Instructions:

- This assignment is meant to help you understand certain concepts we will use in the course.

1. Simple Derivatives

(a) Find the derivative of the sigmoid function with respect to \( x \) where the sigmoid function \( \sigma(x) \) is given by,

\[
\sigma(x) = \frac{1}{1 + e^{-x}}
\]

**Solution:** The derivative of the sigmoid function is as follows:

\[
\sigma'(x) = \frac{d\sigma(x)}{dx}
= d\left(\frac{1}{1 + e^{-x}}\right)
= \frac{d}{dx}(1 + e^{-x})^{-1}
= -(1 + e^{-x})^{-2} \frac{d}{dx}(1 + e^{-x})
= -(1 + e^{-x})^{-2}(-e^{-x})
\]

We can simplify the above answer as follows:

\[
-(1 + e^{-x})^{-2}(-e^{-x}) = \frac{e^{-x}}{(1 + e^{-x})^2}
= \left(\frac{1}{1 + e^{-x}}\right)\left(\frac{e^{-x}}{1 + e^{-x}}\right)
= \left(\frac{1}{1 + e^{-x}}\right)\left(\frac{1 - 1 + e^{-x}}{1 + e^{-x}}\right)
= \left(\frac{1}{1 + e^{-x}}\right)(1 - \frac{1}{1 + e^{-x}})
= \sigma(x)(1 - \sigma(x))
\]

Therefore, the derivative of the sigmoid function is:

\[
\sigma'(x) = \sigma(x)(1 - \sigma(x))
\]
(b) Given two gaussian functions

\[ y = \mathcal{N}(0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]

and

\[ \hat{y} = \mathcal{N}(1, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \]

we define,

\[ \mathcal{L} = (y - \hat{y})^2 \]

Find \( \frac{d\mathcal{L}}{dx} \) at \( x = 1 \).

**Solution:** Given,

\[ \mathcal{L} = (y - \hat{y})^2 = \frac{1}{2\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right)^2 \]

The derivative of \( \mathcal{L} \) w.r.t \( x \) is given by \( \frac{d\mathcal{L}}{dx} = \mathcal{L}' \), which can be found as follows:

\[
\mathcal{L}' = \frac{1}{2\pi} \frac{d}{dx} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right)^2 \\
= \frac{2}{2\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \frac{d}{dx} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \\
= \frac{1}{\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \left( \frac{d}{dx} \left( e^{-\frac{x^2}{2}} \right) - \frac{d}{dx} \left( e^{-\frac{(x-1)^2}{2}} \right) \right) \\
= \frac{1}{\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \left( e^{-\frac{x^2}{2}} \frac{d}{dx} \left( -\frac{x^2}{2} \right) - e^{-\frac{(x-1)^2}{2}} \frac{d}{dx} \left( -\frac{(x-1)^2}{2} \right) \right) \\
= \frac{1}{\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \left( xe^{-\frac{x^2}{2}} - (x-1)e^{-\frac{(x-1)^2}{2}} \right) \\
= \frac{1}{\pi} \left( e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}} \right) \left( xe^{-\frac{x^2}{2}} - (x-1)e^{-\frac{(x-1)^2}{2}} \right)
\]

By substituting \( x = 1 \), we get :

\[
\left. \frac{d\mathcal{L}}{dx} \right|_{x=1} = \frac{-1}{\pi} \left( e^{-\frac{1}{2}} - e^{-\frac{(1-1)^2}{2}} \right) \left( e^{-\frac{1}{2}} - (1-1)e^{-\frac{(1-1)^2}{2}} \right) \\
= \frac{-1}{\pi} \left( e^{-\frac{1}{2}} - (1-1)e^{-\frac{1}{2}} \right)
\]

(c) Find the derivative of \( f(\rho) \) with respect to \( \rho \) where \( f(\rho) \) is given by,

\[ f(\rho) = \rho \log \frac{\rho}{\hat{\rho}} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}} \]

(Hint : You can treat \( \hat{\rho} \) as a constant.)
Solution: The derivative of \( f(\rho) \) with respect to \( \rho \) can be found as follows:

\[
f'(\rho) = \frac{d}{d\rho}(f(\rho)) = \frac{d}{d\rho}\left(\rho \log\left(\frac{\rho}{\hat{\rho}}\right) + (1 - \rho) \log(1 - \rho) \right)
\]

Treating \( \hat{\rho} \) as a constant and using product rule of derivatives, we get,

\[
f'(\rho) = (\rho - \rho \log(\rho)) + \frac{d}{d\rho}\left((1 - \rho) \log(1 - \rho) - (1 - \rho) \log(1 - \hat{\rho})\right)
\]

2. Chain Rule

Using the chain rule of derivatives, find the derivative of \( f(x) \) with respect to \( x \) where

(a) \( f(x) = x \log(3^x) \)

Solution: Let,

\[
z = 3^x
\]

\[
\therefore \frac{dz}{dx} = \frac{d}{dx} 3^x = 3^x \log 3
\]

Also let,

\[
y = \log(z)
\]

\[
\therefore \frac{dy}{dz} = \frac{d}{dz} \log z = \frac{1}{z} = \frac{1}{3^x}
\]

Therefore, we can write \( f(x) \) in terms of \( y \) which itself can be written in terms of \( z \), i.e.,

\[
f(x) = xy
\]
The derivative of $f(x)$ can be found as follows:

$$f'(x) = \frac{d}{dx}(f(x))$$

$$= \frac{d}{dx}(xy)$$

$$= x\frac{dy}{dx} + y\frac{d}{dx}x$$

$$= x\frac{dy}{dz}\frac{dz}{dx} + y$$

(By Product Rule)

$$= x\frac{1}{z}3^x\log 3 + \log 3^x$$

$$= x\log 3 + \log 3^x$$

$$= \log 3^x + \log 3^x$$

$$= 2\log 3^x$$

(b) $f(x) = \sigma(w_1(\sigma(w_0x + b_0)) + b_1)$,
where $w_1, w_0, b_0, b_1$ are constants and $\sigma(x)$ is the sigmoid function defined in Q1(a).

Solution: Using change of variables we can write $f(x)$ as:

$$f(x) = \sigma(w_1(\sigma(w_0x + b_0)) + b_1)$$

$$= \sigma(z)$$

where,

$$z = w_0x + b_0$$

$$\therefore \frac{dz}{dx} = \frac{d}{dx}(w_0x + b_0) = w_0$$

and

$$y = w_1(\sigma(z)) + b_1$$

$$\therefore \frac{dy}{dz} = w_1\frac{d\sigma(z)}{dz} = w_1\sigma(z)(1 - \sigma(z))$$

Therefore, we can write $f(x)$ in terms of $y$ which itself can be written in terms of $z$, i.e.,

$$f(x) = \sigma(y)$$
The derivative of \( f(x) \) can be found as given below. Also, recall from Q1(a), the derivative of \( \sigma(x) \) w.r.t x is given by \( \sigma'(x) = \sigma(x)(1 - \sigma(x)) \).

\[
\begin{align*}
    f(x) &= \sigma(y) \\
    f'(x) &= \frac{d}{dx}\sigma(y) \\
    &= \frac{d}{dy}\sigma(y) \frac{dy}{dx} \tag{By Chain rule} \\
    &= \sigma(y)(1 - \sigma(y)) \frac{dy}{dz} \frac{dz}{dx} \tag{By Chain rule} \\
    &= \sigma(y)(1 - \sigma(y))w_1\sigma(z)(1 - \sigma(z))w_0
\end{align*}
\]

3. Taylor Series

(a) Consider \( x \in \mathbb{R} \) and \( f(x) \in \mathbb{R} \). Write down the Taylor series expansion of \( f(x) \).

**Solution:** A function \( f(x) \) can be expanded around a given point \( x \) by the Taylor Series:

\[
f(x + \delta x) = f(x) + f'(x)(\delta x) + \frac{f''(x)}{2!}(\delta x)^2 + \ldots + \frac{f^{(n)}(x)}{n!}(\delta x)^n + \ldots
\]

where \( \delta x \) is very small, \( f'(x) \) is the first derivative of \( f(x) \) with respect to \( x \) and \( f^{(n)}(x) \) is the \( n^{th} \) derivative of \( f(x) \) with respect to \( x \).

(b) Consider \( x \in \mathbb{R}^n \) and \( f(x) \in \mathbb{R} \). Write down the Taylor series expansion of \( f(x) \).

**Solution:** A function \( f(x) \) where \( x \) is a vector in \( \mathbb{R}^n \), can be expanded by the Taylor series as follows:

\[
f(x + \delta x) = f(x) + \nabla_x f(x)\delta x + \frac{1}{2!}\delta x^T \nabla_x^2 f(x)\delta x + \ldots
\]

where,

\[
\delta x = [\delta x_1, \ldots, \delta x_n]^T \\
\nabla_x f(x) = \begin{bmatrix}
    \frac{\partial f(x)}{\partial x_1} \\
    \vdots \\
    \frac{\partial f(x)}{\partial x_n}
\end{bmatrix}
\]

\[
\nabla_x^2 f(x) = \begin{bmatrix}
    \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\
    \vdots & \vdots & \ddots & \vdots \\
    \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2}
\end{bmatrix}
\]
4. Softmax Function

(a) How is the softmax function defined?

Solution: Softmax function squashes a $K$-dimensional vector $\mathbf{v}$ of arbitrary real values to a $K$-dimensional vector $\text{softmax}(\mathbf{v})$ of real values, where each entry is in the range $(0, 1)$, and all the entries add up to 1. The softmax function is defined as:

$$\text{softmax}(\mathbf{v})_j = \frac{e^{v_j}}{\sum_{k=1}^{K} e^{v_k}} \quad j = 1, 2, \ldots, K$$

For example:
Let $\mathbf{v} = [2.1 \ 4.8 \ 3.5]$, then the softmax of it will be:

$$\begin{align*}
\text{softmax}(\mathbf{v})_1 &= \frac{e^{v_1}}{\sum_{k=1}^{3} e^{v_k}}, \text{ note that here } K = 3 \\
&= \frac{e^{2.1}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.0502 \\
\text{softmax}(\mathbf{v})_2 &= \frac{e^{v_2}}{\sum_{k=1}^{3} e^{v_k}} \\
&= \frac{e^{4.8}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.7464 \\
\text{softmax}(\mathbf{v})_3 &= \frac{e^{v_3}}{\sum_{k=1}^{3} e^{v_k}} \\
&= \frac{e^{3.5}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.2034
\end{align*}$$

Therefore, $\text{softmax}(\mathbf{v}) = [0.0502 \ 0.7464 \ 0.2034]$.

(b) Can you think of any concept which is similar to what the softmax function computes? (Hint: You probably learnt it in high school)

Solution: The output of the softmax function can be used to represent the probability distribution over $K$ components of the input vector.

5. Matrix Multiplication

(a) What are the four ways of multiplying two matrices?

Solution:

1. The most common way of finding the product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is to compute the $ij$-th element of the resultant product matrix $\mathbf{C}$ using the...
\(i^{th}\) row of \(A\) and \(j^{th}\) column of \(B\). For example, suppose matrix \(A\) is of size \(m \times n\) with elements \(a_{ij}\) and a matrix \(B\) of size \(n \times p\) with elements \(b_{jk}\), then multiplying matrices \(A\) and \(B\) will produce matrix \(C\) of size \(m \times p\). The \(ij\)-th element of this matrix will be computed as,

\[
c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}
\]

2. The second way is to realise that the columns of \(C\) are the linear combinations of columns of \(A\). To get the \(i^{th}\) column of \(C\), multiply the whole matrix \(A\) with the \(i^{th}\) column of \(B\). (Remember that a matrix times column is a column.)

Example: Let \(A\) be a \(3 \times 2\) matrix and \(B\) be a \(2 \times 3\) matrix. Then,

\[
C = AB
\]

\[
= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}
\]

\[
= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \end{bmatrix}
\]

3. The third way is to realise that the rows of \(C\) are the linear combinations of rows of \(B\). To get the \(i^{th}\) row of \(C\), multiply the \(i^{th}\) row of \(A\) with the whole matrix \(B\). (Remember that a row times matrix is a row.)
Example: Let $A$ be a $3 \times 2$ matrix and $B$ be a $2 \times 3$ matrix.

$$C = AB$$

$$= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} b_{11} + a_{12} b_{21} & a_{11} b_{12} + a_{12} b_{22} & a_{11} b_{13} + a_{12} b_{23} \\ a_{21} b_{11} + a_{22} b_{21} & a_{21} b_{12} + a_{22} b_{22} & a_{21} b_{13} + a_{22} b_{23} \\ a_{31} b_{11} + a_{32} b_{21} & a_{31} b_{12} + a_{32} b_{22} & a_{31} b_{13} + a_{32} b_{23} \end{bmatrix}$$

4. The fourth way is to look at the product of $AB$ as a sum of (columns of $A$) times (rows of $B$).

Example: Let $A$ be a $3 \times 2$ matrix and $B$ be a $2 \times 3$ matrix. Then,

$$C = AB$$

$$= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} a_{11} b_{11} & a_{11} b_{12} & a_{11} b_{13} \\ a_{21} b_{11} & a_{21} b_{12} & a_{21} b_{13} \\ a_{31} b_{11} & a_{31} b_{12} & a_{31} b_{13} \end{bmatrix} + \begin{bmatrix} a_{12} b_{21} & a_{12} b_{22} & a_{12} b_{23} \\ a_{22} b_{21} & a_{22} b_{22} & a_{22} b_{23} \\ a_{32} b_{21} & a_{32} b_{22} & a_{32} b_{23} \end{bmatrix}$$

(b) Consider a matrix $A$ of size $m \times n$ and a vector $x$ of size $n$. What is the result of
the matrix-vector multiplication $Ax$. Is it a vector or a matrix? What are the dimensions of the product.

**Solution:** It will be a vector of size $m$.

(c) Consider two vectors $x$ and $y \in \mathbb{R}^n$. What is $xy^T$? Is it a matrix of size $n \times n$, a vector of size $n$ or a scalar?

**Solution:** It will be a matrix of size $n \times n$.

6. **L2-norm**

(a) What is meant by L2-norm of a vector?

**Solution:** L2 norm of a vector $v = [v_1, v_2, \ldots, v_n]$ is defined as the square root of the sum of squares of the absolute values of the vector components and is written as,

$$||v||_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2}$$

(b) Given a vector $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \in \mathbb{R}^3$, find it’s L2-norm, i.e. $||v||_2$.

**Solution:** $||v||_2 = \sqrt{v_1^2 + v_2^2 + v_3^2}$

(c) Given a vector $v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{R}^n$, find it’s L2-norm, i.e $||v||_2$.

**Solution:** $||v||_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$

7. **Euclidean Distance**

Consider two vectors $x$ and $y \in \mathbb{R}^n$. How would you compute the Euclidean distance between the two vectors?

**Solution:** Let, $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ be the two vectors. The Euclidean distance,
8. Consider two vectors \( x \) and \( y \in \mathbb{R}^n \). How do you compute the dot product between the two vectors? Is it a matrix of size \( n \times n \), a vector of size \( n \) or a scalar?

**Solution:** Let, \( x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \) and \( y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \) be the two vectors. Then, the dot product between them is defined as follows:

\[
x \cdot y = x^T y = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = \sum_{i=1}^{n} x_i y_i
\]

9. Consider two vectors \( x \) and \( y \in \mathbb{R}^n \). How do you compute the cosine of the angle between the two vectors?

**Solution:** Let, \( x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \) and \( y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \) be the two vectors and \( \theta \) be the angle between them. Then, the cosine of the angle between the two vectors is given by:

\[
\cos \theta = \frac{x \cdot y}{|x||y|}
\]

10. **Basic Geometry**

(a) What is the equation of a line?
Solution: The equation of line can be written as:

\[ y = mx + b \]

Note that it also can be re-written as:

\[ a_1 x_1 + a_2 x_2 = b \]

where, \( x_1 = x, x_2 = y, a_1 = -m, a_2 = 1 \)

(b) What is the equation of a plane in 3 dimensions (assume the axes are \( x_1, x_2, x_3 \))?

Solution: The equation of a plane in 3 dimensions is:

\[ a_1 x_1 + a_2 x_2 + a_3 x_3 = b \]

where, \( x_1, x_2, x_3 \) are the axes and \( a_1, a_2, a_3, b \) are the coefficients.

(c) What is the equation of a plane in \( n \) dimensions (assume the axes are \( x_1, x_2, \ldots, x_n \))?

Solution: The equation of a plane in \( n \) dimensions is:

\[ \sum_{i=1}^{n} a_i x_i = b \]

where, \( x_i \) are the axes and \( a_i, b \) are the coefficients.

11. Basis Consider a set of vectors \( S = \{v_1, v_2, \ldots, v_n\} \in \mathbb{R}^n \). When do you say that these vectors form a basis in \( \mathbb{R}^n \)?

Solution: A set of vectors \( S = \{v_1, v_2, \ldots, v_n\} \in \mathbb{R}^n \) forms a basis in \( \mathbb{R}^n \) if and only if following conditions are satisfied:

1. \( v_1, v_2, \ldots, v_n \) are linearly independent vectors
2. \( S \) spans \( \mathbb{R}^n \) i.e. every vector in \( \mathbb{R}^n \) can be represented as a linear combination of vectors in \( S \).

For example, if \( x \in \mathbb{R}^n \) then we can write,

\[ x = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n \]

where \( v_i \in S \) form the basis of \( \mathbb{R}^n \) and \( c_i \) are co-efficients, \( \forall i \in \{1, 2, \ldots, n\} \).
For example:

The unit basis vectors for $\mathbb{R}^3$ are $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Note that you can represent any vector $\mathbf{v} \in \mathbb{R}^3$ as the linear combination of these three basis vectors.

12. **Orthogonal Vectors**

(a) When are two vectors $\mathbf{u}$ and $\mathbf{v} \in \mathbb{R}^n$ said to be orthogonal?

**Solution:** Two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be orthogonal vectors when their dot-product is zero i.e. $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = 0$.

(b) Are the following vectors orthogonal to each other?

$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

**Solution:** From part (a) of this question, we know that two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be orthogonal if their dot product is zero. Therefore, to check whether $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are orthogonal, we have to find the dot product between them. We do this by taking two vectors at a time.

$\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_1^T \mathbf{v}_2$

$$= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= 0$$

$\mathbf{v}_2 \cdot \mathbf{v}_3 = \mathbf{v}_2^T \mathbf{v}_3$

$$= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= 0$$

$\mathbf{v}_1 \cdot \mathbf{v}_3 = \mathbf{v}_1^T \mathbf{v}_3$

$$= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= 0$$
As we can see, we can take any subset of the above 3 vectors and compute the dot product and the result will be zero. Therefore, \( \mathbf{v}_1, \mathbf{v}_2 \) and \( \mathbf{v}_3 \) are orthogonal to each other.

13. Consider two vectors \( \mathbf{a} \) and \( \mathbf{b} \in \mathbb{R}^n \). What is the vector projection of \( \mathbf{b} \) onto \( \mathbf{a} \)?

**Solution:** The vector projection of \( \mathbf{b} \) onto \( \mathbf{a} \) will have the same direction as vector \( \mathbf{a} \) but it will be either a scaled up or down version of \( \mathbf{a} \) depending on the vector \( \mathbf{b} \). The vector projection of \( \mathbf{b} \) onto \( \mathbf{a} \) is given by,

\[
\left( \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||^2} \right) \mathbf{a} = \left( \frac{\mathbf{a}^T \mathbf{b}}{||\mathbf{a}||^2} \right) \mathbf{a}
\]

14. Consider a matrix \( \mathbf{A} \) and a vector \( \mathbf{x} \). We say that \( \mathbf{x} \) is an eigen vector of \( \mathbf{A} \) if ______ ?

**Solution:** \( \mathbf{x} \) is an eigenvector of \( \mathbf{A} \) if \( \mathbf{A} \mathbf{x} = \lambda \mathbf{x} \) where \( \lambda \) is a scalar and is called the corresponding eigenvalue.

15. Consider a set of vectors \( x_1, x_2, \ldots, x_n \in \mathbb{R}^n \). We say that \( x_1, x_2, \ldots, x_n \) form an orthonormal basis in \( \mathbb{R}^n \) if ______ ?

**Solution:** \( \{x_1, x_2, \ldots, x_n\} \) form an orthonormal basis in \( \mathbb{R}^n \) if \( \{x_1, x_2, \ldots, x_n\} \) are orthogonal to each other and have unit length.

16. Consider a set of vectors \( x_1, x_2, \ldots, x_n \in \mathbb{R}^n \). We say that \( x_1, x_2, \ldots, x_n \) are linearly independent if ______ ?

**Solution:** We say that \( x_1, x_2, \ldots, x_n \) are linearly independent if any vector in the set cannot be written as a linear combination of the remaining vectors in the set. On the other hand, a vector \( x_i \) is said to be linearly dependent on vectors \( x_1 \) to \( x_n \) if it can be written as a linear combination of these vectors as:

\[
c_1 x_1 + \ldots + c_{i-1} x_{i-1} + c_{i+1} x_{i+1} + \ldots + c_n x_n = x_i
\]

\[
\Rightarrow c_1 x_1 + \ldots + c_{i-1} x_{i-1} + c_{i+1} x_{i+1} + \ldots + c_n x_n + (-1) \cdot x_i = 0
\]

\[
\Rightarrow \sum_{k=1}^{n} c_k x_k = 0, \text{ where } c_i = -1
\]
But for a set of linearly independent vectors no vector in the set can be written as a linear combination of the remaining vectors in the set. An alternate way of saying this is that, a set of vectors is linearly independent if the only solution to the equation
\[ \sum_{k=1}^{n} c_k x_k = 0, \text{ is } \forall k = \{1, 2, \ldots, n\} \]

17. Consider a vector \( \mathbf{x} \in \mathbb{R}^n \) and a matrix \( \mathbf{A} \in \mathbb{R}^{n \times n} \). The product \( \mathbf{x}^T \mathbf{A} \mathbf{x} \) can be written as \( \sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ji} x_j \)?

**Solution:** \( \sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ji} x_j \)

18. **KL Divergence**

(a) Consider a discrete random variable \( X \) which can take one of \( k \) values from the set \( \{x_1, \ldots, x_k\} \). A distribution over \( X \) defines the value of \( Pr(X = x) \) \( \forall x \in \{x_1, \ldots, x_n\} \). Consider two such distributions \( P \) and \( Q \). How do you compute the KL divergence between \( P \) and \( Q \).

**Solution:** The KL Divergence between two distributions \( P \) and \( Q \) can be calculated as:

\[
D_{KL}(P||Q) = - \sum_x P(x) \log \frac{Q(x)}{P(x)} = \sum_x P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}_{X \sim P} \left[ \log \frac{P(x)}{Q(x)} \right]
\]

For example, consider a discrete random variable \( X \) which can take one of 3 values from the set \( \{x_1, x_2, x_3\} \). A distribution over \( X \) defines the value of \( Pr(X = x) \) \( \forall x \in \{x_1, x_2, x_3\} \). Consider two such distributions \( P \) and \( Q \) which are defined as follows:

\[
P = \begin{bmatrix} 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}
\]

\[
Q = \begin{bmatrix} 0.228 & 0.619 & 0.153 \\ 0 & 0 & 1 \end{bmatrix}
\]
Then, the KL divergence between $P$ and $Q$ can be calculated as:

$$D_{KL}(P||Q) = (0.0 \times \log \left( \frac{0}{0.228} \right) + 1.0 \times \log \left( \frac{1}{0.619} \right) + 0.0 \times \log \left( \frac{0}{0.153} \right))$$

$$= 0.691$$

(b) Is KL Divergence symmetric?

**Solution:** KL divergence is not symmetric as $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, which can be shown as follows:

$$D_{KL}(Q||P) = -\sum_x Q(x) \log \frac{P(x)}{Q(x)}$$

$$= \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

$$= \mathbb{E}_{X \sim Q} \left[ \log \frac{Q(x)}{P(x)} \right]$$

$$\neq D_{KL}(P||Q)$$

19. Cross Entropy

Given two distributions $P$ and $Q$ defined over a discrete random variable $X$, how do you compute the cross entropy between the two distributions?

**Solution:** The cross entropy between two distributions $P$ and $Q$ is given by,

$$H(P, Q) = -\sum_x P(x) \log Q(x)$$

For example,

Consider a discrete random variable $X$ which can take one of 3 values from the set $\{x_1, x_2, x_3\}$. A distribution over $X$ defines the value of $Pr(X = x) \forall x \in \{x_1, x_2, x_3\}$. Consider two such distributions $P$ and $Q$ which are defined as follows:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ Pr(X = x_1) & Pr(X = x_2) & Pr(X = x_3) \end{bmatrix}$$

$$Q = \begin{bmatrix} 0.228 & 0.619 & 0.153 \\ Pr(X = x_1) & Pr(X = x_2) & Pr(X = x_3) \end{bmatrix}$$

Then, the cross-entropy between $P$ and $Q$ can be calculated as:

$$H(P, Q) = -(0.0 \times \log(0.228) + 1.0 \times \log(0.619) + 0.0 \times \log(0.153))$$

$$= 0.691$$