{3}} 2\frac{2}{\pi(1+x^2)}dx \\ &= \frac{2}{\pi}\left[4\left(\tan^{-1}\frac{1}{\sqrt{3}} - \tan^{-1}0\right) + 3\left(\tan^{-1}1 - \tan^{-1}\frac{1}{\sqrt{3}}\right) + 2\left(\tan^{-1}\sqrt{3} - \tan^{-1}1\right)\right] \\ &= \boxed{\frac{13}{6}} \end{aligned}$$

Problem 2: MLE.

Assume that the random variable X has the exponential distribution

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

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: We'll solve the general case for the MLE of an exponential distribution, then plug in the numbers we have.

$$\begin{aligned}\mathcal{L}(x_1, x_2, ..., x_n; \theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right)\end{aligned}$$

Finding the log-likelihood:

$$\ell(x_1, x_2, ..., x_n; \theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

Taking the derivative with respect to θ:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^{n} x_i}\end{aligned}$$

Plugging in our values for x_i, we get $\hat{\theta} = 0.85$.

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)

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

(b) Let $i \in [1, n]$, and let e_i be the i^{th} standard basis vector (that is, the vector of all zeros except for a single 1 in the i^{th} position). Since A is positive definite, we have $e_i^\top A e_i = a_{ii} > 0$.

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) Let $x \neq 0$. Then

$$\begin{aligned} x^\top (A + \gamma I) x &= x^\top A x + x^\top \gamma I x \\ &= x^\top A x + \gamma \|x\|^2 \\ &> 0 \end{aligned}$$

because $x^\top A x \geq 0$ (since A is positive semidefinite) and $\|x\|^2 > 0$ (because $x \neq 0$). Hence $A + \gamma I$ is positive definite.

(b) We know $x^\top A x > 0$ for any x. Consider v to be any eigenvector with $Av = \lambda v$. Then $v^\top A v = \lambda v^\top v > 0$. Since $v^\top v > 0$ (by definition, v is non-zero), we must have $\lambda > 0$.

(c) Since A is positive definite, all eigenvalues are positive. But then if A is not invertible, 0 is an eigenvalue, which is a contradiction. Thus A must be invertible.

(d) Because A is symmetric positive definite, we diagonalize to obtain $A = P^\top D P$ with orthogonal P and diagonal matrix D with eigenvalues on the diagonal. Since all eigenvalues are positive, we can define $E = D^{1/2}$. Then $A = P^\top E E P = P^\top E^\top E P = (EP)^\top EP$. We thus define $x_1, x_2, ..., x_n$ as the columns of the matrix EP, which are linearly independent since P is orthogonal. As we desired, $A_{ij} = x_i^\top x_j$.

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) Let $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$ and $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix}$.

$$\mathbf{x}^T\mathbf{a} = \sum_{i=1}^n x_i a_i$$

Taking partial derivative wrt a component, we get

$$\frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial x_k} = a_k$$

Placing all partial derivatives into a single vector, we get

$$\frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial x_1} \\ \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial x_2} \\ \cdots \\ \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = \mathbf{a}$$

(b) Let $\mathbf{A} = [a_{ij}]_{n\times n}$. We can write

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i,j=1}^n a_{ij} x_i x_j$$

Taking partial derivative wrt a component, we get

$$\begin{aligned}
\frac{\partial\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\partial x_k} &= \frac{\partial}{\partial x_k}\left(\sum_{i,j=1}^{n} a_{ij}x_i x_j\right) \\
&= \frac{\partial}{\partial x_k}\left(x_1\sum_{j=1}^{n} a_{1j}x_j + x_2\sum_{j=1}^{n} a_{2j}x_j + \cdots + x_k\sum_{j=1}^{n} a_{kj}x_j + \cdots + x_n\sum_{j=1}^{n} a_{nj}x_j\right) \\
&\quad \text{(Use product rule of differentiation, i.e. } (fg)' = f'g + fg'\text{), on each term)} \\
&= x_1a_{1k} + x_2a_{2k} + \cdots + x_ka_{kk} + \sum_{j=1}^{n} a_{kj}x_j + \cdots + x_na_{nk} \\
&= (x_1a_{1k} + x_2a_{2k} + \cdots + x_ka_{kk} + \cdots + x_na_{nk}) + \left(\sum_{j=1}^{n} a_{kj}x_j\right) \\
&= \left(\sum_{i=1}^{n} a_{ik}x_i\right) + \left(\sum_{j=1}^{n} a_{kj}x_j\right) \\
&= \left(k^{th} \text{ column of } \mathbf{A}\right)^T\mathbf{x} + \left(k^{th} \text{ row of } \mathbf{A}\right)^T\mathbf{x}
\end{aligned}$$

Placing all partial derivatives into a single vector, we get

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

(c) Let $\mathbf{A} = [a_{ij}]_{n\times n}$ and $\mathbf{X} = [x_{ij}]_{n\times n}$. We can write

$$\text{Trace}(\mathbf{XA}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}a_{ji}$$

Taking partial derivative wrt a component, we get

$$\begin{aligned}
\frac{\partial\text{Trace}(\mathbf{XA})}{\partial x_{ij}} &= \frac{\partial}{\partial x_{ij}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}a_{ji}\right) \\
&= a_{ji}
\end{aligned}$$

Placing all partial derivatives into the matrix, we get

$$\frac{\partial\text{Trace}(\mathbf{XA})}{\partial\mathbf{X}} = [a_{ji}]_{n\times n} = \mathbf{A}^T$$

(d) First let's prove the left-hand side inequality as follows:

$$\begin{aligned}
\|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} \\
&= \sqrt{x_1^2 + x_2^2 \cdots + x_n^2} \\
&\quad \text{(adding positive terms)} \\
&\leq \sqrt{x_1^2 + x_2^2 \cdots + x_n^2 + 2\Big(\sum_{1 \leq i < j \leq n} |x_i||x_j|\Big)} \\
&= \sqrt{(|x_1| + |x_2| + \cdots + |x_n|)^2} \\
&= |x_1| + |x_2| + \cdots + |x_n| \\
&= \|\mathbf{x}\|_1 \\
\implies \|\mathbf{x}\|_2 &\leq \|\mathbf{x}\|_1
\end{aligned}$$

Let's now prove the right-hand side inequality as follows:

$$\begin{aligned}
&\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n| \\
\implies &\|\mathbf{x}\|_1 = \underbrace{(|x_1|, |x_2|, \ldots, |x_n|)^T}_{\text{call this vector } \mathbf{x}'} \bullet (1, 1, \ldots 1) \\
\implies &\|\mathbf{x}\|_1 = \mathbf{x}'^T \bullet \mathbf{1} \\
&\quad \text{(Using Cauchy}-\text{Schwarz inequality on the right)} \\
\implies &\|\mathbf{x}\|_1 \leq \|\mathbf{x}'\|_2 \|\mathbf{1}\|_2 \\
&\quad \text{Note: } \|\mathbf{x}'\|_2 = \|\mathbf{x}\|_2 \text{ and } \|\mathbf{1}\|_2 = \sqrt{n} \\
\implies &\|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2
\end{aligned}$$

Thus, we have shown

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

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) We can re-write the risk function $R[\mathbf{w}]$ in matrix notation as follows

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

(b) We minimize the risk function, i.e.,

$$\begin{aligned}
&\min_{\mathbf{w}} \; R[\mathbf{w}] \\
=&\min_{\mathbf{w}} \; (\mathbf{X}\mathbf{w}-\mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w}-\mathbf{Y}) \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}\mathbf{Y}-\mathbf{Y}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} \\
&\quad \text{(Note that } \mathbf{\Lambda}^T=\mathbf{\Lambda}\text{)} \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{Y}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w} \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-\mathbf{w}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})-(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w} \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w} \qquad (1)
\end{aligned}$$

To take the derivative w.r.t. $\mathbf{w}$, we use results from 5(a) and 5(b).

$$\begin{aligned}
\implies &\frac{\partial}{\partial\mathbf{w}}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w}\right) \\
&= (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}+(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^T)\mathbf{w}+0-2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&= 2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w}-2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&\quad \text{(Set derivative equal to zero and solve)} \\
&= 0 \\
\implies &(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w}=\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&\quad \text{(Given that } (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}) \text{ is full rank and thus invertible)} \\
\implies &\mathbf{w}=(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}
\end{aligned}$$

(c) For the risk function with a regularizer term, we have

$$\begin{aligned}
R[\mathbf{w}] &= \sum_{i=1}^{n}\lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)}-y^{(i)})^2+\gamma\|\mathbf{w}\|_2^2 \\
&= (\mathbf{X}\mathbf{w}-\mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w}-\mathbf{Y})+\gamma\mathbf{w}^T\mathbf{w} \qquad (2)
\end{aligned}$$

We can minimize it as in part (b).

$$\begin{aligned}
&\min_{\mathbf{w}} \; R[\mathbf{w}] \\
=&\min_{\mathbf{w}} \; (\mathbf{X}\mathbf{w}-\mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w}-\mathbf{Y})+\gamma\mathbf{w}^T\mathbf{w} \\
&\quad \text{(Using similar calculation as in part (b), we obtain)} \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w}+\gamma\mathbf{w}^T\mathbf{w} \\
=&\min_{\mathbf{w}} \; \mathbf{w}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}+\gamma\mathbf{I})\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w}
\end{aligned}$$

This expression is very similar to equation (1) from part (b), so we solve for $\mathbf{w}$ similarly.

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}+\gamma\mathbf{I})\mathbf{w}=\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
\implies &\mathbf{w}=(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \qquad (3)
\end{aligned}$$

(d) With regularization, we have essentially added a penalty to prevent the magnitude of $\mathbf{w}$ from becoming large. This is reflected in our solution in equation (3), where γ appears in the inverse. Thus, the larger the value of γ is, the smaller the magnitude of $\mathbf{w}$ will be. This can help prevent over-fitting on the training data, where the hyperparameter γ is obtained using cross-validation.

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) Define $\lambda_{ij} = \ell(f(x) = i, y = j)$. The risk of classifying a new datapoint as class i is

$$R(\alpha_i|x) = \sum_j \lambda_{ij} P(\omega_j|x) = \lambda_s(1 - P(\omega_i|x)),$$

and the risk of classifying the new datapoint as doubt is

$$R(\alpha_{c+1}|x) = \lambda_r \sum_j P(\omega_j|x) = \lambda_r.$$

For choosing doubt to be better than choosing any of the classes, the ratio of the risks must satisfy

$$1 > \frac{R(\alpha_{c+1}|x)}{R(\alpha_i|x)} = \frac{\lambda_r}{\lambda_s(1 - P(\omega_i|x))} \implies P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s}$$

for all i. This means that any particular i for which $P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ should not be assigned doubt. In this case, the class to choose must be

$$\arg\min_{1 \leq i \leq c} R(\alpha_i|x) = \arg\min_{1 \leq i \leq c} \lambda_s(1 - P(\omega_i|x)) = \arg\max_{1 \leq i \leq c} P(\omega_i|x),$$

as required.

(b) If $\lambda_r = 0$, then doubt will always be assigned, since for all i, $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s = 1$ is not satisfied unless $P(\omega_i|x) = 1$.

If $\lambda_r > \lambda_s$, then doubt will never be assigned, since for all i, $P(\omega_i|x) \geq 0 > 1 - \lambda_r/\lambda_s$ always holds.

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) $P(\omega_1 \mid x) = P(\omega_2 \mid x) \rightarrow P(x \mid \omega_1)P(\omega_1) = P(x \mid \omega_2)P(\omega_2) \rightarrow P(x \mid \omega_1) = P(x \mid \omega_2) \rightarrow \mathcal{N}(\mu_1, \sigma^2) = \mathcal{N}(\mu_2, \sigma^2) \rightarrow (x - \mu_1)^2 = (x - \mu_2)^2 \rightarrow x = \dfrac{\mu_1 + \mu_2}{2}$. The decision rule is to select ω_1 if $x < \dfrac{\mu_1 + \mu_2}{2}$, and ω_2 otherwise.

(b)

$$P_e = \frac{1}{2} \int_{-\infty}^{(\mu_1+\mu_2)/2} \mathcal{N}(\mu_2, \sigma^2)\ du + \frac{1}{2} \int_{(\mu_1+\mu_2)/2}^{\infty} \mathcal{N}(\mu_1, \sigma^2)\ du$$

We normalize each of these to obtain:

$$P_e = \frac{1}{2} P(\mathcal{N}(0,1) \leq \frac{\mu_1 - \mu_2}{2\sigma}) + \frac{1}{2} P(\mathcal{N}(0,1) \geq \frac{\mu_2 - \mu_1}{2\sigma})$$

$$= P(\mathcal{N}(0,1) \geq \frac{\mu_2 - \mu_1}{2\sigma})$$

Finally, we plug in the PDF of the standard normal to observe that $P_e = P(\mathcal{N}(0,1) \geq \dfrac{\mu_2 - \mu_1}{2\sigma}) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_a^\infty e^{-z^2/2} dz$, where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.