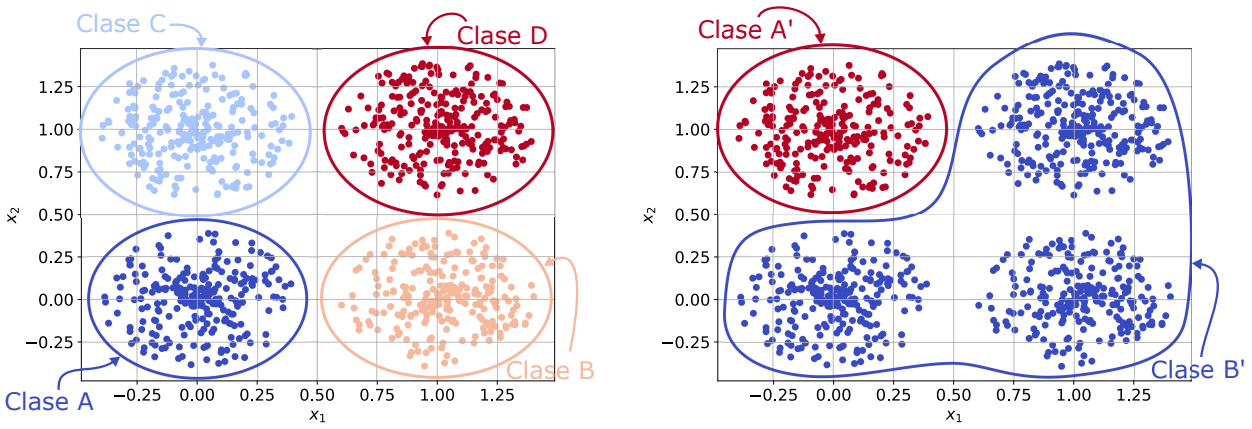
Tema 6: Clasificación Multiclase (Activación sigmoide, Frontera de decisión)

En este tema se explicará el proceso de la clasificación multiclase. Se toma como referencia al conjunto de datos mostrado en la Figura 1a. Este conjunto se divide el cuatro clases: Clase A, B, C y D y la intención es encontrar una función que prediga a que clase pertenece alguna de las muestras (originales o nuevas). Entonces, un método para realizar esta clasificación para múltiples clases consiste en realizar múltiples clasificaciones binarias. Por ejemplo, para obtener un modelo que sea capaz de discriminar la clase C del resto de clases, es necesario tratar esta clase C como una nueva clase A' y el resto de las clases como B' (Figura 1b). Así, existirá un modelo C que indique la probabilidad de que una muestra pertenezca a la clase A'. Nótese que este proceso debe realizarse para cada clase teniendo para cada una de ellas una nueva asignación de clases A' y B'.



(a) (b)

Figura 1: Conjunto de muestras etiquetadas en cuatros clases tomadas como ejemplo para un problema de clasificación multiclase a) Etiquetado original b) Nuevo etiquetado para discriminación binaria de la clase A contra el resto.

Debido a que son cuatro clases, es posible proponer cuatro modelos (subhipótesis) que realicen la clasificación binaria para cada una de las clases. Sin embargo, es necesario considerar un solo modelo para realizar la clasificación multiclase. Este modelo esta representado en la Figura 2. Para cada clasificación binaria existe un polinomio de grado g ($f(\mathbf{x})$) que toma el conjunto $\mathbf{x}$ de n características ($x_0^{(i)} = 1$) para generar la salida $y_d^{(i)}$ a partir de la operación:

$$y_d^{(i)} = \omega_{0,d}x_0 + \omega_{1,d}x_1 + \ldots + \omega_{n,d}x_n \tag{1}$$

donde $d \equiv k$ y es el indicador del modelo que realiza una clasificación sobre la d-esima clase (Cuantitativamente, Clase A $\longrightarrow$ Clase 0, Clase B $\longrightarrow$ Clase 1, etc) sabiendo que existen S clases. Posteriormente una función de activación $g(y_d^{(i)})$ genera la predicción $P_d^{(i)}$.

Debido a que existe una hipótesis que entrega cuatro valores de salida, existe un objetivo $T_d^{(i)}$ para cada una de las muestras en las S clases. Asimismo, es posible evaluar un costo C_d sobre cada salida de la hipótesis empleando las m muestras de tal forma que si existen cuatro clases se pueden calcular cuatro costos. No obstante, este estrategia no es útil puesto que para el proceso numérico para obtener los coeficientes ω es necesario optimizar una sola función, no varias. Por ello, la función a optimizar será la suma de las S funciones costo ilustradas en la Figura 2, teniendo que:

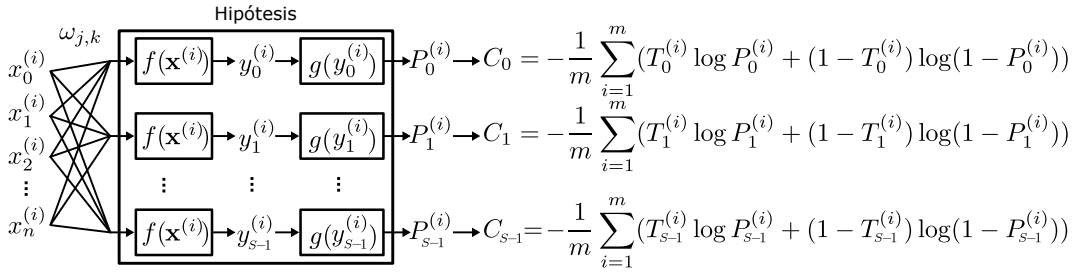


Figura 2: Representación del modelo de hipótesis para la clasificación multiclase.

$$C(\mathbf{W}) = C_0 + C_1 + \ldots + C_{S-1} \tag{2}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\sum_{d=0}^{S-1}(T_d^{(i)} \log P_d^{(i)} + (1 - T_d^{(i)}) \log(1 - P_d^{(i)})) \tag{3}$$

A partir de esta función de costo, es necesario obtener las derivadas con respecto a cada uno de los coeficientes $\omega_{j,k}$, donde $j \in [0, n]$ y $k \in [0, S-1]$. Entonces, las derivadas resultan ser una derivada de una doble sumatoria por lo que se puede derivar por separado cada uno de los términos:

$$\frac{\partial C(\mathbf{W})}{\partial \omega_{j,k}} = \frac{\partial}{\partial \omega_{j,k}} - \frac{1}{m}\sum_{i=1}^{m}\sum_{d=0}^{S-1}(T_d^{(i)} \log P_d^{(i)} + (1 - T_d^{(i)}) \log(1 - P_d^{(i)})) \tag{4}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\sum_{d=0}^{S-1}\left(\frac{\partial}{\partial \omega_{j,k}} T_d^{(i)} \log P_d^{(i)} + \frac{\partial}{\partial \omega_{j,k}}(1 - T_d^{(i)}) \log(1 - P_d^{(i)})\right) \tag{5}$$

y cada termino es en realidad una función que a su vez depende de otras funciones por lo que es posible emplear la regla de la cadena a fin de realizar derivadas mas simples:

$$\frac{\partial C(\mathbf{W})}{\partial \omega_{j,k}} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{d=0}^{S-1}\left(T_d^{(i)} \frac{\partial \log P_d^{(i)}}{\partial P_d^{(i)}} \frac{\partial P_d^{(i)}}{\partial y_d^{(i)}} \frac{\partial y_d^{(i)}}{\partial \omega_{j,k}} + (1 - T_d^{(i)}) \frac{\partial \log(1 - P_d^{(i)})}{\partial P_d^{(i)}} \frac{\partial P_d^{(i)}}{\partial y_d^{(i)}} \frac{\partial y_d^{(i)}}{\partial \omega_{j,k}}\right) \tag{6}$$

Entonces, el resultado de la primera derivada resulta como:

$$\frac{\partial \log P_d^{(i)}}{\partial P_d^{(i)}} = \boxed{\frac{1}{P_d^{(i)}}} \tag{7}$$

Para el segundo termino, el resultado de su primera derivada es:

$$\frac{\partial \log(1 - P_d^{(i)})}{\partial P_d^{(i)}} = \boxed{\frac{-1}{1 - P_d^{(i)}}} \tag{8}$$

La derivada $\frac{\partial P_d^{(i)}}{\partial y_d^{(i)}}$ corresponde a la derivada de la función de activación y ya que en este caso se asume el uso del la función sigmoide, la derivada resulta como:

$$\frac{\partial P_d^{(i)}}{\partial y_d^{(i)}} = \boxed{\sigma_d^{(i)}(1 - \sigma_d^{(i)})} \tag{9}$$

Por ultimo, el resultado de la ultima derivada a resolver queda de la siguiente manera:

$$\frac{\partial y_d^{(i)}}{\partial \omega_{j,k}} = \frac{\partial}{\partial \omega_{j,k}} \sum_{j=0}^{n} \omega_{j,d} x_j^{(i)} = \frac{\partial}{\partial \omega_{j,k}} \omega_{0,d} x_0^{(i)} + \omega_{1,d} x_1^{(i)} + \ldots + \omega_{n,d} x_n^{(i)} = \begin{cases} \boxed{x_j^{(i)}} & \text{si } d = k \\ 0 & \text{si } d \neq k. \end{cases} \tag{10}$$

Sustituyendo los resultados de las derivadas se tiene que:

$$\frac{\partial C(\mathbf{W})}{\partial \omega_{j,k}} = -\frac{1}{m} \sum_{i=1}^{m} \left(T_k^{(i)} \frac{1}{P_k^{(i)}} \sigma_k^{(i)} (1 - \sigma_k^{(i)}) x_j^{(i)} + (1 - T_k^{(i)}) \left(\frac{-1}{1 - P_k^{(i)}} \right) \sigma_k^{(i)} (1 - \sigma_k^{(i)}) x_j^{(i)} \right) \tag{11}$$

Y dado que $\sigma_k^{(i)} = P_k^{(i)}$, es posible simplificar la expresión para llegar al resultado final, tal que:

$$\frac{\partial C(\mathbf{W})}{\partial \omega_{j,k}} = -\frac{1}{m} \sum_{i=1}^{m} \left(T_k^{(i)} (1 - P_k^{(i)}) x_j^{(i)} - (1 - T_k^{(i)}) P_k^{(i)} x_j^{(i)} \right) \tag{12}$$

$$= -\frac{1}{m} \sum_{i=1}^{m} \left(T_k^{(i)} x_j^{(i)} - T_k^{(i)} P_k^{(i)} x_j^{(i)} - P_k^{(i)} x_j^{(i)} + T_k^{(i)} P_k^{(i)} x_j^{(i)} \right) \tag{13}$$

$$= -\frac{1}{m} \sum_{i=1}^{m} \left(T_k^{(i)} x_j^{(i)} - P_k^{(i)} x_j^{(i)} \right) \tag{14}$$

$$= \boxed{\frac{1}{m} \sum_{i=1}^{m} (P_k^{(i)} - T_k^{(i)}) x_j^{(i)}} \tag{15}$$

A partir de estas expresiones (Función de costo y sus derivadas), se puede implementar el algoritmo de optimización (Gradiente descendente por ejemplo). Empleando un valor de $alpha = 0.5$, 1000 iteraciones un polinomio de grado uno como función $f(\mathbf{x})$ se logran los siguientes resultados:

Mapas de probabilidades Por cada salida que tenga la hipótesis (por cada clase) es posible generar un mapa de probabilidades (Figura 3). En cada uno de ellos se visualiza el valor que dará la k-esima salida de la hipótesis para una muestra nueva. Nótese que empleando un polinomio de grado uno, existen muestras ubicadas la zona de incertidumbre indicando que la hipótesis en su salida correspondiente puede predecir un valor menor a 0.5. Lo anterior implica que la muestra en cuestión pertenece a la clase B' significando, en principio, una clasificación incorrecta...

Frontra de decisión Sin embargo la clasificación, esta dada por el valor mayor existente en las salidas de la hipótesis, es decir la clase que se asignará a cada muestra es la posición k para P donde exista el valor mayor. Tomando los cuatro mapas de probabilidades y evaluando para una misma coordenada donde se encuentra el valor mayor es posible genera la Figura 4a, la cual es la representación de la *frontera de decisión multiclase*. En esta representación es evidente que la clasificación es exitosa, teniendo una precisión de 100%

Costo por iteracion Finalmente, se provee información del costo obtenido en cada iteración de algoritmo GD, mostrando que a pesar de no lograr por completo la convergencia se logra una clasificación exitosa.

Como información adicional y punto de comparación, en la Figura 5 y Figura 6 se muestran los resultados correspondientes a lo ya descrito pero usando un polinomio de grado dos. Note como las fronteras de decisión por clase no son homogéneas y no no tiene por que serlo puesto que no hay ningún mecanismo que controle este aspecto.

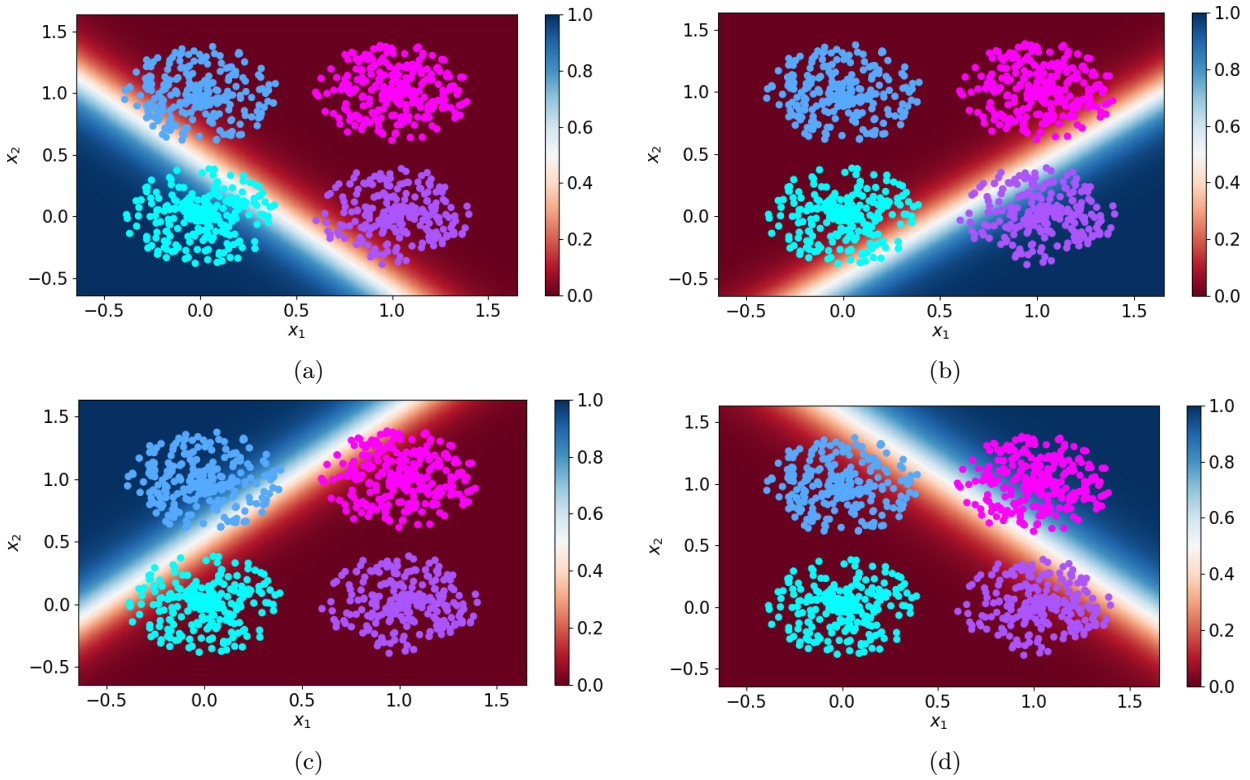


Figura 3: Mapas de probabilidades (MP) para una hipótesis que usa un polinomio de grado uno a) MP para la clase A b) MP para la clase B c) MP para la clase C d) MP para la clase D.

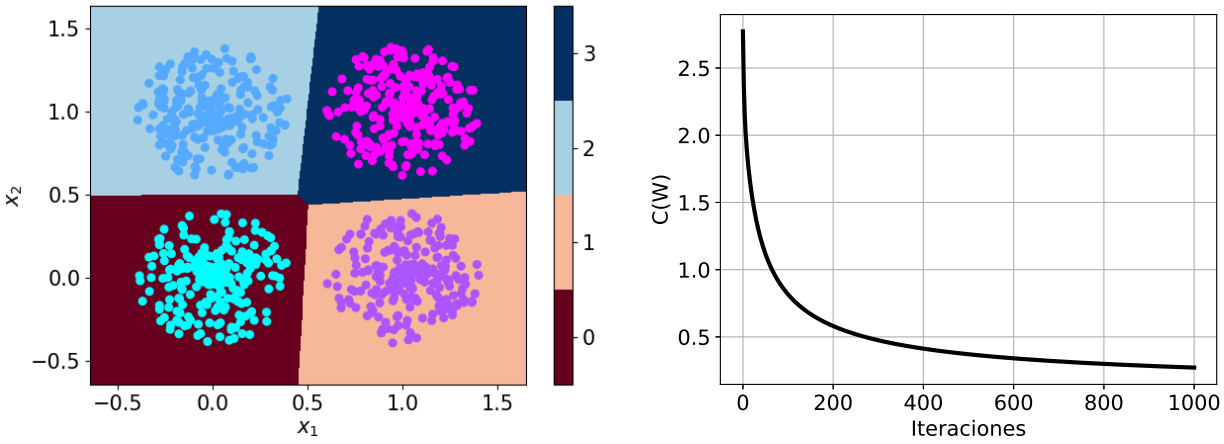


Figura 4: Resultado de la clasificación a partir de una hipótesis que usa un polinomio de grado uno a) Frontera de decisión multiclase (Clase $A \longrightarrow 0$, Clase $B \longrightarrow 1$, Clase $C \longrightarrow 2$, Clase $D \longrightarrow 3$) b) Costo obtenido en cada iteración del GD.

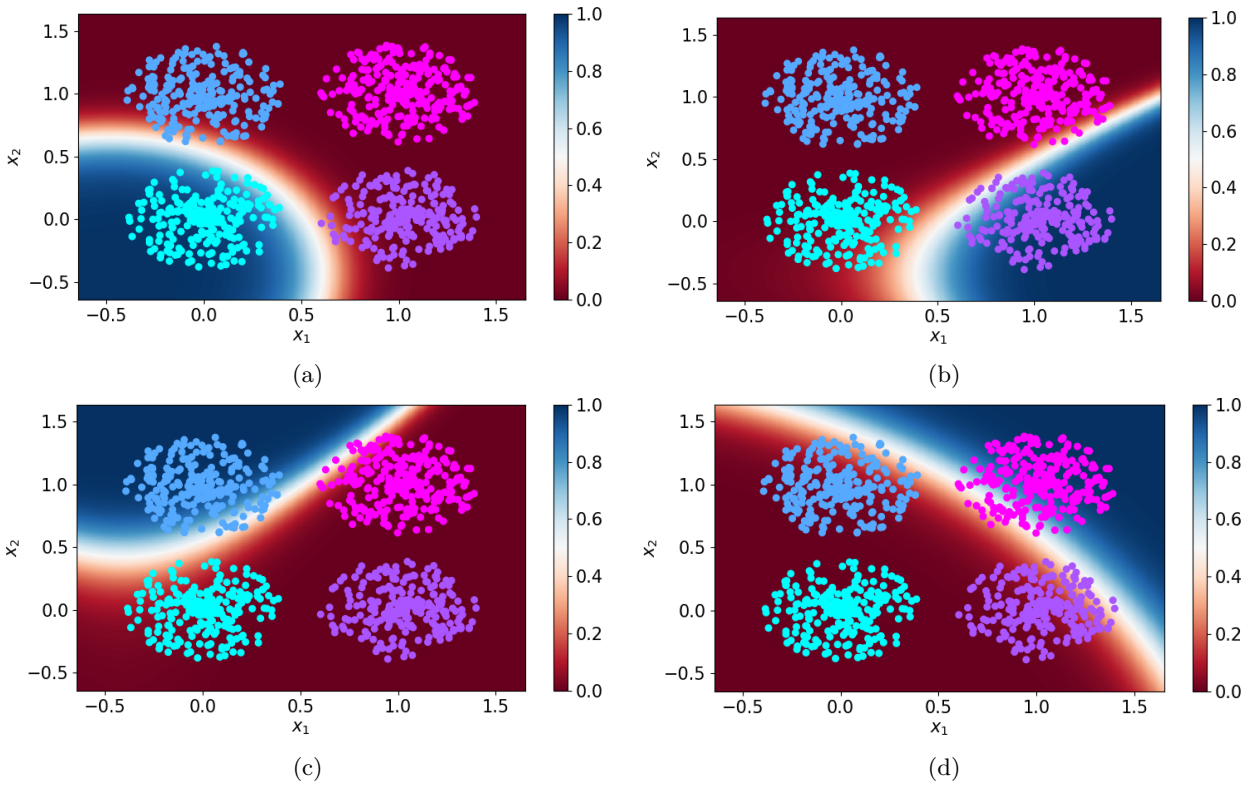


Figura 5: Mapas de probabilidades (MP) para una hipótesis que usa un polinomio de grado dos a) MP para la clase A b) MP para la clase B c) MP para la clase C d) MP para la clase D.

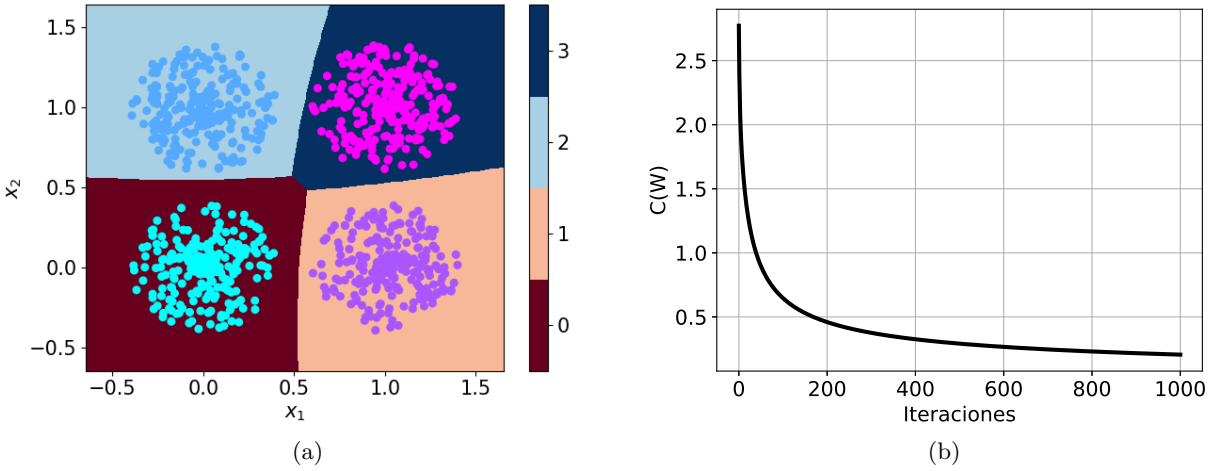


Figura 6: Resultado de la clasificación a partir de una hipótesis que usa un polinomio de grado dos a) Frontera de decisión multiclase (Clase $A \longrightarrow 0$, Clase $B \longrightarrow 1$, Clase $C \longrightarrow 2$, Clase $D \longrightarrow 3$) b) Costo obtenido en cada iteración del GD.

Ejercicios

1. Elaborar un algoritmo de regresión logística donde se repliquen los resultados mostrados en este tema. Se deberán realizar dos implementaciones:

   (a) Empleando un polinomio de grado uno en $f(\mathbf{x}^{(i)})$

   (b) Empleando un polinomio de grado dos en $f(\mathbf{x}^{(i)})$