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:

$$\begin{aligned}
\sigma'(x) &= \frac{d\sigma(x)}{dx} \\
&= \frac{d}{dx}(\frac{1}{1+e^{-x}}) \\
&= \frac{d}{dx}(1+e^{-x})^{-1} \\
&= -(1+e^{-x})^{-2}\frac{d}{dx}(1+e^{-x}) \\
&= -(1+e^{-x})^{-2}(-e^{-x})
\end{aligned}$$

We can simplify the above answer as follows :

$$\begin{aligned}
-(1+e^{-x})^{-2}(-e^{-x}) &= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{1-1+e^{-x}}{1+e^{-x}}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(1-\frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}$$

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,

$$\begin{aligned}\mathcal{L} &= (y-\hat{y})^2 \\ &= \frac{1}{2\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)^2\end{aligned}$$

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:

$$\begin{aligned}\mathcal{L}' &= \frac{1}{2\pi}\frac{d}{dx}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)^2 \\ &= \frac{2}{2\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)\frac{d}{dx}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big) \\ &= \frac{1}{\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)\Big(\frac{d}{dx}\big(e^{-\frac{x^2}{2}}\big) - \frac{d}{dx}\big(e^{-\frac{(x-1)^2}{2}}\big)\Big) \\ &= \frac{1}{\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)\Big(e^{-\frac{x^2}{2}}\frac{d}{dx}\big(-\frac{x^2}{2}\big) - e^{-\frac{(x-1)^2}{2}}\frac{d}{dx}\big(-\frac{(x-1)^2}{2}\big)\Big) \\ &= \frac{1}{\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)\big(e^{-\frac{x^2}{2}}(-x) - e^{-\frac{(x-1)^2}{2}}(-(x-1))\big) \\ &= \frac{-1}{\pi}\big(e^{-\frac{x^2}{2}} - e^{-\frac{(x-1)^2}{2}}\big)\big(xe^{-\frac{x^2}{2}} - (x-1)e^{-\frac{(x-1)^2}{2}}\big)\end{aligned}$$

By substituting $x = 1$, we get :

$$\begin{aligned}\frac{d\mathcal{L}}{dx}\bigg|_{x=1} &= \frac{-1}{\pi}\big(e^{-\frac{1}{2}} - e^{-\frac{(1-1)^2}{2}}\big)\big(e^{-\frac{1}{2}} - (1-1)e^{-\frac{(1-1)^2}{2}}\big) \\ &= \frac{-1}{\pi}\big(e^{-\frac{1}{2}} - 1\big)\big(e^{-\frac{1}{2}}\big)\end{aligned}$$

(c) Find the derivative of $f(\rho)$ with respect to ρ 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 ρ can be found as follows:

$$
\begin{aligned}
f'(\rho) &= \frac{d}{d\rho}(f(\rho)) \\
&= \frac{d}{d\rho}\Big(\rho log(\frac{\rho}{\hat{\rho}}) + (1-\rho)log(\frac{1-\rho}{1-\hat{\rho}})\Big) \\
&= \frac{d}{d\rho}\Big(\rho log(\rho) - \rho log(\hat{\rho}) + (1-\rho)log(1-\rho) - (1-\rho)log(1-\hat{\rho})\Big) \\
&= \frac{d}{d\rho}(\rho log(\rho)) - \frac{d}{d\rho}(\rho log(\hat{\rho})) + \frac{d}{d\rho}((1-\rho)log(1-\rho)) - \frac{d}{d\rho}((1-\rho)log(1-\hat{\rho}))
\end{aligned}
$$

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

$$
\begin{aligned}
f'(\rho) &= (\rho.\frac{1}{\rho} + log(\rho)(1)) - log(\hat{\rho})(1) + ((1-\rho).\frac{-1}{(1-\rho)} + log(1-\rho)(-1)) - log(1-\hat{\rho})(-1) \\
&= 1 + log(\rho) - log(\hat{\rho}) - 1 - log(1-\rho) + log(1-\hat{\rho}) \\
&= log(\frac{\rho}{\hat{\rho}}) - log(\frac{1-\rho}{1-\hat{\rho}}) \\
&= log(\frac{\rho(1-\hat{\rho})}{\hat{\rho}(1-\rho)})
\end{aligned}
$$

2. **Chain Rule**

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

(a) $f(x) = xlog(3^x)$

Solution: Let,

$$
\begin{aligned}
z &= 3^x \\
\therefore \frac{dz}{dx} &= \frac{d}{dx}3^x = 3^x log3
\end{aligned}
$$

Also let,

$$
\begin{aligned}
y &= log(z) \\
\therefore \frac{dy}{dz} &= \frac{d}{dz}logz = \frac{1}{z} = \frac{1}{3^x}
\end{aligned}
$$

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:

$$\begin{aligned}
f'(x) &= \frac{d}{dx}(f(x)) \\
&= \frac{d}{dx}(xy) \\
&= x\frac{dy}{dx} + y\frac{d}{dx}x && \text{(By Product Rule)} \\
&= x\frac{dy}{dz}\frac{dz}{dx} + y && \text{(By Chain Rule)} \\
&= x\frac{1}{3^x}3^x log3 + log3^x \\
&= xlog3 + log3^x \\
&= log3^x + log3^x \\
&= 2log3^x
\end{aligned}$$

(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:

$$f(x) = \sigma(w_1\underbrace{(\sigma(\underbrace{w_0x + b_0}_{\text{= z}})) + b_1}_{\text{= y}})$$

where,

$$\begin{aligned}
z &= w_0x + b_0 \\
\therefore \frac{dz}{dx} &= \frac{d}{dx}(w_0x + b_0) = w_0
\end{aligned}$$

and

$$\begin{aligned}
y &= w_1(\sigma(z)) + b_1 \\
\therefore \frac{dy}{dz} &= w_1\frac{d\sigma(z)}{dz} = w_1\sigma(z)(1 - \sigma(z))
\end{aligned}$$

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 $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

$$\begin{aligned}
f(x) &= \sigma(y) \\
f'(x) &= \frac{d}{dx}\sigma(y) \\
&= \frac{d}{dy}\sigma(y)\frac{dy}{dx} && \text{(By Chain rule)} \\
&= \sigma(y)(1-\sigma(y))\frac{dy}{dz}\frac{dz}{dx} && \text{(By Chain rule)} \\
&= \sigma(y)(1-\sigma(y))w_1\sigma(z)(1-\sigma(z))w_0
\end{aligned}$$

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 :

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

where δx 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 $\mathbf{x} \in \mathbb{R}^n$ and $f(\mathbf{x}) \in \mathbb{R}$. Write down the Taylor series expansion of $f(x)$.

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

$$f(\mathbf{x} + \delta\mathbf{x}) = f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})\delta\mathbf{x} + \frac{1}{2!}\delta\mathbf{x}^T\nabla^2_{\mathbf{x}} f(\mathbf{x})\delta\mathbf{x} + \ldots$$

where,

$$\begin{aligned}
\delta\mathbf{x} &= [\delta x_1, \ldots, \delta x_n]^T \\
\nabla_{\mathbf{x}} f(\mathbf{x}) &= \begin{bmatrix} \frac{\partial \mathbf{x}}{x_1} \\ \vdots \\ \frac{\partial \mathbf{x}}{x_n} \end{bmatrix} \\
\nabla^2_{\mathbf{x}} f(\mathbf{x}) &= \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \ldots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \ldots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}
\end{aligned}$$

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 $\mathbf{softmax}(\mathbf{v})$ 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:
>
> $$softmax(v)_j = \frac{e^{v_j}}{\sum_{k=1}^{K} e^{v_k}} \qquad j = 1, 2, \dots, K$$
>
> For example :
> Let $\mathbf{v} = [2.1 \;\; 4.8 \;\; 3.5]$, then the softmax of it will be:
>
> $$\begin{aligned} softmax(v)_1 &= \frac{e^{v_1}}{\sum_{k=1}^{3} e^{v_k}}, \text{ note that here } K = 3 \\ &= \frac{e^{2.1}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.0502 \\ softmax(v)_2 &= \frac{e^{v_2}}{\sum_{k=1}^{3} e^{v_k}} \\ &= \frac{e^{4.8}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.7464 \\ softmax(v)_3 &= \frac{e^{v_3}}{\sum_{k=1}^{3} e^{v_k}} \\ &= \frac{e^{3.5}}{e^{2.1} + e^{4.8} + e^{3.5}} = 0.2034 \end{aligned}$$
>
> Therefore, $softmax(\mathbf{v}) = [0.09 \; 0.2447 \; 0.6652]$

(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 $\mathbf{A}$ and j^{th} column of $\mathbf{B}$. For example, suppose matrix $\mathbf{A}$ is of size $m \times n$ with elements a_{ij} and a matrix $\mathbf{B}$ of size $n \times p$ with elements b_{jk}, then multiplying matrices $\mathbf{A}$ and $\mathbf{B}$ will produce matrix $\mathbf{C}$ 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 $\mathbf{C}$ are the linear combinations of columns of $\mathbf{A}$. To get the i^{th} column of $\mathbf{C}$, multiply the whole matrix $\mathbf{A}$ with the i^{th} column of $\mathbf{B}$. (Remember that a matrix times column is a column.)
Example: Let $\mathbf{A}$ be a 3×2 matrix and $\mathbf{B}$ be a 2×3 matrix. Then,

$$\begin{aligned}
\mathbf{C} &= \mathbf{AB} \\
&= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\
&= \begin{bmatrix} \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix}}_{1^{st} \text{ column of } \mathbf{C}} & \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix}}_{2^{nd} \text{ column of } \mathbf{C}} & \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{13} \\ b_{23} \end{bmatrix}}_{3^{rd} \text{ column of } \mathbf{C}} \end{bmatrix} \\
&= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \end{bmatrix}
\end{aligned}$$

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

Example: Let $\mathbf{A}$ be a 3×2 matrix and $\mathbf{B}$ be a 2×3 matrix.

$$
\begin{aligned}
\mathbf{C} &= \mathbf{AB} \\
&= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\
&= \begin{bmatrix} \left.\begin{bmatrix} a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \right\} 1^{st} \text{ row of } \mathbf{C} \\ \\ \left.\begin{bmatrix} a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \right\} 2^{nd} \text{ row of } \mathbf{C} \\ \\ \left.\begin{bmatrix} a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \right\} 3^{rd} row of \mathbf{C} \end{bmatrix} \\
&= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \end{bmatrix}
\end{aligned}
$$

4. The fourth way is to look at the product of $\mathbf{AB}$ as a sum of (columns of A) times (rows of B).
   Example: Let $\mathbf{A}$ be a 3×2 matrix and $\mathbf{B}$ be a 2×3 matrix. Then,

$$
\begin{aligned}
\mathbf{C} &= \mathbf{AB} \\
&= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}}_{1^{st} \text{ column of A}} \underbrace{\begin{bmatrix} b_{11} & b_{12} & b_{13} \end{bmatrix}}_{1^{st} \text{ row of B}} + \underbrace{\begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}}_{2^{nd} \text{ column of A}} \underbrace{\begin{bmatrix} b_{21} & b_{22} & b_{23} \end{bmatrix}}_{2^{nd} \text{ row of B}} \\
&= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} \\ a_{31}b_{11} & a_{31}b_{12} & a_{31}b_{13} \end{bmatrix} + \begin{bmatrix} a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} \\ a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} \\ a_{32}b_{21} & a_{32}b_{22} & a_{32}b_{23} \end{bmatrix} \\
&= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \end{bmatrix}
\end{aligned}
$$

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

Solution: It will be a vector of size m.

(c) Consider two vectors $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$. What is $\mathbf{xy^T}$? 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 $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is defined as the square root of the sum of squares of the absolute values of the vector components and is written as,

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

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

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

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

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

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,

d, between the two vectors can then be calculated as:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

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:

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= \mathbf{x}^T \mathbf{y} \\ &= x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \\ &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

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 θ be the angle between them. Then, the cosine of the angle between the two vectors is given by:

$$cos\ \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

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:

$$a_1x_1 + a_2x_2 = b$$

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:

$$a_1x_1 + a_2x_2 + a_3x_3 = b$$

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 $\mathbf{x} \in \mathbb{R}^n$ then we can write,

   $$x = c_1v_1 + c_2v_2 + \ldots + c_nv_n$$

   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. $\mathbf{u} \cdot \mathbf{v} = \mathbf{u^T v} = \mathbf{0}$.

(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.

$$\begin{aligned}\mathbf{v_1} \cdot \mathbf{v_2} &= \mathbf{v_1^T v_2}\\ &= \begin{bmatrix}1 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\\ &= 0\end{aligned}$$

$$\begin{aligned}\mathbf{v_2} \cdot \mathbf{v_3} &= \mathbf{v_2^T v_3}\\ &= \begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}0\\0\\1\end{bmatrix}\\ &= 0\end{aligned}$$

$$\begin{aligned}\mathbf{v_1} \cdot \mathbf{v_3} &= \mathbf{v_1^T v_3}\\ &= \begin{bmatrix}1 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\0\\1\end{bmatrix}\\ &= 0\end{aligned}$$

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 a and $b \in \mathbb{R}^n$. What is the vector projection of b onto a ?

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

$$(\frac{a \cdot b}{||a||^2}) \cdot a = (\frac{a^T b}{||a||^2}) \cdot a$$

14. Consider a matrix A and a vector x. We say that x is an eigen vector of A if _____ ?

Solution: x is an eigenvector of A if $Ax = \lambda x$ where λ 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 :

$$\begin{aligned} c_1x_1 + \ldots + c_{i-1}x_{i-1} + c_{i+1}x_{i+1} + \ldots c_nx_n &= x_i \\ \implies c_1x_1 + \ldots + c_{i-1}x_{i-1} + c_{i+1}x_{i+1} + \ldots c_nx_n + (-1)x_i &= 0 \\ \implies \sum_{k=1}^{n} c_kx_k &= 0, \text{where } c_i = -1 \end{aligned}$$

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

$$\sum_{k=1}^{n} c_k x_k = 0, \text{is}, \;\; c_k = 0 \;\forall k = \{1, 2, \ldots, n\}$$

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

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

18. **KL Divergence**

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

Solution: The KL Divergence between two distributions P and Q can be calculated as :

$$\begin{aligned} D_{KL}(P||Q) &= -\sum_x P(x) \log \frac{Q(x)}{P(x)} \\ &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= \mathbb{E}_{X \sim P}\Big[\log \frac{P(x)}{Q(x)}\Big] \end{aligned}$$

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

$$P = \Big[\underbrace{0}_{\Pr(X=x_1)} \quad \underbrace{1}_{\Pr(X=x_2)} \quad \underbrace{0}_{\Pr(X=x_3)}\Big]$$

$$Q = \Big[\underbrace{0.228}_{\Pr(X=x_1)} \quad \underbrace{0.619}_{\Pr(X=x_2)} \quad \underbrace{0.153}_{\Pr(X=x_1)}\Big]$$

Then, the KL divergence between P and Q can be calculated as:

$$\begin{aligned} D_{KL}(P||Q) &= (0.0 * log\Big(\frac{0}{0.228}\Big) + 1.0 * log\Big(\frac{1}{0.619}\Big) + 0.0 * log\Big(\frac{0}{0.153}\Big)) \\ &= 0.691 \end{aligned}$$

(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:

$$\begin{aligned} D_{KL}(Q||P) &= -\sum_x Q(x)\log\frac{P(x)}{Q(x)} \\ &= \sum_x Q(x)\log\frac{Q(x)}{P(x)} \\ &= \mathbb{E}_{X\sim Q}\Big[\log\frac{Q(x)}{P(x)}\Big] \\ &\neq D_{KL}(P||Q) \end{aligned}$$

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

$$P = \Big[\underbrace{0}_{\Pr(X=x_1)} \quad \underbrace{1}_{\Pr(X=x_2)} \quad \underbrace{0}_{\Pr(X=x_3)}\Big]$$
$$Q = \Big[\underbrace{0.228}_{\Pr(X=x_1)} \quad \underbrace{0.619}_{\Pr(X=x_2)} \quad \underbrace{0.153}_{\Pr(X=x_1)}\Big]$$

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

$$\begin{aligned} H(P,Q) &= -(0.0 * \log(0.228) + 1.0 * \log(0.619) + 0.0 * \log(0.153)) \\ &= 0.691 \end{aligned}$$