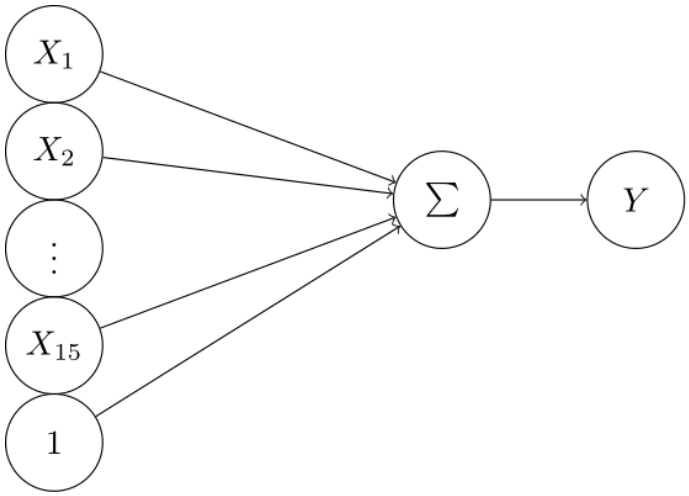
**Pregunta 3** **[20 Puntos]**

# 1. Optimización en Machine Learning

## 1.1. Descripción del problema

La empresa Cabo de Hornos, es de las principales productoras de almohadas de alta gama. La empresa le vende a cada cliente una almohada con cierta cantidad de relleno (Y) en gramos de plumas. Cada persona, por sus características físicas, tiene sólo una almohada que es máximamente cómoda por su cantidad de relleno. La empresa gasta muchos recursos en tiempo cuando los clientes se devuelven a cambiar el relleno de la almohada, que se debe a que la cantidad de relleno que se le recomendó al cliente inicialmente no era totalmente cómodo. Es por lo anterior que Cabo de Hornos ha decidido contratarlo a usted para que con sus conocimientos de optimización elabore y resuelva un modelo matemático que permita entregarle a sus clientes una recomendación óptima de relleno de almohada basado en ciertas mediciones de sus dimensiones corporales.
Basado en lo anterior, la empresa ha decidido hacer una encuesta pagada a 100 de sus clientes que están completamente satisfechos con su almohada. Se les pide que tomen quince medidas de sus características corporales, las cuales son: peso ($x_1$), estatura ($x_2$), edad ($x_3$), contorno de la cabeza ($x_4$), distancia hombro-cuello ($x_5$), altura del cuello($x_5$),.. etc. También se les pide que indiquen los gramos de plumas de su almohada ($y$).
El objetivo de este ejercicio de machine learning es construir una función Y = f($X$) que permita entregar una recomendación Y para cada persona basado en sus medidas corporales. Para esto se entrenará una red neuronal del tipo regresión lineal usando la información de las encuestas. El modelo es el siguiente:




Este diagrama del modelo a utilizar explica cómo es el recorrido de transformaciones para pasar desde los inputs X al output Y. Todos y cada uno de los arcos (menos el final) tienen asociados unos valores denominados pesos. Llámese $w_i$ a los pesos de los arcos que conectan los inputs $x_i$ con el nodo de la sumatoria. Llámese además $b$ al peso que conecta la constante 1 a la sumatoria final. Para elaborar la función deseada se debe conseguir el objetivo de asignarle valores a todos los pesos descritos anteriormente. Pero no pueden ser cualquier valor, pues debieran representar de buena manera la función que conecta X con Y. Si, por ejemplo, todos los pesos tuvieran valor cero, el modelo siempre predeciría erróneamente que la almohada óptima es de Y = 0 gramos, para cualquier persona de características X. Esto es un problema porque va en contradicción con los datos recopilados, ya que ningún relleno de almohada en la encuesta es de valor cero, sino que son mayores. La idea es encontrar pesos que minimicen la diferencia entre lo predicho por el modelo para una entrada X y el verdadero output Y de esa entrada X.

## 1.2. Formulación Matemática 1

El primer problema busca encontrar los pesos $w_i$ y la constante $b$ del modelo a entrenar descrito anteriormente. Los parámetros de este problema serán los valores X e Y, los cuales se obtienen de la encuesta. Sea $r \in \{1, ..., 100\}$ el subíndice para cada una de las muestras de clientes que respondieron la encuesta.

Se deberá resolver el siguiente problema de optimización no lineal irrestricto para encontrar los pesos y constante óptimos:

$$\min_{w,b} \quad \sum_{r=1}^{100}(b + \sum_{i=1}^{15} w_i \cdot x_{i,r} - y_r)^2 \tag{1}$$

## Parte 1

Se le pide que elabore un código main1.py y explicite en su consola los valores de cada uno de los pesos $w_i$ y $b$ para cada uno de las características. Debe mencionar el nombre de la característica y su respectivo ponderador en orden. Recuerde también imprimir en consola el valor del término $b$ (es una constante en gramos).

## 1.3. Formulación Matemática 2

Después de resolver el modelo anterior, la empresa se dio cuenta de que era mucho tomarle 15 medidas a cada cliente para determinar la almohada óptima a usar. Pero... ¿Qué características usar? ¿Cuáles son más importantes? Está claro que los pesos $w_i$ tienen relación con la respuesta. Se propone agregar un factor $\lambda$ que penalice la existencia de valores $w_i$ (ya sea positivos o negativos). Se definen variables $z_i$ que representen la norma 1 (valor absoluto) de las variables $w_i$.

$$\min_{w,z,b} \quad \sum_{r=1}^{100}(b + \sum_{i=1}^{15} w_i \cdot x_{i,r} - y_r)^2 + \lambda \cdot \sum_{i=1}^{15} z_i \tag{2}$$

$$-z_i \leq w_i \quad \forall i \in \{1, ..., 15\} \tag{3}$$

$$w_i \leq z_i \quad \forall i \in \{1, ..., 15\} \tag{4}$$

## Parte 2

Se le pide que elabore un código main2.py en que implemente la formulación 2. El parámetro $\lambda$ es un regularizador en la regresión lineal. Éste permite castigar la magnitud de los valores de $w_i$ si es que no está realizando cambios significativos en minimizar el error cuadrático, y si son fácilmete reemplazables por los otros pesos. Note que si $\lambda = 0$, no hay castigo y si $\lambda = \infty$ todos los $w_i$=0. En su código debe ir aumentando progresivamente los valores de $\lambda$, y viendo los valores de los pesos $w_i$, hasta que sean distintos de cero exactamente 5 de éstos últimos. Esos representarán las características más significativas que determinan el relleno óptimo en una regresión lineal, según los datos. Determine el conjunto $E \subseteq \{1, ..., 15\}$ de 5 características más importantes y que su código lo imprima en pantalla al correr. Imprima en pantalla el valor $\lambda$ utilizado para encontrar $E$. Considere que, en Python, el cero a veces se expresa como una potencia de 10 de exponente muy negativo (-11 o incluso menor).

### Parte 3

Se le pide que elabore un código main3.py en que implemente de nuevo la formulación 1 del problema (sin el $\lambda$ ni $z$). Sin embargo, esta vez, debe modificar la formulación para incluir solamente aquellas características que resultaron ser mejores predictoras en el modelo, y que estén en el conjunto $E$ de la parte 2. Note que el resultado de la parte 3 es la que realmente le importa a la empresa Cabo de Hornos para predecir el relleno óptimo. Después de resolver el modelo, su código debe imprimir en la consola un string que represente el modelo final de la regresión que usted le entregará a la empresa.
Debe imprimirse siguiendo un formato:

$$Y = \sum_{i \in E} w_i \cdot x_i + b \tag{5}$$

Y mencionando qué representa cada uno de los 5 $x_i$, qué representa la variable Y, y la unidad de medida en que están cada uno.

## 1.4. Base de datos:

Junto a la tarea se encontrará un archivo.csv que tiene los datos recopilados en la encuesta. El formato de los datos es tal que cada fila representa la información de un usuario y en la primera fila está la información de lo que representa cada columna. Ese es el archivo con que se verificará su código.