The goal of this problem is to formulate and solve a classification problem, in which we would like to know whether an individual is diabetic (Class 1) or non-diabetic (Class 0). Diabetes is one of the most commonly diseases in the U.S. Among the main factors causing diabetes, two main characteristics include:

1. Elevated body mass index (BMI), noted as variable d_1
2. Family history of diabetes, noted as variable d_2

For this reason, we have obtained data from three individuals, "samples" as they are called in machine learning, for which we have measured their BMI and the number of relatives suffering from diabetes.

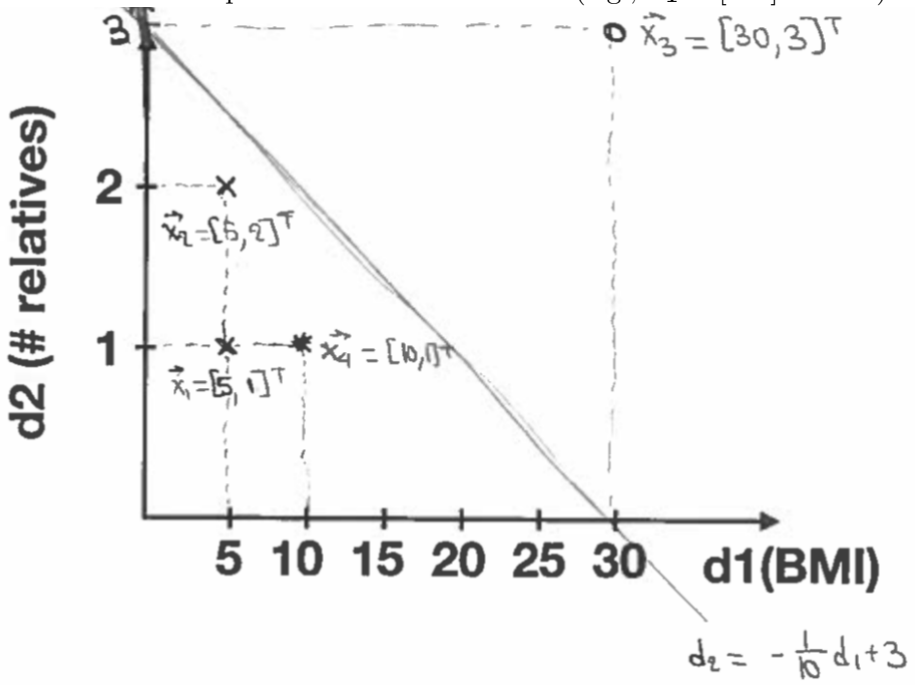
1. Sample 1 ($\mathbf{x_1}$) is non-diabetic, has a BMI of 5 and 1 relative with diabetes
2. Sample 2 ($\mathbf{x_2}$) is non-diabetic, has a BMI of 5 and 2 relatives with diabetes
3. Sample 3 ($\mathbf{x_3}$) is diabetic, has a BMI of 30 and 3 relatives with diabetes

In this problem, we are going to see:

- How to express the aforementioned information in a vector notation (Part 1)
- How to obtain a mathematical expression to classify between the diabetics and non-diabetics (Part 2)
- How to obtain a decision for an "unknown" sample, for which we do not whether he/she is diabetic or not (Parts 3 and 4)

Part 1: Forming a 2D vector space from the data

Consider the above data in a 2D space, where the horizontal axis d_1 is BMI and vertical axis d_2 is the number of relatives with diabetes. Please plot the three data points $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{x_3}$ and write the three data points in a vector notation (e.g., $\mathbf{x_1} = [?\ \ ?] \in \mathbb{R}^{2\times 1}$).



Part 2: Forming a linear decision boundary

We would like to find a line that separates diabetic from non-diabetic samples. This is called "decision boundary," and since we are interested in a line, we are looking at a "linear decision boundary."

Consider that the line passing from $[30, 0]$ and $[0, 3]$ as a decision boundary for classifying between diabetic and non-diabetic individuals.

(2.a) Please plot the line in the above figure and find its equation.
Hint: You will need to find the equation of the line in terms of d_1 and d_2.

$$d_2 = \underbrace{-\frac{1}{10}}_{slope} d_1 + \underbrace{3}_{y-intercept}$$

(2.b) Using the above line, form a classification function, i.e., $f(d_1, d_2)$ which will decide whether a new sample $\mathbf{x} = [d_1, d_2]^T$ belongs to Class 0 (non-diabetic) or Class 1 (diabetic).
Hint: You will need to express f with a rule-based equation (e.g., if $f(d_1, d_2) > 0$ then Class 1, else ...)

A sample $\mathbf{x} = [d_1, d_2]^T$ belongs to Class 1, if it lies on the right side of the line $d_2 = -\frac{1}{10}d_1 + 3$, therefore it meets the condition $d_2 > -\frac{1}{10}d_1 + 3 \;\; \Rightarrow -3 + \frac{1}{10}d_1 + d_2 > 0$.

Similarly, a sample $\mathbf{x} = [d_1, d_2]^T$ belongs to Class 0, if it lies on the left side of the line $d_2 = -\frac{1}{10}d_1 + 3$, therefore it meets the condition $d_2 < -\frac{1}{10}d_1 + 3 \;\; \Rightarrow -3 + \frac{1}{10}d_1 + d_2 < 0$.

Therefore, the classification function should be:

$$\begin{cases} \text{if } f(d_1, d_2) = -3 + \frac{1}{10}d_1 + d_2 > 0 \text{ , then } \mathbf{x} = [d_1, d_2] \in \text{ Class 1} \\ \text{if } f(d_1, d_2) = -3 + \frac{1}{10}d_1 + d_2 < 0 \text{ , then } \mathbf{x} = [d_1, d_2] \in \text{ Class 0} \end{cases}$$

(2.c) Express the three given data points in a matrix notation, such that $\mathbf{D} = \begin{bmatrix} 1 & \mathbf{x_1} \\ 1 & \mathbf{x_2} \\ 1 & \mathbf{x_3} \end{bmatrix} \in \mathbb{R}^{3\times 3}$, by substituting the values of $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{x_3}$ in the above equation. Multiply $\mathbf{D}$ with the weight $\mathbf{w} = [w_0, w_1, w_3]^T = [-3, \frac{1}{10}, 1]^T \in \mathbb{R}^{3\times 1}$. What is the final dimension of $\mathbf{D} \cdot \mathbf{w}$? What do you observe based on question **(2.b)**?

$$\mathbf{D} = \begin{bmatrix} 1 & \mathbf{x_1} \\ 1 & \mathbf{x_2} \\ 1 & \mathbf{x_3} \end{bmatrix} = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 30 & 3 \end{bmatrix}$$

$$\mathbf{D} \cdot \mathbf{w} = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 30 & 3 \end{bmatrix} \cdot \begin{bmatrix} -3 \\ \frac{1}{10} \\ 1 \end{bmatrix} = \begin{bmatrix} -3 + \frac{1}{2} + 1 \\ -3 + \frac{1}{2} + 2 \\ -3 + 3 + 3 \end{bmatrix} = \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{2} \\ 3 \end{bmatrix} \begin{matrix} < 0 \\ < 0 \\ > 0 \end{matrix}$$

We observe that $\mathbf{Dw}$ is the vector notation of the decision rule from question (2.b), since for a given $\mathbf{x} = [d_1, d_2]^T$ we have:

$$\underbrace{\begin{bmatrix} 1 & d_1 & d_2 \end{bmatrix}}_{[1\ \mathbf{x}]} \cdot \underbrace{\begin{bmatrix} -3 \\ \frac{1}{10} \\ 1 \end{bmatrix}}_{\mathbf{w}} = -3 + \tfrac{1}{10}d_1 + d_2 = f(d_1, d_2)$$

(2.d) Now assume that you would like to find whether a person $\mathbf{x_4}$ whose BMI is 10 and has 1 relative with diabetes, is a diabetic or not. How would you write $\mathbf{x_4}$ in a vector notation? How would you classify this person based on the rule in question **(2.b)**?

$\mathbf{x_4} = [10, 1]^T$

$f(\mathbf{x_4}) = -3 + \frac{1}{10} + 1 = -1 < 0$

Therefore, $\mathbf{x_4}$ belongs to Class 0 (non-diabetic).

Part 3: 1-Nearest Neighbor with Euclidean distance

(3.a) Compute the Euclidean distance ($l2-$norm) between sample $\mathbf{x_4}$ and samples $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$.
Hint: If $\mathbf{a} = [a_1, a_2]^T$ and $\mathbf{b} = [b_1, b_2]^T$, then $l_2(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{(|a_1 - b_1|^2 + |a_2 - b_2|^2)}$.

$\mathbf{x_1} = [5, 1]^T$, $\mathbf{x_2} = [5, 2]^T$, $\mathbf{x_3} = [30, 3]^T$, $\mathbf{x_4} = [10, 1]^T$

$d(\mathbf{x_1}, \mathbf{x_4}) = \sqrt{(5-10)^2 + (1-1)^2} = 5$

$d(\mathbf{x_2}, \mathbf{x_4}) = \sqrt{(5-10)^2 + (2-1)^2} = \sqrt{26}$

$d(\mathbf{x_1}, \mathbf{x_4}) = \sqrt{(30-10)^2 + (3-1)^2} = \sqrt{404}$

(3.b) Which sample is closest to $\mathbf{x_4}$ based on the Euclidean distance?

$d(\mathbf{x_1}, \mathbf{x_4}) < d(\mathbf{x_2}, \mathbf{x_4}) < d(\mathbf{x_3}, \mathbf{x_4})$

Therefore, $\mathbf{x_4}$ is closest to $\mathbf{x_1}$ based on the Euclidean distance.

(3.c) If we were to classify $\mathbf{x_4}$ based on the class of its closest sample according to the Euclidean distance, would it come from a diabetic or a non-diabetic person?

Since $\mathbf{x_1}$ is non-diabetic and $\mathbf{x_4}$ is closest to $\mathbf{x_1}$ based on the Euclidean distance, $\mathbf{x_4}$ is *non-diabetic*.

Part 4: 1-Nearest Neighbor with cosine similarity

(4.a) Compute the cosine similarity between sample $\mathbf{x_4}$ and samples $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$.
Hint: If $\mathbf{a} = [a_1, a_2]^T$ and $\mathbf{b} = [b_1, b_2]^T$, then $cos(\mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a},\mathbf{b})}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} = \frac{a_1 \cdot b_1 + a_2 \cdot b_2}{\sqrt{(|a_1|^2+|a_2|^2)}\sqrt{(|b_1|^2+|b_2|^2)}}$.

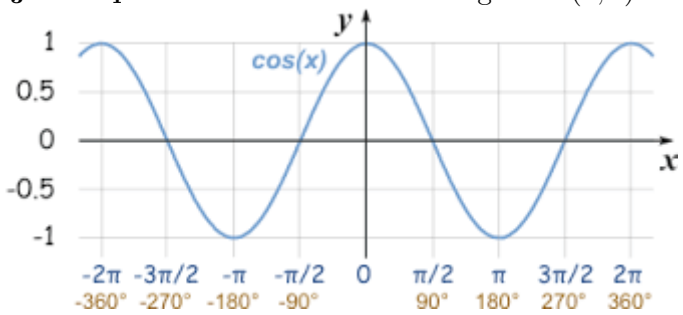
$\mathbf{x_1} = [5, 1]^T$, $\mathbf{x_2} = [5, 2]^T$, $\mathbf{x_3} = [30, 3]^T$, $\mathbf{x_4} = [10, 1]^T$

$cos(\mathbf{x_1}, \mathbf{x_4}) = \frac{5\times10+1\times1}{\sqrt{(5^2+1^2)(10^2+1^2)}} \approx 0.99 \Rightarrow \theta(\mathbf{x_1}, \mathbf{x_4}) \approx 0.14\pi \approx 44^o$

$cos(\mathbf{x_2}, \mathbf{x_4}) = \frac{5\times10+2\times1}{\sqrt{(5^2+2^2)(10^2+1^2)}} \approx 0.96 \Rightarrow \theta(\mathbf{x_2}, \mathbf{x_4}) \approx 0.28\pi \approx 88^o$

$cos(\mathbf{x_3}, \mathbf{x_4}) = \frac{30\times10+3\times1}{\sqrt{(30^2+3^2)(10^2+1^2)}} = 1 \Rightarrow \theta(\mathbf{x_2}, \mathbf{x_4}) = 0 \approx 0^o$

In fact, we see that $\mathbf{x_3}$ and $\mathbf{x_4}$ are in the same line starting from $(0, 0)$.



(4.b) Which sample is closest to $\mathbf{x_4}$ based on the cosine distance?

Based on the above, $\mathbf{x_4}$ forms a 0^o angle with $\mathbf{x_3}$, while it forms a 44^o and 88^o angle with $\mathbf{x_1}$ and $\mathbf{x_2}$, respectively. Therefore, based on the cosine similarity $\mathbf{x_4}$ is closest to $\mathbf{x_3}$.

(4.c) If we were to classify $\mathbf{x_4}$ based on the class of its closest sample according to the cosine distance, would it come from a diabetic or a non-diabetic person? What do you observe compared to question **(3.c)**?

Sample $\mathbf{x_3}$ belongs to the diabetic class, therefore based on the cosine similarity, sample $\mathbf{x_4}$ is *diabetic*. This contradicts the decision based on the Euclidean distance (Question 3.c), which indicates that the *distance metric matters* when comparing samples in the $\mathbb{R}^N$ space.