

9 10 10 4



NOMBRE y DNI: JULIETA MERYNANDA / 93456832

PROBLEMA 1:

Dado el siguiente conjunto de datos aplicar el algoritmo de clustering *kmeans* con k=2.

- Utilizar como centroides iniciales los primeros dos datos.
- Indicar los centros de cluster encontrados, etiquetar los datos y calcular la *distancia introcluster*.

	x1	x2
a	6	9
b	2	6
c	1	0
d	8	5
e	0	4

PROBLEMA 2:

Grafe las rectas de discriminación de los resultados que dieron los cromosomas de la siguiente población de un Algoritmo Genético que entrenó un perceptrón simple. Se codificaron los parámetros del perceptrón como $[w_1 \ w_2 \ b]$ y la función de evaluación intentó minimizar el error de clasificación.

Indique si alguno obtuvo una solución exitosa.

x_1	x_2	CLASE
-2	1	A
0	2	A
1	-1	B
2	0	B
-1	-2	B

$$\text{POBLACIÓN} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 3 & 2 \\ -2 & 5 & 0 \end{bmatrix}$$

PROBLEMA 3:

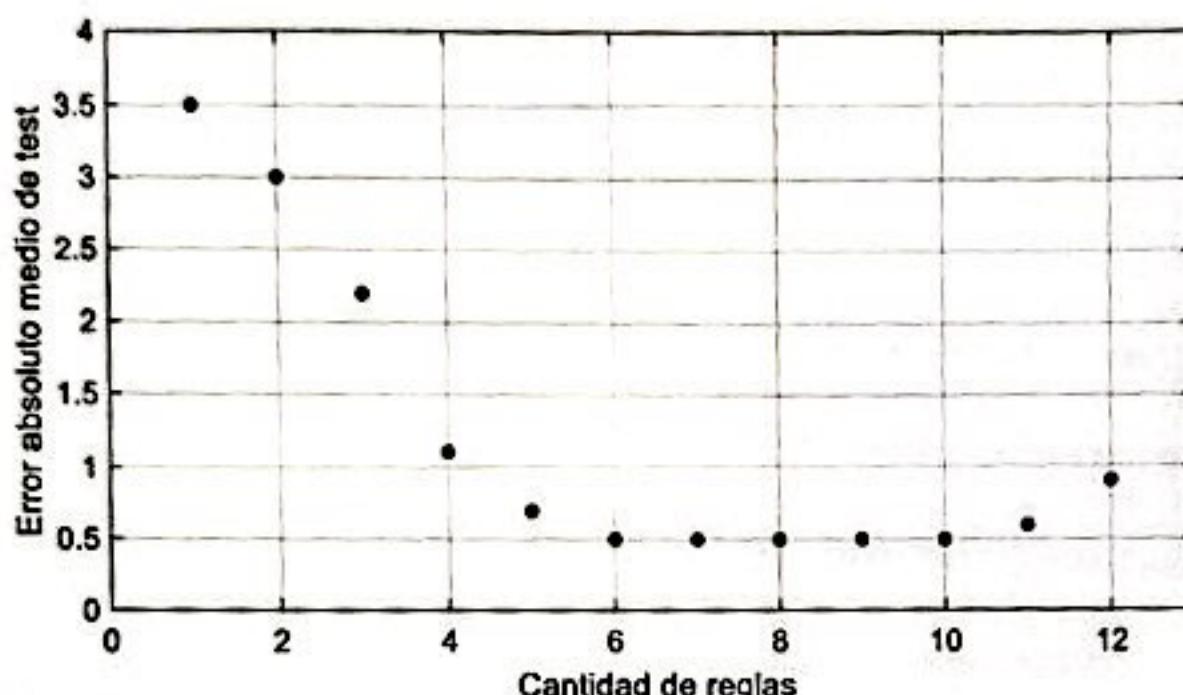
Se intenta diseñar un modelo de Sugeno para estimar un valor de alarma de enfermedad renal entre 0 y 10 (KDA, *kidney disease alarm*), dadas dos variables: la tasa de filtración glomerular (GFR, *glomerular filtration rate*) y la cantidad de albúmina en sangre (ALB).

Para generar el modelo se dispone de 160 casos en los que se conoce GFR, ALB y su respectivo valor KDA.

Se ha decidido utilizar el 5% para datos de test y el resto para datos de "entrenamiento" (generación del modelo). Se utiliza el error absoluto medio para evaluar el comportamiento del modelo. Se utiliza el método de hold-out para la evaluación con 20 iteraciones.

a) Indique las dimensiones de las matrices que contienen los datos para entrenamiento y los datos para test.

b) Se varió la cantidad de reglas de diversos modelos y se observa lo siguiente:



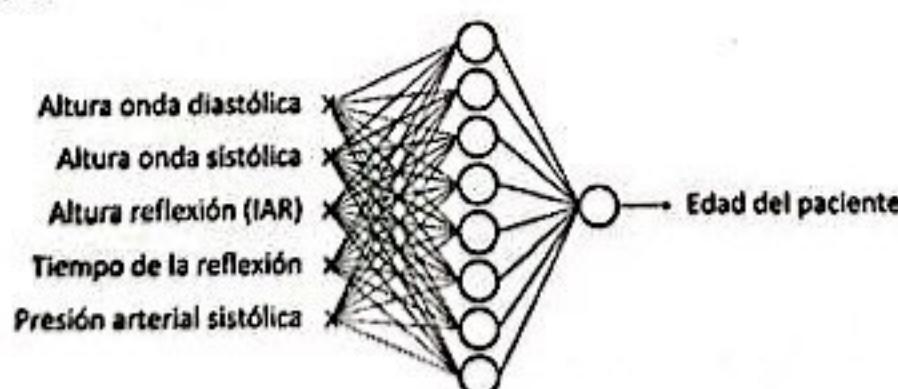
Seleccione la cantidad de reglas que estima adecuada para elegir uno de los modelos. Explique cuál puede ser la causa del aumento del error en la última parte.

c) Calcule el error para los siguientes 3 datos, según lo que estimó el modelo elegido:

GFR	ALB	KDA estimado	KDA real
110	24	1.2	0.5
12	52	9.7	9.4
55	31	2.2	3.0

PROBLEMA 4:

Se tiene la siguiente red neuronal:



- Indique la dimensión de la matriz de datos de entrenamiento, si se dispone de 1000 datos.
- ¿Qué es el error de entrenamiento? Se entrenó la red y se obtiene un error de entrenamiento demasiado alto. Indique qué puede cambiar de la red para mejorar los resultados. ¿Un error de entrenamiento bajo asegura un error de test bajo?
- Indique una fórmula para conocer cuántos pesos se intentan optimizar durante el entrenamiento en la red como la mostrada, en función de la cantidad de entradas y de la arquitectura dada. Utilice la notación que prefiera.

①

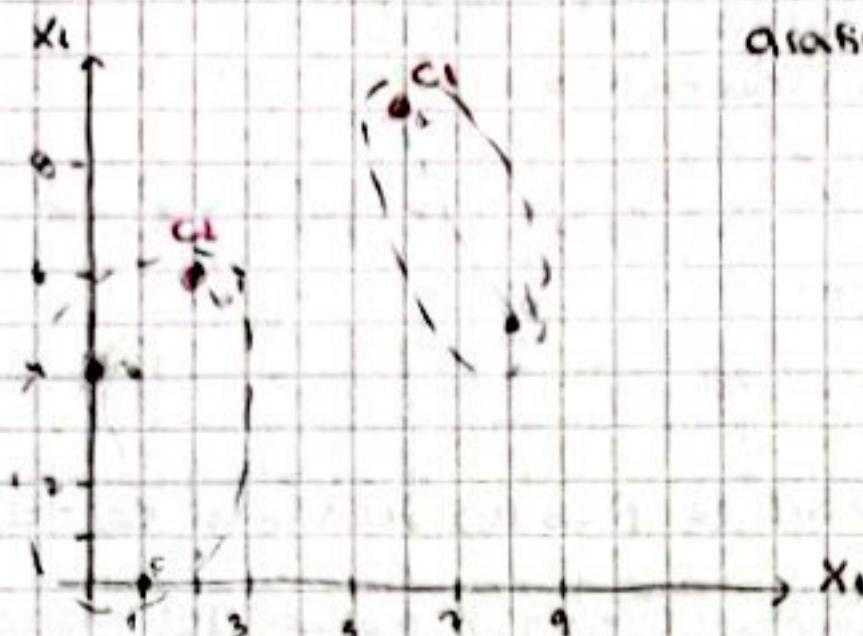


grafico (x_1, x_2) inicialmente.



El primer paso del algoritmo de k-means es la asignación de los clusters.

El segundo paso es la asignación de cada punto a un centro de cluster.

El tercero paso es recalcular los centros centroides de clusters; para ello, busco minimizar la distancia intracluster y maximizar la distancia intercluster.

USO LA DISTANCIA EUCLIDIANA: $d(x_i, x_{j,i}) =$

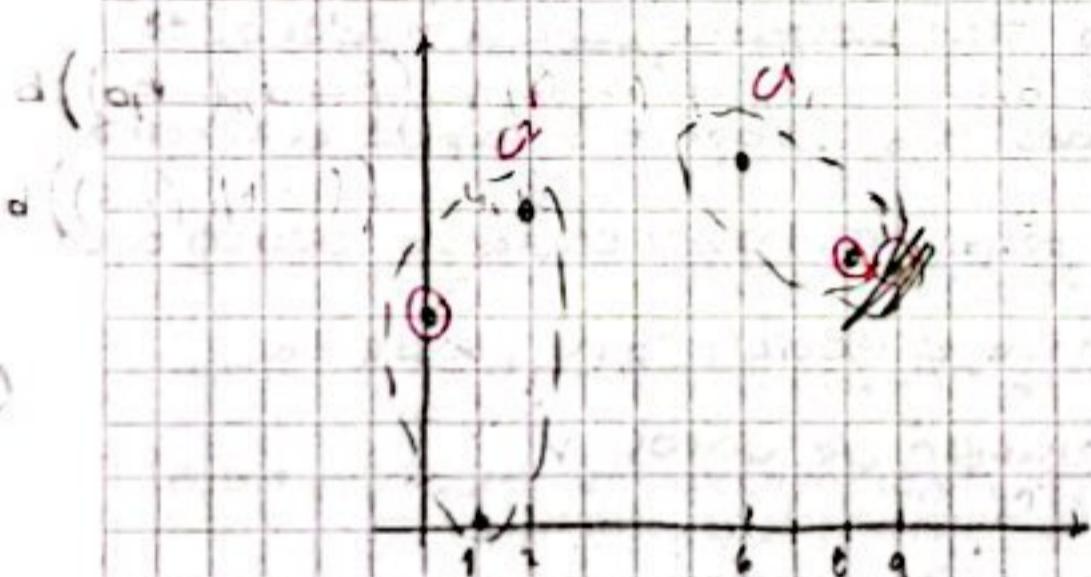


grafico (x_1, x_2) final

x_1	x_2	k
6	9	2
2	6	2
1	0	2
8	5	1
0	4	2



$$d_{\text{intracluster}} = \sum_{k=1}^2 \left[\sum_{\text{dato} \in C_k} d_{\text{eucl}}(\text{dato}, c_k(k, i)) \right] = \underbrace{(2.83 + 4.1)}_{\text{para } C_2} + \underbrace{(4.47)}_{\text{para el } C_1} = 11.54$$

se debe minimizar la distancia.

Cohiciente

NOTA

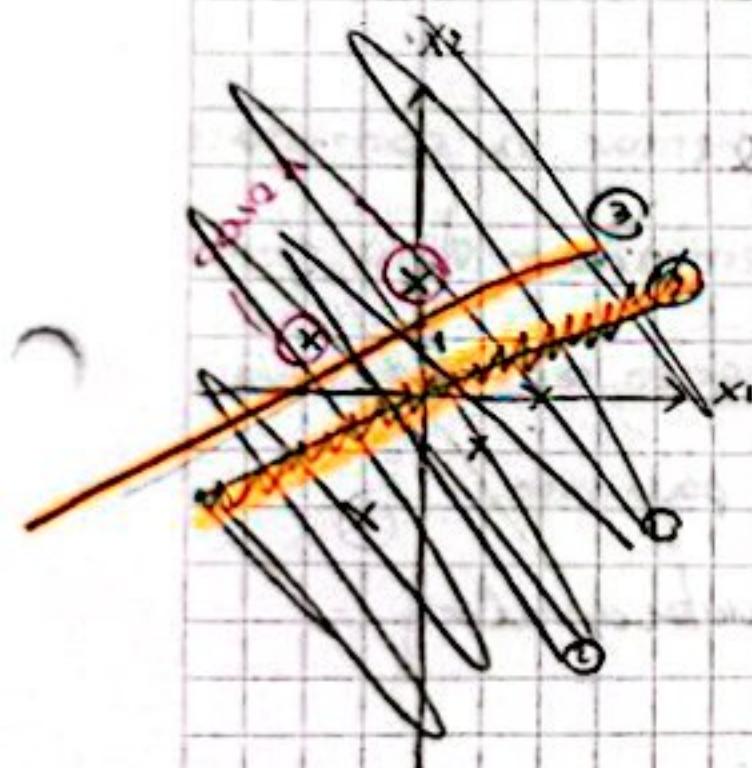
NOTA

C_1
 C_2

$$② Y = w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0$$

Población: $\begin{bmatrix} w_1 & w_2 & b \\ 1 & 2 & 3 \\ 3 & 3 & 2 \\ -2 & 5 & 0 \end{bmatrix}$

Para el cálculo de los rectos se debe, con la función $X_2 = -\frac{w_1 \cdot x_1 + b}{w_2}$, reemplazar en ella con todos los valores de (x_1, x_2) utilizando cada $[w_1 \ w_2 \ b]$ de la matriz de población. Así se obtienen los 3 rectos. De ellos, se debe seleccionar aquella que separa correctamente ambas clases.

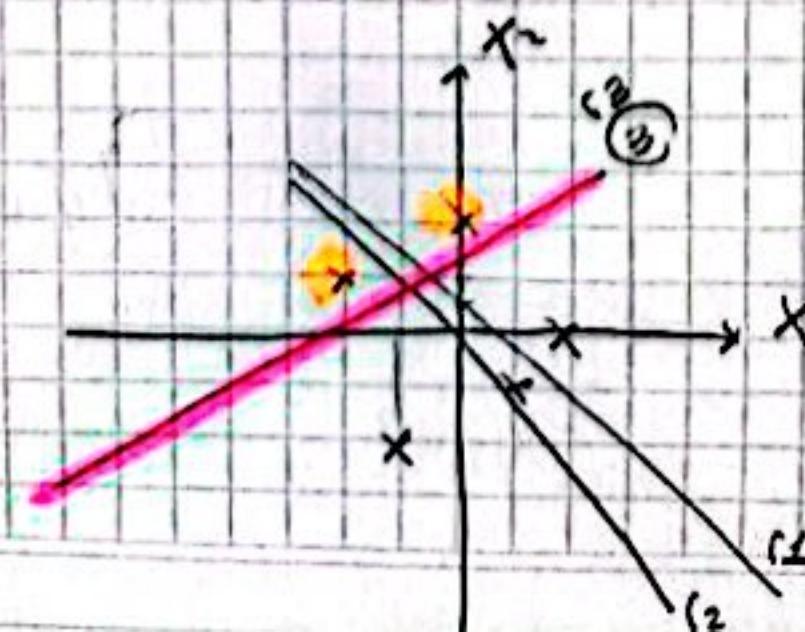


Los rectos 1 y 2 no separan linealmente los conjuntos, mientras que el recto 3 sí los separa, siendo ésta la condición necesaria para ser tomado como la recta de discriminación.

$$\text{Población 1: } \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} -2 & 0 & 1 & 2 & -1 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix} \Rightarrow X_2 = -\frac{1}{2}x_1 + \frac{3}{2}$$

$$\text{Población 2: } \begin{bmatrix} 3 & 3 & 2 \end{bmatrix} \begin{bmatrix} \text{idem} \end{bmatrix} \Rightarrow X_2 = x_1 + \frac{2}{3}$$

$$\text{Población 3: } \begin{bmatrix} -2 & 5 & 0 \end{bmatrix} \begin{bmatrix} \text{idem} \end{bmatrix} \Rightarrow X_2 = -\frac{2}{5}x_1$$



③ 160 casos \rightarrow 5% \Rightarrow 8 ✓, mae , variables $\xrightarrow{\text{GFR}}$ kDD .
 95% \Rightarrow 152 ✓, mae , variables $\xrightarrow{\Delta \text{LB}}$

a) Dimension de la matriz de entrenamiento.

152 filas x 3 columnas. ✓

Dimension de la matriz de test:

8 filas x 3 columnas. ✓

Para ambas matrices, las cantidad de filas representa la cantidad de casos que se utilizaron para la generalización y prueba del modelo. Mencionar que la cantidad de columnas representan los 2 inputs y el único output que tiene ese sistema. ✓

b). La cantidad adecuada de reglas se determina a partir de la que presente el menor error absoluto medio. En este caso, para el rango de [6, 10] reglas el error es el mismo, conociendo cualquier cantidad de reglas perteneciente a ese intervalo sería adecuado para el modelo. En mi opinión, escogería 6 reglas para hacer más eficiente y rápido el sistema. El aumento del error en la última parte se puede haber causado debido a un overtraining - Aquí se complejiza tanto el modelo que pierde la capacidad de predecir la interpretación de datos. ✓

c) error: $|T(x_i) - \text{valor obtenido}|$

$$\text{mae} = \frac{1}{n} \sum_{i=1}^n |T(x_i) - y_i|$$

kDD estimado	kDD real	error	mae
1.2	0.5	0.7	✓
4.7	9.4	0.3	
2.2	3.0	0.8	0.6

4) dimensión de los datos de entrenamiento:

1000 x 6 ✓

↳ siendo 5 inputs y un output,

↳ siendo mil los datos.

b) El error de entrenamiento es el error sufrido cuando hay discrepancias entre los predicciones del modelo y los valores utilizados en el conjunto de entrenamiento. ✓

- Para mejorar los resultados una opción podría ser incrementar la cantidad de datos suministrados al sistema para que tenga un entrenamiento más preciso para cada caso en específico. También se pueden agregar más variables de entrada que determinen la edad de un paciente.

c).
↳ no es la red lo que está cambiando.
↳ no responde.