Tema 6: Clasificación (Activación sigmoide)

En este tema se describe el procedimiento para realizar un clasificación binaria sobre las muestras graficadas en la Figura 1. Estas se dividen en dos clases: Clase A (Rojo) y Clase B (Azul), cada una con 300 muestras ($m = 600$). La intención es encontrar un modelo matemático que prediga la clase a la que pertenece cualquiera de las muestras de la Figura 1 o incluso nuevas. Los pasos a seguir para lograr esta clasificación son mismo que los que se han seguido en ejemplos anteriores, a excepción de que se presenta el uso de una nueva función de activación mas robusta que la función escalón.

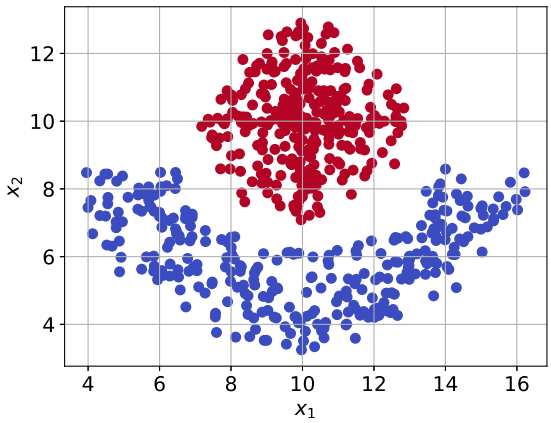


Figure 1: Ejemplo de un conjunto de muestras repartido en dos clases tomado como referencia para un problema de clasificación binaria.

Sobre una clasificación binaria, donde los objetivos son $T = 1$ para la clase A y $T = 0$ para la clase B es posible emplear la *función sigmoide* como función de activación, la cual indica una estimación de probabilidad de que un par x_1 y x_2 pertenezca a la clase 1. Esta función esta graficada Figura 2 y se expresa como:

$$P = g(z) = \frac{1}{1 + e^{(-z)}} \tag{1}$$

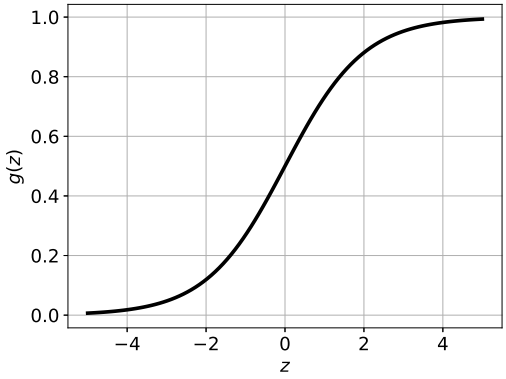


Figure 2: Función de activación sigmoide.

Así, una hipótesis recibe como entrada las características de cada muestra (en este ejemplo son dos características, $x_1^{(i)}$ y $x_2^{(i)}$) generando un resultado parcial $z^{(i)}$ a partir de un polinomio de grado g expresado para dos características como:

$$z^{(i)} = \omega_0 x_0^{(i)} + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \omega_3 (x_1^{(i)})^2 + \omega_4 (x_2^{(i)})^2 + \ldots + \omega_{(g\times 2-1)} (x_1^{(i)})^g + \omega_{(g\times 2)} (x_2^{(i)})^g \tag{2}$$

O de manera expresado de forma general para n características:

$$\begin{aligned} z^{(i)} =& \omega_0 x_0^{(i)} + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \ldots + \omega_n x_n^{(i)} + \\ & \omega_{(n+1)} (x_1^{(i)})^2 + \omega_{(n+2)} (x_2^{(i)})^2 + \ldots + + \omega_{(n\times 2)} (x_n^{(i)})^2 + \ldots + \\ & \omega_{((n\times(g-1))+1)} (x_1^{(i)})^g + \omega_{((n\times(g-1))+2)} (x_2^{(i)})^g + \ldots + \omega_{(n\times g)} (x_n^{(i)})^g \end{aligned} \tag{3}$$

y partir de $z^{(i)}$ la función sigmoide aporta una probabilidad tal que:

$$\mathbb{P}(T = 0|\mathbf{x}) + \mathbb{P}(T = 1|\mathbf{x}) = 1 \tag{4}$$

Así:

$$P^{(i)} = \frac{1}{1 + e^{(-\mathbf{w}^\top \mathbf{x}^{(i)})}} = \mathbb{P}(T = 1|\mathbf{x}^{(i)}) \tag{5}$$

En cuanto a la función de costo a emplear, el error medio cuadrático (MSE) es una opción que puede resultar funcional para propósitos de clasificación. Sin embargo, la *entropia binaria cruzuda* (EBC) es una opción mas adecuada cuando a la salida de la hipótesis se espera idealmente un valor discreto, es decir cero o uno (Si bien la función sigmoide puede entregar valores en el rango $(0, 1)$ se pretende que con ajustando los coeficientes ω el valor de salida sea ≈ 0 o ≈ 1). La EBC para evaluar un costo sobre la i-esima muestra puede ser expresada como:

$$C(P^{(i)}) = \begin{cases} -\log P^{(i)} & \text{si} \quad T^{(i)} = 1 \\ -\log(1 - P^{(i)}) & \text{si} \quad T^{(i)} = 0 \end{cases} \tag{6}$$

donde, dependiendo el valor de $T^{(i)}$ será la expresión a emplear para evaluar este costo. La intucion del funcionamiento de la EBC se ilustra en la Figura 3. Cuando $T^{(i)} = 1$ se aplica la expresión $C(P^{(i)}) = -\log P^{(i)}$ (Curva azul) que arroja un valor de cero cuando $P^{(i)} = 1$, esto implica un costo cero cuando la predicción P coincide con su objetivo T, por el contrario, cuando $P^{(i)} = 0$ siendo su objetivo $T^{(i)} = 1$, el costo tiende a infinito $C(P^{(i)}) \longrightarrow \infty$. Esta misma lógica aplica en el escenario contrario, cuando $T^{(i)} = 0$; si la predicción $P^{(i)} = 0$, el costo sera cero puesto que la predicción coincide con el objetivo $(P^{(i)} = T^{(i)})$, por otro lado, el costo tendera a ser infinito $C(P^{(i)}) \longrightarrow \infty$ cuando $P^{(i)} = 1$ dado que la predicción es el valor discreto contrario al que se espera $(P^{(i)} = 1, T^{(i)} = 0)$. Esta función de costo EBC permite tener un costo que se eleva de manera exponencial a medida la predicción se aleja de su objetivo.

Para fines prácticos y afín de optimizar su implementación en un lenguaje de programación, la Ecuación 6 puede ser expresada como:

$$C^{(i)} = -T^{(i)} \log P^{(i)} - (1 - T^{(i)}) \log(1 - P^{(i)}) \tag{7}$$

la cual, es la suma de dos términos que dependiendo el valor de $T^{(i)}$ se anulará uno u otro. Por ejemplo, si $T^{(i)} = 1$ solo se aplicará el termino $-(1) \log P^{(i)}$ ya que el segundo contendrá una multiplicación por cero $(-(0) \log(1 - P^{(i)}))$. Asimismo, cuando $T^{(i)} = 0$, el primer termino queda anulado ya que se tiene $-(0) \log P^{(i)}$ y por tanto solo se calcula el costo en base a $-(1 - 0) \log(1 - P^{(i)})$.

Entonces si el desea calcular el costo global considerando las m muestras, es necesario obtener el promedio de los costos por muestra, esto es:

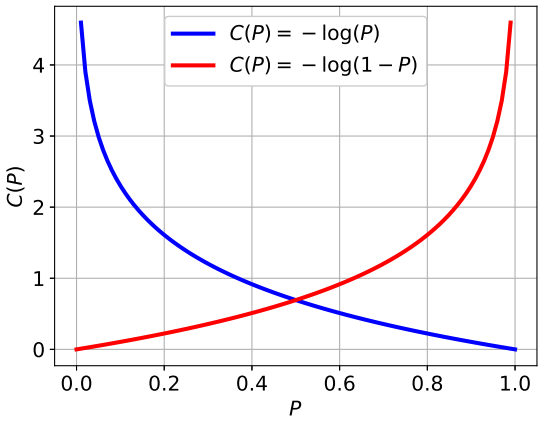


Figure 3: Comparación visual de las funciones $C(P) = -\log(P)$ y $C(P) = -\log(1-P)$.

$$C = -\frac{1}{m}\sum_{i=1}^{m} T^{(i)} \log P^{(i)} + (1 - T^{(i)}) \log(1 - P^{(i)}) \tag{8}$$

Siendo esta Ecuación 8 la función de costo a optimizar con el uso del GD. Para ello, se requiere obtener sus derivadas con respecto a todos coeficientes ω. Considerando una hipótesis compuesta por la Figura 2 y Ecuación 2 se tiene el calculo de las derivadas en los pasos mostrados a continuación:

Dado que la derivada de la EBC es la derivada de una sumatoria, se pueden derivar individualmente cada uno de los términos:

$$\frac{\partial C(\mathbf{w})}{\partial \omega_j} = -\frac{1}{m}\frac{\partial}{\partial \omega_j}\sum_{i=1}^{m} T^{(i)} \log P^{(i)} + (1 - T^{(i)}) \log(1 - P^{(i)}) \tag{9}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial \omega_j} T^{(i)} \log P^{(i)} + \frac{\partial}{\partial \omega_j}(1 - T^{(i)}) \log(1 - P^{(i)}) \tag{10}$$

ahora, dado que que los términos $\log P^{(i)}$ y $\log(1-P^{(i)})$ son funciones que dependen de la variable $P^{(i)}$, a su vez el valor $P^{(i)}$ depende de la función $g(z)$ (Figura 2), la cual tiene como argumento de entrada el resultado $z^{(i)}$ dado por la función mostrada en la Ecuación 2 es posible expresar cada derivada como una multiplicación de derivadas usando la regla de la cadena. Considerando que $l(P^{(i)}) = \log P^{(i)}$ y $h(P^{(i)}) \log(1 - P^{(i)})$ se tiene que:

$$\frac{\partial C(\mathbf{w})}{\partial \omega_j} = -\frac{1}{m}\sum_{i=1}^{m} T^{(i)} \frac{\partial l(P^{(i)})}{\partial P^{(i)}} \frac{\partial P^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial \omega_j} + (1 - T^{(i)}) \frac{\partial h(P^{(i)})}{\partial P^{(i)}} \frac{\partial P^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial \omega_j} \tag{11}$$

De esta manera, se pueden realizar seis derivadas mas simples y al final sustituir los resultados en la Ecuación 11. Así la primera derivada queda como:

$$\frac{\partial l(P^{(i)})}{\partial P^{(i)}} = \frac{\partial \log P^{(i)}}{\partial P^{(i)}} = \boxed{\frac{1}{P^{(i)}}} \tag{12}$$

mientras que las derivada para $h(P^{(i)})$ resulta:

$$\frac{\partial h(P^{(i)})}{\partial P^{(i)}} = \frac{\partial \log(1-P^{(i)})}{\partial P^{(i)}} = \boxed{\frac{-1}{1-P^{(i)}}} \tag{13}$$

Posteriormente se realiza una derivada que aparece dos veces en la Ecuación 11, que es la derivada de la función sigmoide:

$$\frac{\partial P^{(i)}}{\partial z^{(i)}} = \frac{\partial}{\partial z^{(i)}} \frac{1}{1+\exp(-z^{(i)})} \tag{14}$$

Utilizado el método para derivar una fracción:

$$\frac{d}{dx}\frac{u}{v} = \frac{vu' - uv'}{v^2} \tag{15}$$

y por propósitos de simplificación visual , se considera que $\sigma^{(i)} = \frac{1}{1+e^{-z^{(i)}}}$, teniendo que:

$$\frac{\partial P^{(i)}}{\partial z^{(i)}} = \frac{\partial}{\partial z^{(i)}} \frac{1}{1+e^{(-z^{(i)})}} = \frac{(1+e^{-z^{(i)}})\frac{\partial}{\partial z^{(i)}}(1) - (1)(\frac{\partial}{\partial z^{(i)}}(1+e^{-z^{(i)}}))}{(1+e^{-z^{(i)}})^2} \tag{16}$$

$$= \frac{(1+e^{-z^{(i)}})(0) - (1)(e^{-z^{(i)}}(-1))}{(1+e^{-z^{(i)}})^2} \tag{17}$$

$$= \frac{0-(-e^{-z^{(i)}})}{(1+e^{-z^{(i)}})(1+e^{-z^{(i)}})} = \frac{1}{1+e^{-z^{(i)}}}\frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}} \tag{18}$$

$$= \frac{1}{1+e^{-z^{(i)}}}\left(\frac{-1+1+e^{-z^{(i)}}}{1+e^{-z^{(i)}}}\right) \tag{19}$$

$$= \frac{1}{1+e^{-z^{(i)}}}\left(\frac{-1}{1+e^{-z^{(i)}}} + \frac{1+e^{-z^{(i)}}}{1+e^{-z^{(i)}}}\right) \tag{20}$$

$$= \frac{1}{1+e^{-z^{(i)}}}\left(1 - \frac{1}{1+e^{-z^{(i)}}}\right) = \boxed{\sigma^{(i)}(1-\sigma^{(i)})} \tag{21}$$

finalmente, se deriva el termino restante que igualmente aparece dos veces en la Ecuación 11. Así:

$$\frac{\partial z^{(i)}}{\partial \omega_j} = \frac{\partial}{\partial \omega_j}\omega_0 x_0^{(i)} + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \omega_3 (x_1^{(i)})^2 + \omega_4 (x_2^{(i)})^2 + \ldots + \omega_{(g\times 2-1)}(x_1^{(i)})^g + \omega_{(g\times 2)}(x_2^{(i)})^g \tag{22}$$

$$= \boxed{d_j^{(i)}} \tag{23}$$

donde $d_j^{(i)}$ corresponde al j-esimo termino del polinomio que multiplica a ω_j. por ejemplo $d_1^{(i)} = \frac{\partial z^{(i)}}{\partial \omega_1} = x_1^{(i)}$, $d_3^{(i)} = \frac{\partial z^{(i)}}{\partial \omega_3} = (x_1^{(i)})^2$, $d_{(g\times 2-1)}^{(i)} = \frac{\partial z^{(i)}}{\partial \omega_{(g\times 2-1)}} = (x_1^{(i)})^g$, etc.

Entonces, sustituyendo los resultados, resulta:

$$\frac{\partial C(\mathbf{w})}{\partial \omega_j} = -\frac{1}{m}\sum_{i=1}^{m} T^{(i)}\frac{1}{P^{(i)}}(\sigma^{(i)}(1-\sigma^{(i)}))d_j^{(i)} + (1-T^{(i)})(\frac{-1}{1-P^{(i)}})\sigma^{(i)}(1-\sigma^{(i)})d_j^{(i)} \tag{24}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} \frac{T^{(i)}}{P^{(i)}}\sigma^{(i)}(1-\sigma^{(i)})d_j^{(i)} - \frac{1-T^{(i)}}{1-P^{(i)}}\sigma^{(i)}(1-\sigma^{(i)})d_j^{(i)} \tag{25}$$

El resto del procedimiento consiste solo en simplificar el resultado. Recordando que $P^{(i)} = \sigma^{(i)}$ se consigue el resultado final:

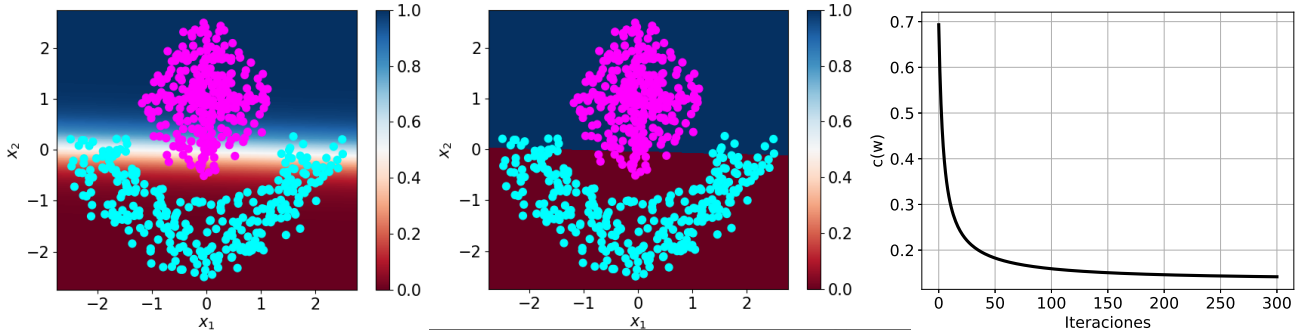
$$\frac{\partial C(\mathbf{w})}{\partial \omega_j} = -\frac{1}{m}\sum_{i=1}^{m} T^{(i)}(1-P^{(i)})d_j^{(i)} - (1-T^{(i)})P^{(i)}d_j^{(i)} \tag{26}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (T^{(i)} - T^{(i)}P^{(i)})d_j^{(i)} - (P^{(i)} - T^{(i)}P^{(i)})d_j^{(i)} \tag{27}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (T^{(i)} - T^{(i)}P^{(i)} - P^{(i)} + T^{(i)}P^{(i)})d_j^{(i)} \tag{28}$$

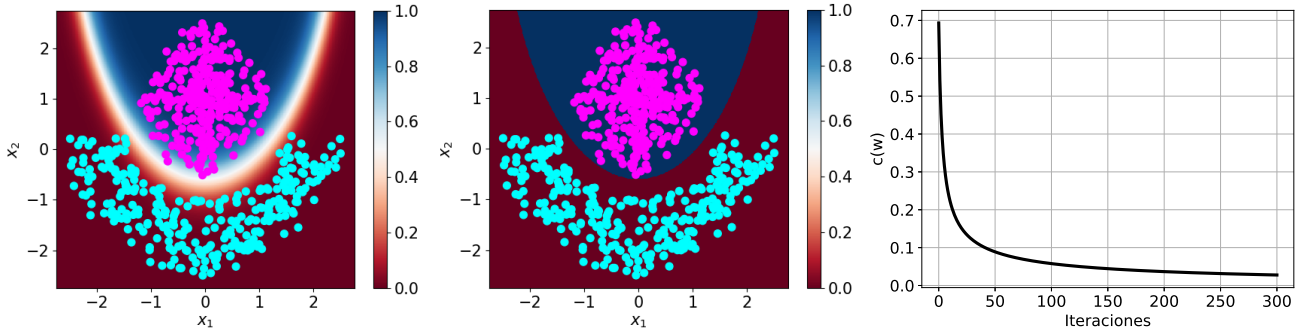
$$= -\frac{1}{m}\sum_{i=1}^{m} (T^{(i)} - P^{(i)})d_j^{(i)} = \boxed{\frac{1}{m}\sum_{i=1}^{m} (P^{(i)} - T^{(i)})d_j^{(i)}} \tag{29}$$

En este punto se tienen los elementos necesarios para implementar el algoritmo de búsqueda basado en GD. Normalizando los valores de x a un rango de $[-2.5, 2.5]$, tomando un valor de $\alpha = 0.4$, trescientas iteraciones y tres valores de g (1,2 y 3) para el polinomio de la Ecuación 2 se logran los resultados de mostrados en la Figura 4 Figura 5 y Figura 6. Para cada valor de g se muestran tres resultados. Primero, el mapa de probabilidades indica el valor P que entrega la hipótesis para cualquier muestra (original o nueva) cuyos valores de sus características x_1 y x_2 se encuentren en un dominio perteneciente al rengo de normalización. Nótese que sobre el mapa de probabilidades se presentan las muestras de la Figura 1 para una mejor interpretación. Así, es claro ver para que valores de g las muestras son separadas correctamente. Esta separación se puede evidenciar el segundo resultado que es la frontera de decisión, donde, al igual que en el mapa de probabilidades, las muestras de Figura 1 se grafican encima. Esta frontera esta definida como por los puntos en el espacio donde $P = 0.5$, es decir, la linea divisoria donde la probabilidad de que una muestra pertenezca a la clase 1 es 50%. Para un valor $g = 1$ se tiene una linea recta como frontera, lo cual, para la distribución que tienen muestras es imposible separa las dos clases por completo. En el resultado de la Figura 4b se tiene un precisión del 93.33 % que indica el porcentaje de muestras que fueron clasificadas correctamente. Para los valores de $g = 2$ y $g = 3$ se logra obtener una frontera que separa a la perfección las muestras representando una precisión en la clasificación de 100%. Sin embargo se debe destacar que lograr un 100% de precisión no lo es todo, es necesario considerar que frontera de decisión puede ser la mejor opción en base a la aplicación que se desee dar o a la posible distribución de nuevas muestras no consideradas en el proceso de optimización. Finalmente, el tercer resultado corresponde al costo obtenido en cada iteración del algoritmo GD.



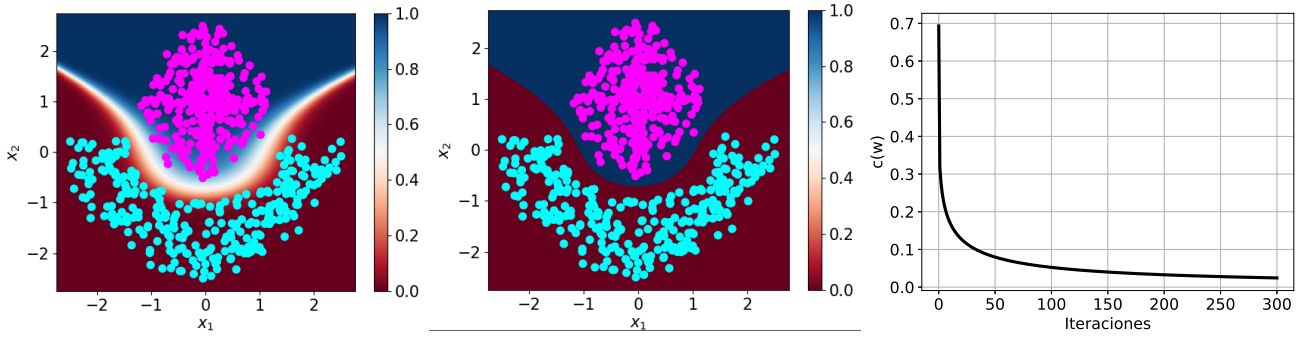
(a) (b) (c)

Figure 4: Resultado del algoritmo de búsqueda para el ejercicio de clasificación para $g = 1$: a) Mapa de probabilidades b) Frontera de decisión c) Costo obtenido en cada iteración.



(a) (b) (c)

Figure 5: Resultado del algoritmo de búsqueda para el ejercicio de clasificación para $g = 2$: a) Mapa de probabilidades b) Frontera de decisión c) Costo obtenido en cada iteración.



(a) (b) (c)

Figure 6: Resultado del algoritmo de búsqueda para el ejercicio de clasificación para $g = 3$: a) Mapa de probabilidades b) Frontera de decisión c) Costo obtenido en cada iteración.

Ejercicios

1. Elaborar un algoritmo de regresión logística que sea capaz de clasificar un conjunto de entrenamiento de dos características generado artificialmente. Para ello se deberán considerar (crear) al menos tres diferentes distribuciones de las características. Como referencia, se pueden crear tres conjuntos de entrenamiento que visualmente se asemejen a los mostrados en la Figura 7. El número total de muestras de be ser de al menos 100 y la posición exacta de cada muestra debe ser generada de forma aleatoria.

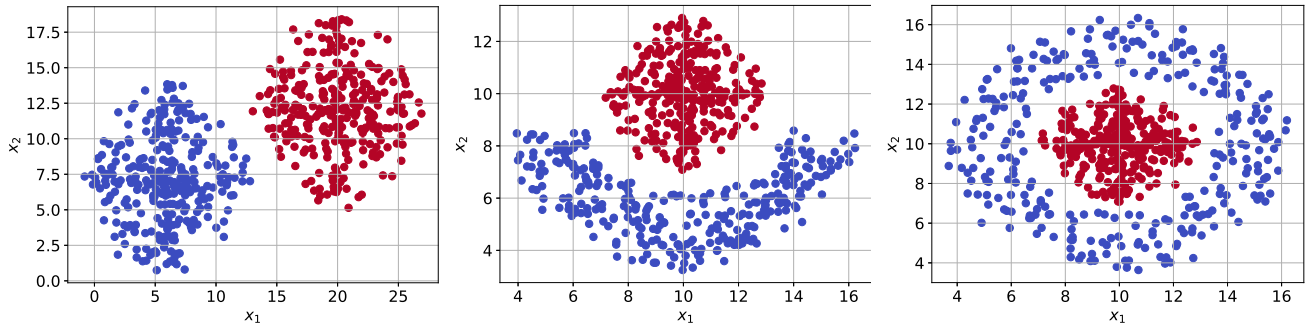
   Para cada conjunto de entrenamiento, los paso a seguir son los siguientes:

   (a) Generar el conjunto de entrenamiento

   (b) Emplear entropía binaria cruzada (Ecuación 8) como función de costo y obtener su derivada con respecto a cada uno de los coeficientes ω, considerando la siguiente hipótesis:

$$P^{(i)} = \frac{1}{1 + e^{-z^{(i)}}} \tag{30}$$

   Pudiendo z_i tener las siguientes formas:



(a) (b) (c)

Figure 7: Ejemplos de conjuntos de muestras para ejercicios de clasificación binaria a)Separable linealmente b) Separable no linealmente c) Separable no linealmente

- $z^{(i)} = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)}$
- $z^{(i)} = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 (x_1^{(i)})^2 + \omega_3 x_2^{(i)} + \omega_4 (x_2^{(i)})^2$
- $z^{(i)} = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \omega_3 x_1^{(i)} x_2^{(i)} + \omega_4 (x_1^{(i)})^2 + \omega_5 (x_2^{(i)})^2 + \omega_6 (x_1^{(i)})^2 (x_2^{(i)})^2$
- Una ecuación (polinomio) propuesta por el estudiante

(c) Para cada variante de hipótesis se deberá:

- Implementar un algoritmo de entrenamiento basado en gradiente descendente.
- Obtener la precisión de clasificación.
- Generar el mapa de predicciones.
- Mostrar la frontera de decisión.

2. Repetir el ejercicio anterior realizando las siguientes cambios:

(a) Utilizar un modelo de hipótesis con dos salidas ($P_1^{(i)}$ y $P_2^{(i)}$) (una para cada clase), es decir:

$$P_h^{(i)} = \frac{1}{1 + e^{-(\sum_{k=0}^{n} \omega_k x_k^{(i)})}} = \frac{1}{1 + e^{-(\omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)})}} \tag{31}$$

donde $h \in [0, 1]$ correspondiente a las salidas. Note como al tener dos salidas por muestra, es necesario también tener dos objetivos por muestras ($T_0^{(i)}$ y $T_1^{(i)}$), donde siempre un objetivo sera '1' y el otro '0'. Asimismo, la función de costo a optimizar será $C = C1 + C2$, donde:

$$C_0 = -\frac{1}{m} \sum_{i=1}^{m} T_0^{(i)} \log P_0^{(i)} + (1 - T_0^{(i)}) \log(1 - P_0^{(i)}) \tag{32}$$

y

$$C_1 = -\frac{1}{m} \sum_{i=1}^{m} T_1^{(i)} \log P_1^{(i)} + (1 - T_1^{(i)}) \log(1 - P_1^{(i)}) \tag{33}$$

(b) comparar los resultados de este ejercicio con respecto al anterior donde se emplea una hipótesis de una salida.