CSE 5819 Assignment #2

by: Aayushi Verma (uef24001)

This assignment is organized as follows:

1 Part 1: ChatGPT Self-Learning (20pts)

You can start from the following prompts, but you need to create subsequent questions that attempt to understand the basic concepts of machine learning.

1.1 What defines a probability function?

A probability function, denoted as P, is a function that assigns probabilities to events in a sample space S. It must satisfy the following axioms:

1. **Non-negativity**:
$$P(E) \geq 0 \quad \text{for all events } E \subseteq S$$
2. **Normalization**:
$$P(S) = 1$$
3. **Additivity** (for disjoint events $E_1, E_2, \ldots$):
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \quad \text{if } E_i \cap E_j = \emptyset \text{ for } i \neq j$$

1.2 What is Bernoulli distribution?

The Bernoulli distribution is a discrete probability distribution for a random variable X that takes on only two possible outcomes: 1 (success) and 0 (failure). The probability mass function (PMF) is given by:

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

where $0 \leq p \leq 1$ is the probability of success. The Bernoulli distribution has the following properties:

1. **Mean**:
$$\mathbb{E}[X] = p$$
2. **Variance**:
$$\text{Var}(X) = p(1 - p)$$

1.3 What is binomial distribution?

The binomial distribution describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success p. If X is a binomial random variable representing the number of successes in n trials, then the probability mass function (PMF) is given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

where:

- n is the number of trials, - k is the number of successes, - p is the probability of success on a single trial, - $\binom{n}{k}$ is the binomial coefficient, defined as:

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}$$

The binomial distribution has the following properties:

1. **Mean**:
$$\mathbb{E}[X] = np$$
2. **Variance**:
$$\text{Var}(X) = np(1 - p)$$

1.4 What is the relationship between Bernoulli and binomial distributions?

The relationship between the Bernoulli and binomial distributions can be understood as follows:

1. A **Bernoulli distribution** is a special case of the binomial distribution where the number of trials $n = 1$. In other words, a Bernoulli random variable represents the outcome of a single trial that can result in either success (with probability p) or failure (with probability $1 - p$).

2. A **binomial distribution** is the sum of n independent Bernoulli random variables, each with the same probability of success p. Specifically, if $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables, then their sum follows a binomial distribution:

$$X = X_1 + X_2 + \cdots + X_n$$

where X is binomially distributed with parameters n and p, i.e., $X \sim \text{Binomial}(n, p)$.

Thus, a binomial distribution can be viewed as the distribution of the number of successes in n independent Bernoulli trials.

1.5 Properties of the expectation of a random variable X, E[X].

The expectation, or expected value, of a random variable X, denoted by $\mathbb{E}[X]$, has the following properties:

1. **Linearity of Expectation**: For any two random variables X and Y, and constants a and b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

This holds regardless of whether X and Y are independent.

2. **Expectation of a Constant**: For any constant c:

$$\mathbb{E}[c] = c$$

3. **Non-Negativity**: If $X \geq 0$ almost surely (i.e., for all outcomes), then:

$$\mathbb{E}[X] \geq 0$$

4. **Expectation of an Indicator Variable**: If I_A is an indicator random variable for event A (i.e., $I_A = 1$ if A occurs, and $I_A = 0$ otherwise), then:

$$\mathbb{E}[I_A] = P(A)$$

5. **Monotonicity**: If $X \leq Y$ almost surely, then:

$$\mathbb{E}[X] \leq \mathbb{E}[Y]$$

6. **Multiplication by a Constant**: For any random variable X and constant a:

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$

7. **Expectation of the Sum of Random Variables**: For random variables $X_1, X_2, \ldots, X_n$:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

This holds regardless of whether the random variables are independent.

1.6 Is there a difference between Gaussian distribution and Normal distribution?

The terms **Gaussian distribution** and **Normal distribution** refer to the same probability distribution. Both terms describe a continuous probability distribution characterized by its bell-shaped curve. The Gaussian distribution is named after the mathematician Carl Friedrich Gauss, who studied its properties, while the term Normal distribution is used because it arises naturally in many practical situations due to the Central Limit Theorem.

The probability density function (PDF) of a Normal (or Gaussian) distribution with mean μ and standard deviation σ is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where: - μ is the mean of the distribution, - σ is the standard deviation of the distribution.

In summary, there is no difference between the Gaussian distribution and the Normal distribution; they are two names for the same distribution.

1.7 What is the density function of a Gaussian distribution?

The probability density function (PDF) of a Gaussian (Normal) distribution with mean μ and standard deviation σ is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where: - μ is the mean of the distribution, - σ is the standard deviation of the distribution, - x is the variable for which the density is being evaluated, - $\exp(\cdot)$ denotes the exponential function.

1.8 What is the p-norm of a vector?

The p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

where x_i denotes the i-th component of the vector $\mathbf{x}$, and $p \geq 1$ is a parameter that determines the type of norm:

1. **For** $p = 1$ (the Manhattan or L1 norm):

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

2. **For** $p = 2$ (the Euclidean or L2 norm):

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$$

3. **For** $p \to \infty$ (the Infinity or L∞ norm):

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

The p-norm is a generalization of various norms used to measure the magnitude of a vector in different ways.

1.9 What is the formula of p-norm when p = 0 (or p=1, p=∞)?

For the p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

1. **For** $p = 0$: The 0-norm is not a true norm, but it is often used to count the number of non-zero components in the vector. It is defined as:

$$\|\mathbf{x}\|_0 = \text{number of non-zero components of } \mathbf{x}$$

2. **For** $p = 1$ (the Manhattan or L1 norm):

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

3. **For** $p = \infty$ (the Infinity or L∞ norm):

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

Note that the 0-norm is not strictly a norm in the mathematical sense because it does not satisfy all the properties required for a norm, such as positive homogeneity and the triangle inequality.

1.10 Why is any vector norm a convex function?

A vector norm $\|\mathbf{x}\|$ is a convex function because it satisfies the definition of convexity. A function f is convex if for all vectors $\mathbf{x}$ and $\mathbf{y}$, and for any $t \in [0, 1]$, the following inequality holds:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

For a vector norm $\|\cdot\|$, this translates to:

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

To see why this is true, consider the properties of norms:
1. **Triangle Inequality**: For any vectors $\mathbf{x}$ and $\mathbf{y}$, and any $t \in [0, 1]$, the norm satisfies:

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\|$$

2. **Homogeneity**: For any vector $\mathbf{x}$ and scalar $a \geq 0$:

$$\|a\mathbf{x}\| = a\|\mathbf{x}\|$$

Using these properties, we have:

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

This inequality shows that a vector norm satisfies the definition of convexity. Thus, any vector norm is a convex function.

1.11 What is the definition of a convex function?

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if, for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and for any scalar $t \in [0, 1]$, the following inequality holds:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

In other words, the value of the function at any point on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ is less than or equal to the weighted average of the function values at $\mathbf{x}$ and $\mathbf{y}$.

This definition ensures that the function lies below the line segment connecting any two points on its graph, giving it a "bowl-shaped" appearance.

1.12 What is the definition of a convex set?

A set $C \subseteq \mathbb{R}^n$ is called **convex** if, for any two points $\mathbf{x}$ and $\mathbf{y}$ in C, the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ is also entirely contained within C. Formally, C is convex if:

$$\text{For all } \mathbf{x}, \mathbf{y} \in C \text{ and for all } t \in [0, 1], \text{ we have } t\mathbf{x} + (1-t)\mathbf{y} \in C.$$

In other words, if you take any two points in the set C and draw a line segment between them, every point on that line segment will also be in C. This definition ensures that the set does not have any "dips" or "holes" between its points.

1.13 What is the definition of convex optimization?

Convex optimization is a branch of optimization where the goal is to minimize a convex objective function over a convex set. Formally, a convex optimization problem is defined as:

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{x} \in C \end{aligned}$$

where:
- $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, meaning that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and for all $t \in [0, 1]$:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

- $C \subseteq \mathbb{R}^n$ is a convex set, meaning that for all $\mathbf{x}, \mathbf{y} \in C$ and for all $t \in [0, 1]$:

$$t\mathbf{x} + (1 - t)\mathbf{y} \in C$$

In convex optimization, the objective is to find the point $\mathbf{x} \in C$ that minimizes the convex function $f(\mathbf{x})$. This type of optimization problem is particularly well-studied because convex functions have desirable properties, such as having no local minima other than the global minimum, which makes solving these problems more tractable and efficient.

1.14 What is the 0-norm?

The 0-**norm** of a vector $\mathbf{x} \in \mathbb{R}^n$, often denoted as $\|\mathbf{x}\|_0$, is not a true norm in the mathematical sense, but it is widely used in sparse modeling and optimization. The 0-norm counts the number of non-zero elements in the vector. It is defined as:

$$\|\mathbf{x}\|_0 = \text{number of non-zero elements in } \mathbf{x}$$

Formally, it can be expressed as:

$$\|\mathbf{x}\|_0 = \sum_{i=1}^{n} \mathbb{I}(x_i \neq 0)$$

where $\mathbb{I}$ is the indicator function, which is 1 if $x_i \neq 0$ and 0 otherwise.

Note that the 0-norm does not satisfy all properties of a true norm, particularly the triangle inequality, which is why it is considered a "quasi-norm" in practice.

2 Problems

2.1 [Basic Statistics]

1. **[Basic Statistics]** (15 pts) If $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$, $Var[X] = \sigma^2$, and $E[X^2] = \mu^2 + \sigma^2$. Further, recall that expectation is linear, so *it* obeys the following three properties:

   $E[X + c] = E[X] + c$ for any constant c,
   $E[X + Y] = E[X] + E[Y]$,
   $E[aX] = aE[X]$ for any constant a.

   We note that if X and X' are independent, then $E[XX'] = E[X]E[X']$.
   Consider two points (sampled independently) from the same class follow: $X \sim N(\mu_1, \sigma^2)$ and $X' \sim N(\mu_1, \sigma^2)$. What is the expected squared distance between them, i.e., $E[(X - X')^2]$?

We are given $E[(X - X')^2]$. Let's expand this.

$$\begin{aligned} E[(X - X')^2] &= E[(X - X')(X - X')] \\ &= E[X^2 - 2XX' + X'^2] \\ &= E[X^2] - E[2XX'] + E[X'^2] \qquad \text{using linearity property} \end{aligned} \tag{1}$$

We are given $E[XX'] = E[X]E[X']$, given X and X' are independent. Using this, and the second expectation property, we substitute back into Eq. 1:

$$\begin{aligned} E[(X - X')^2] &= E[X^2] - 2E[XX'] + E[X'^2] \\ &= E[X^2] - 2E[X]E[X'] + E[X'^2] \end{aligned} \tag{2}$$

We are also given $E[X^2] = \mu^2 + \sigma^2$, and $E[X] = \mu$. Substituting these back into Eq. 2, we have:

$$E[(X - X')^2] = (\mu_X^2 + \sigma_X^2) - 2\mu_X\mu_{X'} + (\mu_{X'}^2 + \sigma_{X'}^2) \tag{3}$$

Remember that both X and X' are independently sampled from the same class: $X, X' \sim N(\mu, \sigma^2)$. Therefore we let $\mu_X = \mu_{X'}$, and $\sigma_X = \sigma_{X'}$. Then Eq. 3 becomes:

$$\begin{aligned} E[(X - X')^2] &= \mu^2 + \sigma^2 - 2\mu^2 + \mu^2 + \sigma^2 \\ &= 2\mu^2 - 2\mu^2 + 2\sigma^2 \\ &= 2\sigma^2 \end{aligned} \tag{4}$$

$\therefore$ The expected squared distance between $E[(X - X')^2]$ is $2\sigma^2$ (which also makes intuitive sense).

2.2 [Basic Linear Algebra] (15 pts) We are given a vector $x = [0, 0.2, 1.0, 2.2]$. Which of the following vector is closest to x and what is the distance from the closest point to x under each of the following vector norms?

$$x_1 = [0.7, 0.2, 0.5, 2.0]$$
$$x_2 = [0, 1.0, 1.5, 2.2]$$
$$x_3 = [0.8, 0.1, 1.2, 2.0]$$

a) 0-norm = b) 1-norm = c) 2-norm = d) ∞-norm =

2.2.1 For $\mathbf{x}_1$:

$$\begin{aligned} ||\mathbf{x} - \mathbf{x}_1|| &= \begin{bmatrix} |0 - 0.7| & |0.2 - 0.2| & |1 - 0.5| & |2.2 - 2| \end{bmatrix} \\ &= \begin{bmatrix} 0.7 & 0 & 0.5 & 0.2 \end{bmatrix} \end{aligned} \tag{5}$$

$$\Rightarrow ||\mathbf{x}_1||_0 = 3 \tag{6}$$

$$\Rightarrow ||\mathbf{x}_1||_1 = \sum ||\mathbf{x} - \mathbf{x}_1|| = 0.7 + 0 + 0.5 + 0.2 = 1.4 \tag{7}$$

$$\begin{aligned} \Rightarrow ||\mathbf{x}_1||_2 &= \sqrt{|0 - 0.7|^2 + |0.2 - 0.2|^2 + |1 - 0.5|^2 + |2.2 - 2|^2} \\ &= \sqrt{0.7^2 + 0^2 + 0.5^2 + 0.2^2} = \sqrt{0.78} = 0.8832 \end{aligned} \tag{8}$$

$$\Rightarrow ||\mathbf{x}_1||_\infty = max||\mathbf{x} - \mathbf{x}_1|| = 0.7 \tag{9}$$

2.2.2 For $\mathbf{x}_2$:

$$\begin{aligned} ||\mathbf{x} - \mathbf{x}_2|| &= \begin{bmatrix} |0 - 0| & |0.2 - 1| & |1 - 1.5| & |2.2 - 2.2| \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0.8 & 0.5 & 0 \end{bmatrix} \end{aligned} \tag{10}$$

$$\Rightarrow ||\mathbf{x}_2||_0 = 2 \tag{11}$$

$$\Rightarrow ||\mathbf{x}_2||_1 = \sum ||\mathbf{x} - \mathbf{x}_2|| = 0 + 0.8 + 0.5 + 0 = 1.3 \tag{12}$$

$$\begin{aligned} \Rightarrow ||\mathbf{x}_2||_2 &= \sqrt{|0 - 0|^2 + |0.2 - 1|^2 + |1 - 1.5|^2 + |2.2 - 2.2|^2} \\ &= \sqrt{0^2 + 0.8^2 + 0.5^2 + 0^2} = \sqrt{0.89} = 0.9434 \end{aligned} \tag{13}$$

$$\Rightarrow ||\mathbf{x}_2||_\infty = max||\mathbf{x} - \mathbf{x}_2|| = 0.8 \tag{14}$$

2.2.3 For $\mathbf{x}_3$:

$$\begin{aligned} ||\mathbf{x} - \mathbf{x}_3|| &= \begin{bmatrix} |0 - 0.8| & |0.2 - 0.1| & |1 - 1.2| & |2.2 - 2| \end{bmatrix} \\ &= \begin{bmatrix} 0.8 & 0.1 & 0.2 & 0.2 \end{bmatrix} \end{aligned} \tag{15}$$

$$\Rightarrow ||\mathbf{x}_3||_0 = 4 \tag{16}$$

$$\Rightarrow ||\mathbf{x}_3||_1 = \sum ||\mathbf{x} - \mathbf{x}_2|| = 0.8 + 0.1 + 0.2 + 0.2 = 1.3 \tag{17}$$

$$\begin{aligned} \Rightarrow ||\mathbf{x}_3||_2 &= \sqrt{|0 - 0.8|^2 + |0.2 - 0.1|^2 + |1 - 1.2|^2 + |2.2 - 2|^2} \\ &= \sqrt{0.8^2 + 0.1^2 + 0.2^2 + 0.2^2} = \sqrt{0.73} = 0.8544 \end{aligned} \tag{18}$$

$$\Rightarrow ||\mathbf{x}_3||_\infty = max||\mathbf{x} - \mathbf{x}_3|| = 0.8 \tag{19}$$

Based on the above norm calculations, we have the following results:

1. by the 0-norm, $\mathbf{x}_2$ is the closest to $\mathbf{x}$.
2. by the 1-norm, both $\mathbf{x}_2$ and $\mathbf{x}_3$ are the closest to $\mathbf{x}$.
3. by the 2-norm, $\mathbf{x}_3$ is the closest to $\mathbf{x}$.
4. by the ∞-norm, $\mathbf{x}_1$ is the closest to $\mathbf{x}$.

2.3 [Convexity] (1)

3. **[Convexity]** (1) (10 pts) Given a vector space V, any vector norm satisfies the following three properties:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$.	**Non-negativity**
(N2) $\|\lambda x\| = \|\lambda\| \, \|x\|$.	**Scaling**
(N3) $\|x + y\| \leq \|x\| + \|y\|$.	**Triangle Inequality**

If we consider this vector norm as a function of x, $f(x) = ||x||$, prove that $f(x)$ is a convex function (using the above properties). Recall the definition of a convex function is:

A real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is called convex if, for any two points $x, y \in \mathbb{R}^n$ and any $\lambda \in [0, 1]$, the following inequality holds:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

We are given the definition of a convex function:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \tag{20}$$

We start with $f(x) = ||x||$:

$$\begin{aligned} f(x) &= ||x|| \\ &= f(\lambda x + (1 - \lambda)y) \\ &= ||\lambda x + (1 - \lambda)y|| && \text{using Eq. 20} \\ &\leq ||\lambda x|| + ||(1 - \lambda)y|| && \text{using triangle inequality} \\ &= |\lambda|||x|| + |(1 - \lambda)|||y|| && \text{using scaling} \end{aligned} \tag{21}$$

Then using Eq. 21, and substituting back into the convex function (Eq. 20), we have:

$$f(\lambda x + (1 - \lambda)y) = ||\lambda x|| + ||(1 - \lambda)y|| \leq |\lambda|||x|| + |(1 - \lambda)|||y|| \tag{22}$$

$\therefore$ We have shown that the vector norm as a function, $f(x) = ||x||$, is indeed convex, as it satisfies the inequality of the convex function (Eq. 20).

2.4 [Convexity] (2) (10 pts). Prove that the convexity is preserved under a linear transformation. Supposed $f(w)$ is convex in terms of w. Prove that $g(w) = f(Xw + b)$ is also convex in terms of w where X is a fixed matrix of appropriate size, and b is a fixed vector of appropriate size. (So X and b are not variables in g). (Hint: you can simply use the definition of convex function)

Once again, we will use the convex function definition, Eq. 20 from the previous question. We will re-write Eq. 20 in terms of $g(w)$:

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y) \tag{23}$$

Let's examine the RHS of Eq. 23, and substitute $g(w) = f(Xw + b)$:

$$\begin{aligned} \lambda g(x) + (1 - \lambda)g(y) &= \lambda f(Xw_1 + b) + (1 - \lambda)f(Xw_2 + b) \\ &= f(\lambda Xw_1 + (1 - \lambda)Xw_2 + b) \end{aligned} \tag{24}$$

Let's check the convexity of Eq. 24:

$$\begin{aligned} f(\lambda Xw_1 + (1 - \lambda)Xw_2 + b) &\leq \lambda f(Xw_1 + b) + (1 - \lambda)f(Xw_2 + b) \\ &= \lambda g(x) + (1 - \lambda)g(y) \end{aligned} \tag{25}$$

Finally we substitute Eq. 25 into Eq. 24:

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) \tag{26}$$

$\therefore$ Convexity is preserved under the linear transformation $g(w) = f(Xw + b)$.

2.5 [Convexity] (3) (10 pts). From (1) and (2), prove the "loss" function $||y - Xw||_2$ is convex in terms of w where $X_{n\times d}$ is a data matrix containing n examples and each example has d features, and y is a column vector of length d. This dataset has been observed and thus fixed, and w is the only variable in the norm.

Let's represent the loss function by a vector $\mathbf{a}$, where $\mathbf{a} = y - Xw$. Then we have:

$$\begin{aligned} f(x) &= \mathbf{a} \\ &= y - Xw \end{aligned} \tag{27}$$

The norm of this vector $\mathbf{a}$ satisfies the vector norm properties, and is given by:

$$\begin{aligned} ||\mathbf{a}|| &= ||y - Xw|| \\ \Rightarrow ||\mathbf{a}||_2 &= \sqrt{\sum_{i=1}^{n} \mathbf{a}_i^2} \end{aligned} \tag{28}$$

As shown in Eq. 22 in Question 2.3, the vector norm $||\mathbf{a}||$ is indeed convex.

Similarly, let $g(w) = f(y - Xw)$. As shown in Eq. 26, this linear transformation is also convex.

Therefore we have shown that a function in the form of a linear norm ($||y - Xw||_2$) is convex.