Evaluación

1. Proporciona una matriz de confusión para el problema de predecir si la inversión de una determinada empresa debe ser “alta”, “baja” o “media” de modo que el Sensibilidad de la clase “alta” sea del 75%, el número de Falsos Positivos (FP) de la clase “baja” sea 50. Calcula la Especificidad de la clase “media”. NOTA: En la matriz de confusión no puede haber 0’s.

   SOLUCIÓN: Para este ejercicio hay múltiples soluciones, una posible es esta. En negrita están los valores según las restricciones del enunciado, el resto son valores libres para que no haya celdas a cero.

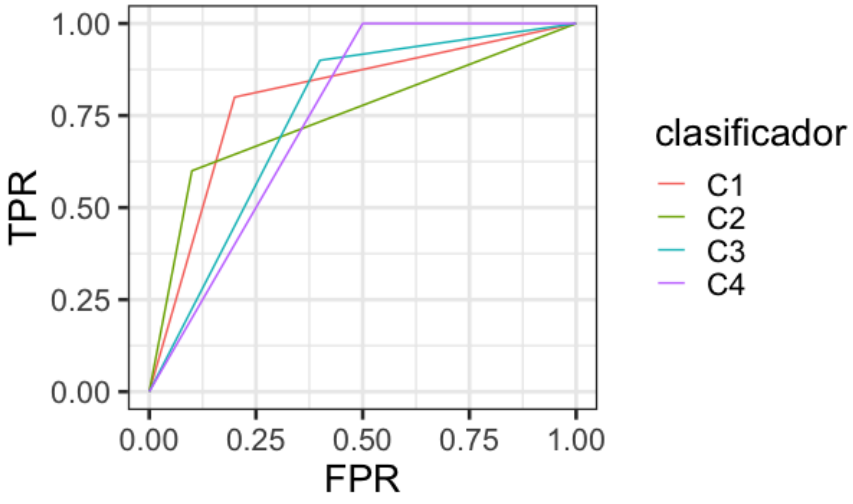
		Real		
		alta	media	baja
Pred.	alta	**75**	3	11
	media	**10**	27	5
	baja	**15**	**35**	16

   Especificidad(media) = (75+11+15+16)/ (75+11+15+16+10+5)=0.886

2. ¿Qué clasificador escogerías y porqué en base a los siguientes resultados?

	C1	C2	C3	C4
Sensibilidad	0.8	0.6	0.9	1.0
Especificidad	0.8	0.9	0.6	0.5

SOLUCIÓN: Nos basamos en la curva ROC, el mejor clasificador será el que tenga su coordenada (FPR,TFP) más ceca del punto (0,1) o, de otra forma, el que deje una mayor área bajo su curva. Por tanto, el mejor clasificador es C1.

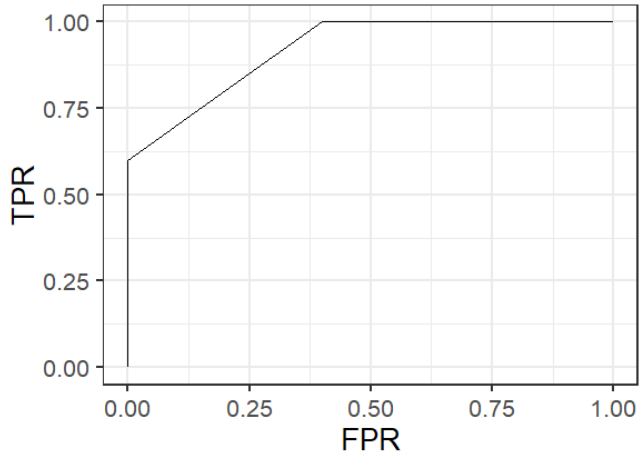


3. Dado un clasificador con las siguientes predicciones:

Real	pos	neg	pos	pos	neg	neg	neg	pos	neg	pos
Score(pos)	0.6	0.1	0.4	0.8	0.6	0.3	0.4	0.7	0.2	0.7

Obtén la curva ROC. ¿Se podría decir que es un buen clasificador?

SOLUCIÓN: Coordenadas FPR=(0.0,0.0,0.0,0.2,0.4,0.6,0.8,1.0); TPR=(0.0,0.2,0.6,0.8,1.0,1.0,1.0,1.0)



4. Supongamos que tenemos 50 libros de minería de datos de un total de 200 libros en una biblioteca. Si un clasificador predice que 10 libros son de minería de datos, pero solo 5 de ellos lo son realmente, ¿Cuál es la Sensibilidad y a Especificidad?

SOLUCIÓN: Sensibilidad=5/50=0.1; Especificidad=(150-5)/150=0.9667

Árboles de decisión

5. Genera un árbol de decisión para las siguientes funciones booleanas:

   a. $X_1 \wedge \neg X_2$

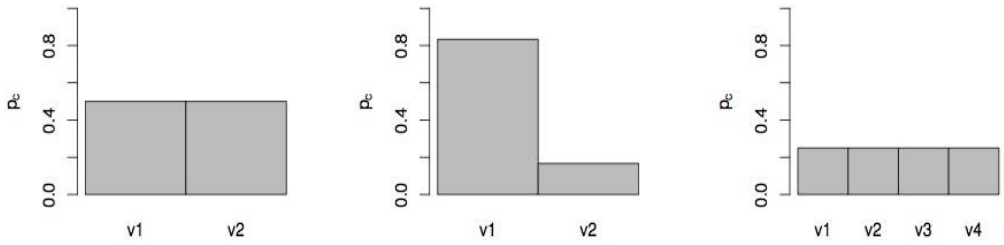
   b. $X_1 \vee (X_2 \wedge X_3)$

SOLUCIÓN: En base a las tablas de verdad buscamos la variable más relacionada con el resultado.

a. X_1=F: F
X_1=T
| X_2=F: T
| X_2=T: F

b. X_1=F
| X_2=F: F
| X_2=T
| | X_3=F: F
| | X_3=T: T
X_1=T: T

6. Dadas las variables a, que toma los valores {v1, v2}, b, que toma los valores {v1,v2}, y c que toma los valores {v1,v2,v3,v4}, responde **razonadamente** a la pregunta ¿qué variable tiene la mayor entropía y cual tiene la menor?



a b c

SOLUCIÓN: La variable con menor entropía es b porque uno de sus valores (v1) es mucho más frecuente que el otro. Las otras dos variables tienen valores equiprobables, pero como c tiene más valores hay más incertidumbre y por tanto esta el la variable con mayor entropía.

7. Dado un conjunto de entrenamiento con 150 ejemplos y tres variables, la variable Color que puede tomar los valores Rojo, Azul o Amarillo, la variable Tamaño que toma los valores Grande o Pequeño y la variable a predecir, COMPRA; que toma los valores SÍ o NO. Suponiendo que:

   - Hay 50 ejemplos donde Color=Rojo, 50 para los que Color=Azul y 50 para los que Color=Amarillo.
   - Hay 20 ejemplos para los que Tamaño=Grande y 130 para los que Tamaño=Pequeño.
   - Hay 100 ejemplos de la clase COMPRA= SI y 50 de la clase COMPRA=NO
   - IG(Color)=IG(Tamaño)

   a) **Razona** qué variable selecciona el algoritmo C4.5 como raíz

SOLUCIÓN: El algoritmo C4.5 usa como métrica para decidir que variable se usa como próximo nodo del árbol el Gain Ratio (GR), como esta métrica se basa en dividir la ganancia de información entre SplitInfo y sabiendo que la ganancia de información de las dos variables es iguale, todo se reduce a calcular el valor de SplitInfo.

$$SplitInfo(Color) = (-0.33 \cdot \log_2 0.33) \cdot 3 = 1.58$$

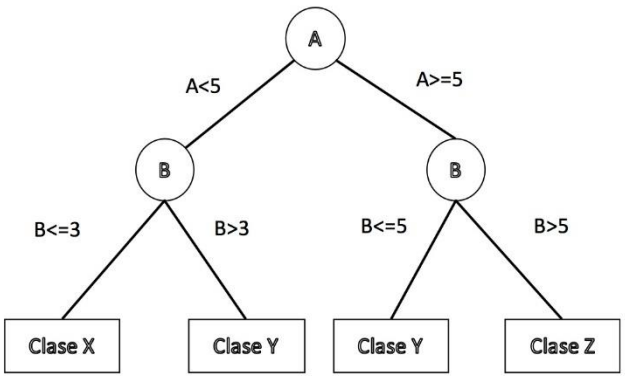
$$SplitInfo(Tamaño) = -0.13 \cdot \log_2 0.13 - 0.86 \cdot \log_2 0.86 = 0.57$$

El algoritmo C4.5 seleccionara la variable Tamaño por delante de la variable Color.

b) ¿Es el 70% un porcentaje de acierto aceptable para el método C4.5?

SOLUCIÓN: La frecuencia de la clase más frecuente es 2 tercios y este es el valor de porcentaje de acierto a batir por cualquier modelo. Un modelo con un 70% de acierto es mejor que nada, con lo que tendríamos un 67%, por tanto es aceptable.

8. Dado el siguiente conjuntos de datos y el siguiente árbol de decisión, asumiendo que los ejemplos {EJ1, ….EJ6} forman el conjunto de entrenamiento y {EJ7, …EJ10} el conjunto de test



Ejemplo	A	B	CLASE
EJ1	3	3	Y
EJ2	2	2	X
EJ3	5	2	X
EJ4	7	6	Y
EJ5	7	7	Y
EJ6	8	6	Z
EJ7	2	4	Y
EJ8	8	9	Z
EJ9	5	5	Z
EJ10	1	4	X

a) ¿Cuál es el error en entrenamiento? ¿Y en test? Utilizando solo los datos de error en entrenamiento y test, responde **razonadamente** si existe sobreajuste o no.

SOLUCIÓN: En el conjunto de entrenamiento se acierta en los ejemplos 2 y 6 mientras que se falla en el resto. Esto da una tasa de error de 4/6=67%. Por otro lado, con los datos de test se acierta en los ejemplos 7 y 8 lo que da una tasa de error de 50%. Dado que el modelo presenta mejor resultado con los datos de testeo no parece que exista sobreajuste.

b) Calcula la matriz de confusión asociada al modelo descrito por el árbol de decisión.

SOLUCIÓN:

	Real		
	X	Y	Z

Pred.	X	1	1	0
	Y	2	1	1
	Z	0	2	2

c) Calcula los valores de TP y TN para cada una de las clases.

SOLUCIÓN: TP(X)=1; TN(X)=6; TP(Y)=1 ;TN(Y)=3 ;TP(Z)=2 ;TN(Z)=5

9. Dado un conjunto de datos con 7 variables binarias, Rain, Wind, Summer, Winter, Day y Night que toman el valor "yes" o "no" y Flight_Delay que toma los valores "Delayed" y "not Delayed", considera la siguiente tabla de datos, que explica cómo se comporta la variable Flight_Delay con respecto a las otras 6.

Feature	Value = yes	Value = no
Rain	Delayed - 30, not Delayed - 10	Delayed - 10, not Delayed - 30
Wind	Delayed - 25, not Delayed - 15	Delayed - 15, not Delayed - 25
Summer	Delayed - 5, not Delayed - 35	Delayed - 35, not Delayed - 5
Winter	Delayed - 20, not Delayed - 10	Delayed - 20, not Delayed - 30
Day	Delayed - 20, not Delayed - 20	Delayed - 20, not Delayed - 20
Night	Delayed - 15, not Delayed - 10	Delayed - 25, not Delayed - 30

Si construimos un árbol de decisión utilizando C4.5, responde **razonadamente** a la pregunta ¿cuál es la variable raíz?

SOLUCIÓN: Hay que ver cuál es la variable con mayor valor de GainRatio que es la relación entre IG y SplitInfo. Se puede ver claramente que las variables Rain, Wind, Summer y Day son equiprobables y por lo tanto su valor de SplitInfo será de 1. De estas variables se puede ver de la tabla que la que mejor separa los valores de Delayed es Summer, así que no será necesario hacer los cálculos para las otras tres.

Partimos de las entropías: $H(Delayed) = 1; H(Delayed|Summer) = 0.54; H(Delayed|Winter) = 0.95; H(Delayed|Night) = 0.99$

Los valores de SplitInfo: $SplitInfo(Winter) = 0.95; SplitInfo(Night) = 0.9$

Por último calculamos GR: $GR(Delayed|Summer) = \frac{1-0.54}{1} = 0.46$
$GR(Delayed|Winter) = \frac{1-0.95}{0.95} = 0.05; GR(Delayed|Night) = \frac{1-0.99}{0.9} = 0.01$

La variable seleccionada por C4.5 será Summer.

10. Dado el conjunto de ejemplos descrito a continuación:

Ejemplo	X1	X2	X3	Clase
EJ1	0	250	36	**A**
EJ2	10	150	34	**B**
EJ3	4	20	1	**B**
EJ4	6	78	8	**B**
EJ5	2	90	10	**A**
EJ6	1	170	70	**A**
EJ7	6	200	45	**A**
EJ8	8	160	41	**B**
EJ9	10	180	38	**A**

a) Completa la siguiente tabla de entropías de las particiones asociadas a los atributos X1, X2, X3:

Entropía	YES	NO
X1<=5	0.811	0.971
X2<=160	0.722	
X3<=40		0.918

SOLUCIÓN: $H(Clase|X2 > 160) = 0; H(Clase|X \leq 40) = 1$

b) Sabiendo que H(Clase)=0.991, y utilizando los datos de la tabla del apartado a), calcula el atributo más informativo según la GANANCIA DE INFORMACIÓN.

SOLUCIÓN: $IG(Clase|X1) = 0.991 - \left(\frac{4}{9} \cdot 0.811 + \frac{5}{9} \cdot 0.971\right) = 0.091;$
$IG(Clase|X2) = 0.991 - \left(\frac{5}{9} \cdot 0.722 + 0\right) = 0.590; IG(Clase|X3) = 0.991 -$
$\left(\frac{6}{9} \cdot 1 + \frac{3}{9} \cdot 0.918\right) = 0.018$

El atributo más informativo es X2.

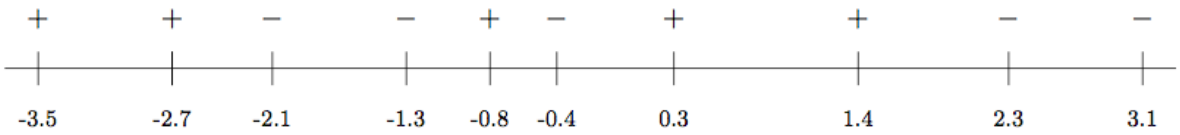
Vecinos más cercanos

11. Dado el conjunto de ejemplos del ejercicio 10 considera el conjunto {EJ1, EJ2, …, EJ8} como entrenamiento, realiza los cálculos para clasificar el ejemplo EJ9 utilizando 3NN con la distancia euclídea. ¿Es un acierto?

SOLUCIÓN: $dist(EJ9, EJ1) = \sqrt{10^2 + 70^2 + 2^2} = 70.74$; $dist(EJ9, EJ2) = 30.27$; $dist(EJ9, EJ3) = 164.33$; $dist(EJ9, EJ4) = 106.4$; $dist(EJ9, EJ5) = 95.54$; $dist(EJ9, EJ6) = 34.71$; $dist(EJ9, EJ7) = 21.56$; $dist(EJ9, EJ8) = 20.32$

Con estas distancias, los tres vecinos más cercanos son EJ2, EJ7 y EJ8. Como dos de ellos tienen clase 'B' esa sería la predicción y sería un fallo.

12. Dado el siguiente conjunto de ejemplos con un atributo real y una variable objetivo que toma dos valores (+,-)



x	-3.5	-2.7	-2.1	-1.3	-0.8	-0.4	0.3	1.4	2.3	3.1
y	+	+	−	−	+	−	+	+	−	−

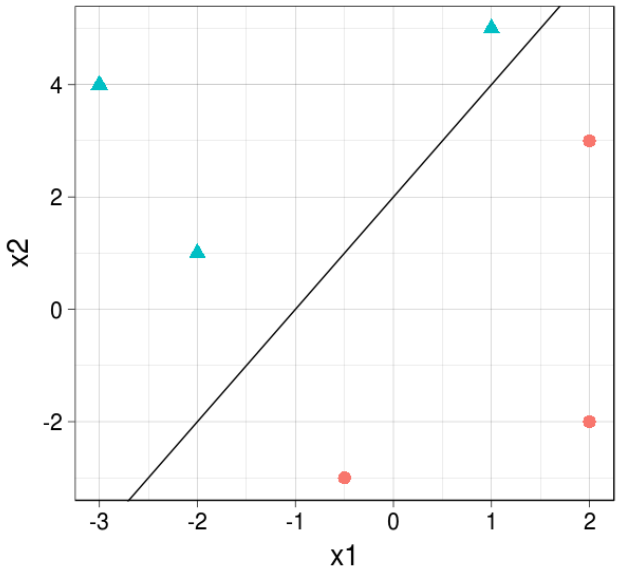
Utilizando como conjunto de test los ejemplos (-3.5, +), (-0.8, +) y (-0.4,-) y como conjunto de entrenamiento el resto, calcula **razonadamente** el porcentaje de error que se obtiene con K-NN, K=3.

SOLUCIÓN: Para -3.5 los vecinos son -2.7, -2.1 y -1.3, la predicción es '-' y es un error. Para -0.8 los vecinos son -1.3, -2.1 y 0.3, la predicción es '-' y es un error. Para -0.4 los vecinos son -1.3, 0.3 y -2.1, la predicción es '-' y es un acierto.

Esto da una tasa de error del 66%.

Redes neuronales

13. Calcula los valores de los pesos w_0, w_1, w_2 para el perceptrón cuya frontera de decisión se ilustra en la siguiente figura.



SOLUCIÓN: Buscamos la recta que muestra la figura y esta es x2=2x1+2. Dado que el valor umbral del perceptrón es 0, dejamos la ecuación de la recta con un 0 a un lado del igual, por ejemplo 2x1-x2+2=0 y de aquí tenemos los valores de los pesos asociados a cada variable: $w_0 = 2; w_1 = 2; w_2 = -1$. Otra alternativa con -2x1+x2-2: $w_0 = -2; w_1 = -2; w_2 = 1$

14. Diseña un perceptrón para la función booleana $A \wedge \neg B$

SOLUCIÓN: En vez de usar el algoritmo de aprendizaje del perceptrón podemos ver en la tabla de verdad de la función si valores de A y B implican una salida alta (1) o baja (-1) para saber si los sus pesos deben ser altos o bajos. Otra opción es hacer una gráfica con los cuatro puntos y resolver el problema como el ejercicio anterior. En cualquier caso una solución puede ser $w_0 = -0.5; w_1 = 1; w_2 = -1$

15. Construye a mano una red de perceptrones para la función booleana XOR.

SOLUCIÓN: Hay que tener en cuenta que la función XOR se puede expresar como $(A \wedge \neg B) \vee (\neg A \wedge B)$, así vemos que podemos generar la red con tres perceptrones dos que hagan los resultados entre paréntesis, a partir del ejercicio anterior, y la salida de estos dos se combina con un último perceptrón que codifique la función OR.

16. Considera una red neuronal con **función de activación lineal**. Tiene dos entradas, una neurona oculta y una unidad de salida. Habrá un total de cinco pesos $(w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0})$ que vamos a inicializar todos a 0.1. Calcula la actualización de los pesos utilizando el algoritmo de propagación hacía atrás con una tasa de aprendizaje $\eta = 0.3$ y los siguientes ejemplos de entrenamiento:

a	b	d
1	0	1
0	1	0

NOTA: Solo es necesario ejecutar una pasada del algoritmo por cada instancia, no se pide llegar a unos valores de convergencia. El usar una función de activación lineal hay que tener en cuenta las expresiones de δ_k y δ_h adecuadas a este caso, que son:

$$\delta_k = (y_k - o_k)$$

$$\delta_h = \sum_k \delta_k w_{kh}$$

SOLUCIÓN: $w_{d0} = 0.24442;\ w_{dc} = 0.12573;\ w_{c0} = 0.10813;\ w_{ca} = 0.1264;\ w_{cb} = 0.08173$

Máquinas de vector soporte

17. Dado el siguiente conjunto de entrenamiento, al que se le ha aplicado SVM con kernel lineal para predecir la variable Y, calcula los parámetros del hiperplano obtenido.

X1	4	5	9	7	2	4	9	2
X2	5	6	4	9	1	4	8	0
Y	1	-1	-1	-1	1	1	-1	1
α	1	1	0	0	0	0	0	0

SOLUCIÓN: $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i = 1 \cdot 1 \cdot \begin{pmatrix}4\\5\end{pmatrix} + 1 \cdot (-1) \cdot \begin{pmatrix}5\\6\end{pmatrix} = \begin{pmatrix}-1\\-1\end{pmatrix}$

$$b = \frac{1}{|VS|}\sum_i y_i - \boldsymbol{w}\boldsymbol{x}_i = \frac{1}{2} \cdot (1 - (-1 \quad -1)\begin{pmatrix}4\\5\end{pmatrix} + (-1) - (-1 \quad -1)\begin{pmatrix}5\\6\end{pmatrix} = 10$$

18. Dados dos puntos con dos dimensiones $p_1 = (1,2)$ y $p_2 = (3,4)$, cada uno pertenece a una clase. Calcula los parámetros y la función discriminante obtenida por un modelo SVM con kernel lineal.

SOLUCIÓN: Si representamos gráficamente este problema vemos que la coordenada equidistante de los dos vectores soporte es el punto (2,3). La pendiente de la recta que une los dos vectores soporte es 1, pero la línea discriminante es perpendicular a esta y su pendiente será -1. De esto tenemos que la ecuación de línea discriminante es $x_2 = -x_1 + 5$. Como la expresión de esta línea según un SVM debe ser $\boldsymbol{w}\boldsymbol{x} + b = 0$ tenemos que $\boldsymbol{w} = \begin{pmatrix}-1\\-1\end{pmatrix}$ y $b = 5$.

Verifiquemos si estos parámetros son conformes a un SVM. La función discriminante debe da 1 o -1 para los vectores soportes. Para p_1 tenemos $g(p_1) = -1 - 2 + 5 = 2$ y para p_2, $g(p_2) = -3 - 4 + 5 = -2$. Además, la anchura del margen es según el valor de $\boldsymbol{w}\ \frac{2}{\|\boldsymbol{w}\|} = 1.4$, pero el valor que debería ser atendiendo a la distancia euclídea entre los puntos es $\sqrt{(3-1)^2 + (4-2)^2} = 2\sqrt{2} = 2.8$

Hay que escalar el valor de los parámetros para que se ajusten a la definición de SVM. En realidad, la función discriminante se puede escalar por cualquier número c:

$$-cx_1 - cx_2 + 5c = 0$$

Entonces, por ejemplo ajustando por el ancho del margen, $\frac{2}{\|\boldsymbol{w}\|} = 2\sqrt{2} = \frac{2}{\sqrt{-c^2+(-c)^2}} = \frac{2}{c\sqrt{2}}$, con lo que $c = \frac{1}{2}$

Valores finales: $\boldsymbol{w} = \begin{pmatrix}-0.5\\-0.5\end{pmatrix}$; $b = 2.5$

19. Dado el siguiente conjunto de entrenamiento {(1,+),(2,+),(4,-),(5,-),(6,+)} al que se la he aplicado SVM con kernel $K(x,y) = (xy+1)^2$, obteniendo $\alpha_1 = 0$, $\alpha_2 = 2.499$, $\alpha_3 = 0$, $\alpha_4 = 7.331$, $\alpha_5 = 4.832$. Calcula los parámetros que definen el hiperplano para separar las clases.

SOLUCIÓN: En clasificación tenemos que $g(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b$. Para cualquier vector soporte se debe cumplir que $|g(VS)| = 1$, entonces: $b = y_k - \sum_i \alpha_i y_i k(\boldsymbol{x_i}, \boldsymbol{x_k})$. Para el vector soporte (2,+) tenemos:

$$b = 1 - (2.499 \cdot 1 \cdot (2 \cdot 2 + 1)^2 + 7.331 \cdot (-1) \cdot (5 \cdot 2 + 1)^2 + 4.832 \cdot 1 \cdot (6 \cdot 2 + 1)^2$$

$$b = 8.968$$

20. Con el conjunto de entrenamiento anterior y el valor del parámetro b redondeado sin decimales, calcula la predicción de este modelo para los valores 2.5 y 6.5.

SOLUCIÓN: Usamos $b = 9$

$$g(2.5) = 2.499 \cdot 1 \cdot (2 \cdot 2.5 + 1)^2 + 7.331 \cdot (-1) \cdot (5 \cdot 2.5 + 1)^2 + 4.832 \cdot 1 \cdot (6 \cdot 2.5 + 1)^2 + 9 = -0.119$$

$$g(6.5) = 2.499 \cdot 1 \cdot (2 \cdot 6.5 + 1)^2 + 7.331 \cdot (-1) \cdot (5 \cdot 6.5 + 1)^2 + 4.832 \cdot 1 \cdot (6 \cdot 6.5 + 1)^2 + 9 = 2.789$$

Con estos resultados el valor de 2.5 será clasificado como '-' y el valor de 6.5 como '+'.