Tema 5: Clasificación

En este apartado se provee una descripción del proceso de clasificación de datos. Considere las muestras graficadas en la Figura 1. Se tiene un conjunto de veinte muestras ($m = 20$), las cuales estas dividas en dos clases; clase A y clase B. Esta asignación inicial de clases se le conoce como *etiquetado* y se realiza en base a la distribución de las características (x_1 y x_2 en este ejemplo) con la que esta representada cada muestra. Debido a que existen dos "nubes" de puntos claramente separados se asignan dos etiquetas distintas (clase A y clase B). El propósito en un ejercicio de clasificación es encontrar un modelo (función matemática) que prediga la clase a la que pertenece cada muestra. Esta función representa lo que se conoce como *frontera de decisión* que divide las zonas en las que se obtendrá una predicción de clase A o clase B para nuevas muestras.

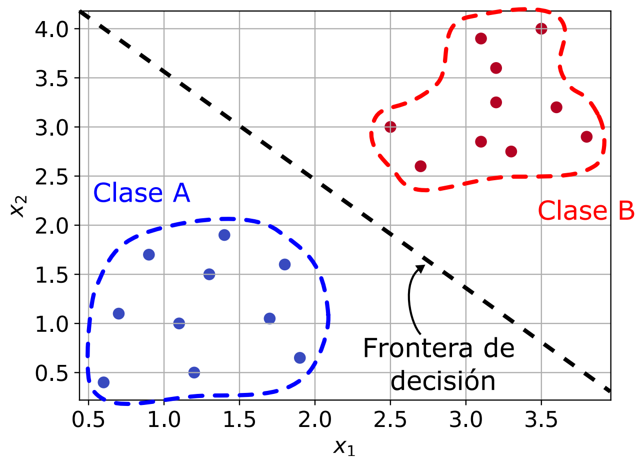


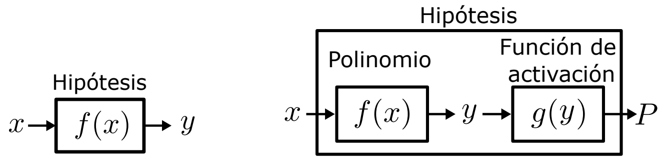
Figura 1: Distribución de muestras etiquetadas en clase A y clase B tomada como referencia para ejemplificar la clasificación de datos.

Al igual que en un ejercicio de regresión lineal, la idea fundamental es proponer una hipótesis que reciba como entradas las características de cada muestra y arroje como resultado un valor que represente la clase a la que pertenece cada muestra. Sin embargo, existen diferencias entre la hipótesis necesaria para un ejercicio de regresión lineal y uno de clasificación. Una hipótesis en regresión lineal (Figura 2a) tiene a su salida y que puede ser avaluada en una función de costo junto con su objetivo T que es una variable numérica, es decir, tanto y como T pueden tener cualquier valor. Por otro lado, en clasificación (Figura 2b) la hipótesis esta conformada por dos funciones. La primera $f(x)$, del mismo modo que en regresión lineal, puede ser cualquier función que arroje como salida un valor numérico y, (aunque para esta explicación se asume que $f(x)$ es un polinomio de grado g). La segunda función $g(x)$ es una función no lineal que idealmente arroje a sus salida un valor discreto P $(0, 1, \ldots, h-1)$ cuyos posibles valores dependen del número de clases (En este ejemplo $h = 2$). Asimismo, P puede ser evaluado en una función de costo junto a un objetivo T que es una variable discreta.

Este proceso de clasificación puede realizarse del mismo modo que en regresión lineal:

Proponer una hipotesis: En este paso se deben definir $f(x)$ y $g(x)$. Como un primer intento de clasificación se establece en este ejemplo:

$$y^{(i)} = f(x_1^{(1)}, x_2^{(i)}) = \omega_0 x_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} \tag{1}$$



(a) (b)

Figura 2: Diagrama de bloques representativo de una hipótesis para a) Regresión linear y b) Clasificación.

y

$$P^{(i)} = g(y^{(i)}) = \begin{cases} 1 & \text{si} \quad y^{(i)} \geq Th \\ 0 & \text{si} \quad y^{(i)} < Th \end{cases} \tag{2}$$

donde $x_0 = 1$ y Th representa un umbral de decisión que será definido arbitrariamente en base a los valores resultantes de y despues de ejecutar el algoritmo de búsqueda.

Inicializar el valor de los coeficientes: El valor inicial de los coeficientes ω se establece arbitrariamente como cero.

Definir una funcion de costo: como ya se ha definido antes, una función de costo cuantifica la diferencia entre el valor de salida de la hipótesis y su objetivo. En este caso, se mostrará el proceso de búsqueda utilizando y y T como argumentos de entrada de esta función. Asimismo se tomara como ejemplo el error medio cuadratico (MSE) como función de costo:

$$C(\mathbf{w}) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)}(x_1^{(1)}, x_2^{(i)}) - T^{(i)})^2 \tag{3}$$

Obtener la derivada de la funcion de costo: La derivada de la Ecuación 3 con respecto a los coeficientes resulta como:

$$\frac{\partial C(\mathbf{w})}{\partial \omega_j} = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)}(x_1^{(1)}, x_2^{(i)}) - T^{(i)}) x_j^{(i)} \tag{4}$$

Implementar el algortimo de optimización: Manteniendo como referencia el algoritmo de búsqueda de gradiente descendente, se tiene su regla de actualización:

$$\omega_{j,t+1} = \omega_{j,t} - \alpha \frac{\partial C(\mathbf{w})}{\partial \omega_{j,t}} \tag{5}$$

Así, ejecutando el GD por diez iteraciones y empleando $\alpha = 0.1$ se tienen los resultados listados en la Tabla 1. En ésta se evidencia que los objetivos T son discretos, teniendo un valor de cero (0) para las muestras correspondientes a la clase A y un valor de uno (1) para aquellas correspondientes a las clase B. Nótese como el valor de y, es menor para la muestras cuyo objetivo es cero y mayor cuando su objetivo es uno. Analizando visualmente estos valores se puede establecer un umbral Th que separa estas dos clases; cuando $T = 0$ el mayor valor de y es 0.4365, mientras que, cuanto $T = 1$ el menor valor de y es 0.7035. En consecuencia, el mejor valor para Th podría ser $Th = (0.4365 + 0.7035)/2 \approx 0.57$.

Los resultados del algoritmo de búsqueda pueden representarse tal como se muestra en la Figura 3. Primero, la Figura 3a corresponde al llamado *Mapa de probabilidades*. Este muestra en una escala de colores el valor de salida de la función $y = f(x)$ para un domino en el cual se encuentran todas las muestras. Para generar este mapa es necesario evaluar la función $f(x)$ para un nuevo conjunto de valores de x_1 y x_2 que tengan un rango similar (ligeramente mayor) a los que describen las muestras del conjunto de entrenamiento

Table 1: Comparación de valores y y T en un ejercicio de clasificación para las muestras de la Figura 1.

y	T	y	T
0.0998	0	0.7035	1
0.1957	0	0.8656	1
0.3124	0	0.7344	1
0.2148	0	0.7940	1
0.2546	0	0.8067	1
0.3432	0	0.8962	1
0.3289	0	0.9462	1
0.3542	0	0.9163	1
0.4365	0	0.9126	1
0.4257	0	1.0150	1

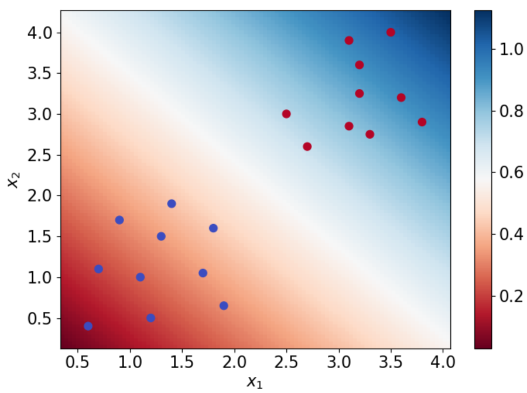
(Figura 1). Cabe destacar que sobre este mapa se grafican las muestras de la Figura 1 para una mejor interpretación del resultado. La selección de colores se selecciona de manera que se evidencien tres zonas importantes: 1) los valores para y cercanos a cero (color rojo) implican que las muestras tienen mayor probabilidad ce pertenecer a la clase A ($T = 0$) mientras que 2) los valores de y mayores, en este caso, alrededor de uno (color azul) señalan las muestras que es mas probable que pertenezcan a la clase B ($T = 1$). Por ultimo, la zona donde y es se encuentra en su valor medio corresponde a la *zona de incertidumbre* (Color blanco) que representa la zona donde alguna nueva muestra tendría aproximadamente la misma probabilidad de pertenecer a cualquiera de las dos clases. En un problema de clasificación, las muestras deberían resultar lo mas alejadas posible de la zona de incertidumbre.

Si a los valores mostrados en el mapa de probabilidades se le aplica la función $g(x)$ utilizando el umbral $Th = 0.57$ es posible generar la Figura 3b la cual muestra la frontera de decisión resaltado en color rojo y azul las zonas donde cualquier muestra será clasificada como clase A o Clase B , respectivamente. Para obtener este resultado se ejecuto el GD durante diez iteraciones y un valor $\alpha = 0.1$ obteniendo los costos mostrados en la Figura 3c asi como los coeficientes $\omega_0 = -0.03954$, $\omega_1 = 0.1356$ y $\omega_2 = 0.1449$ en la ultima iteración.

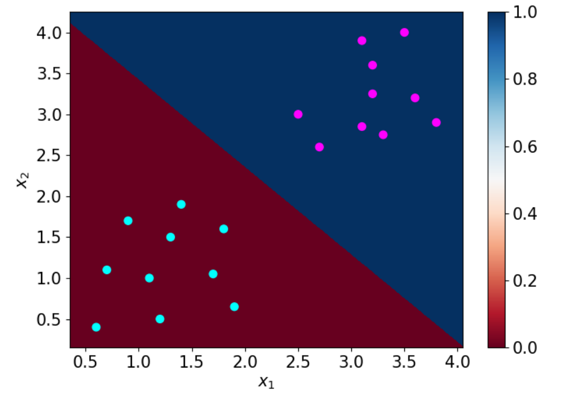
Para eliminar la necesidad de seleccionar arbitrariamente el umbral Th se debería tratar la hipótesis como una función A dependiente de una función B, en este caso $g(f(x))$ (Figura 2b). Teniendo que la hipótesis sería ahora:

$$P^{(i)} = g(y^{(i)}(x_1^{(1)}, x_2^{(i)})) = \begin{cases} 1 & \text{si} \quad y^{(i)}(x_1^{(1)}, x_2^{(i)}) & \geq 0 \\ 0 & \text{si} \quad y^{(i)}(x_1^{(1)}, x_2^{(i)}) & < 0 \end{cases} \tag{6}$$

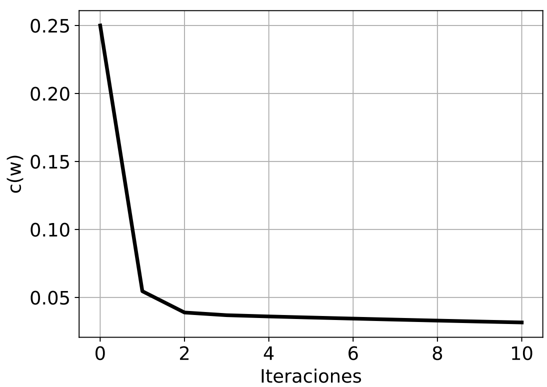
Así, el umbral Th sería inexistente y la frontera de decisión dependería solo de los coeficientes ω. Sin embargo, hasta este punto la función de activación empleada corresponde a un *función escalón* (Figura 4) y su inconveniente para incluirla como parte de la hipótesis es que su derivada es indefinida en $x = 0$ ademas de que derivada es cero en el resto del dominio, lo que implica que el algoritmo GD no seria capaz de actualizar los coeficientes en cada iteración.



(a)



(b)



(c)

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

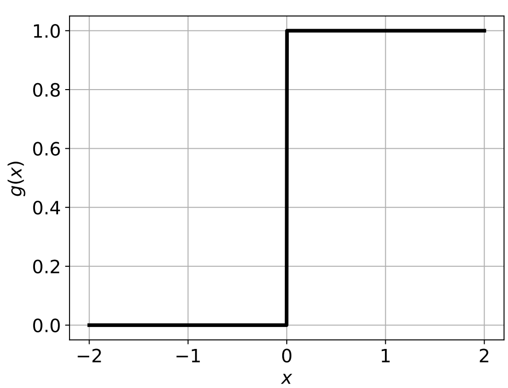


Figura 4: Función escalón. $y = g(x) = 1 \quad \text{si} \quad x \geq 0$ y $g(x) = 0 \quad \text{si} \quad x < 0$

Ejercicios

1. Implementar un algoritmo de clasificación donde se repliquen los resultados mostrados en este tema.