Introducción al reconocimiento de patrones: Clasificación

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación, bachillerato en Ingeniería en Computación,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

20 de marzo de 2019

En el presente trabajo práctico se estudiarán distintos enfoques básicos de clasificación en dos clases supervisados, de modo que el estudiante pueda estudiar su comportamiento y limitaciones. Los modelos a estudiar e implementar en este trabajo práctico son lineales, e implementan el enfoque discriminativo de clasificación (una muestra pertenece o no pertenece a una clase), prescindiendo de una definición probabilística de pertenencia, la cual separa las etapas de inferencia y de toma de decisiones. Los modelos lineales de clasificación suponen que los datos son linealmente separables, es decir, es posible trazar una superficie de decisión lineal que clasifique correctamente todas las muestras.

1. Modelos lineales para la clasificación en dos clases

Anteriormente se exploró el problema de regresión lineal donde un conjunto de datos a la entrada y salida de un fenómeno $\mathcal{D} = \{\vec{x}, \vec{t}\}$ fueron utilizados para aproximar un modelo específico $y(x, \vec{w})$. Con tal modelo es posible predecir la salida del fenómeno estudiado frente a nuevas muestras $\vec{x}'$.

El objetivo de la clasificación consiste en tomar una muestra de dimensión D, $\vec{m'} = [m_1, \ldots, m_D]^t$ y asignarle una etiqueta k de las $k = 1, \ldots, K$ clases posibles. El espacio de las entradas de dimensión D es dividido en **regiones de decisión,** cuyos límites se denominan superficies de decisión o límites de decisión. Para los clasificadores con modelos lineales, las superficies de decisión se definen como una función lineal respecto a la muestra a evaluar $\vec{m}$, tal función es refererida como **función de activación** f:

$$y(\vec{m}) = f\left(\vec{w'}^T \vec{m}' + w_0\right) = f\left(y(\vec{m})\right), \tag{1}$$

con lo que la posición de la muestra $\vec{m'}$ respecto al hiperplano definido por $y(\vec{m})$ define la clasificación (etiqueta k) de la muestra $\vec{m'}$. La dimensionalidad del vector de pesos, en este caso, es la misma que los datos, por lo que $\vec{w'} \in \mathbb{R}^D$. La función $y(\vec{m'})$ es lineal respecto a $\vec{m'}$.

Haciendo la analogía respecto al problema de regresión, las muestras de entrada $\vec{x}$ son comparables con el conjunto de muestras de entrada $M' = \left\{\vec{m'}_1, \vec{m'}_2, \ldots, \vec{m'}_N\right\}$ para construir el modelo de clasificación, y por cada muestra a la salida real del fenómeno t es comparable con la etiqueta a asignar por muestra $\vec{q}$. El problema de la clasificación se puede enfocar entonces como la **discretización del problema de ajuste de curvas**, donde la salida t pasa a tener un número finito de valores K. Por ejemplo, si el objetivo es clasificar en $K = 2$ clases el conjunto de muestras M', $\vec{q} \in \{0, 1\}$. Si $K > 2$, la codificación de la etiqueta puede implementarse usando un esquema 1-de-*K*, el cual define a

$$\vec{q} = [q_0, q_1, \ldots, q_K],$$

poniendo en uno el valor q_j de la etiqueta j a representar y en cero los demás como un vector de largo K.

Por ejemplo, si la muestra $\vec{m}_i$ corresponde a la clase $k = 3$ de $K = 5$, se tiene que:

$$\vec{q} = [0, 0, 1, 0, 0],$$

Para el problema de clasificación en 2 clases, la función lineal discriminante más sencilla es dada por:

$$y(\vec{m}) = \vec{w'}^T \vec{m'} + w_0, \tag{2}$$

donde $\vec{w}$ es el vector de pesos, y w_0 **es el sesgo o su negativo es también llamado umbral**. Un vector de entrada $\vec{m}$ es asignado a la clase $\mathcal{C}_1$ si $y\left(\vec{m'}\right) \geq 0$ y a la clase $\mathcal{C}_2$ de otro modo. Lo anterior se resume como:

$$\begin{aligned} y\left(\vec{m'}\right) \geq 0 \quad \vec{m'} \in \mathcal{C}_1 \quad k = 0 \\ y\left(\vec{m'}\right) < 0 \quad \vec{m'} \in \mathcal{C}_2 \quad k = 1, \end{aligned}$$

donde k corresponde a la etiqueta de la clase. Como se puede observar en la Figura 1, el vector de pesos $\vec{w'}$ define la orientación del límite de decisión, y el negativo del sesgo w_0, la distancia respecto al origen.

Tomando en cuenta como **dos puntos** $\vec{m}_1$ y $\vec{m}_2$ los cuales están sobre la superficie de decisión, por lo que entonces

$$y\left(\vec{m'}_1\right) = y\left(\vec{m'}_2\right) = 0 \Rightarrow \vec{w'}^T \vec{m'_1} + w_0 - \vec{w'}^T \vec{m'_2} - w_0 = 0$$

$$\Rightarrow \vec{w}^T \left(\vec{m'_1} - \vec{m'_2}\right) = 0$$

lo cual significa que el vector $\vec{w}$ es perpendicular a todos los vectores paralelos o sobre la superficie de decisión, y además define la orientación de tal superficie, como lo muestra la Figura 1 con el vector verde.

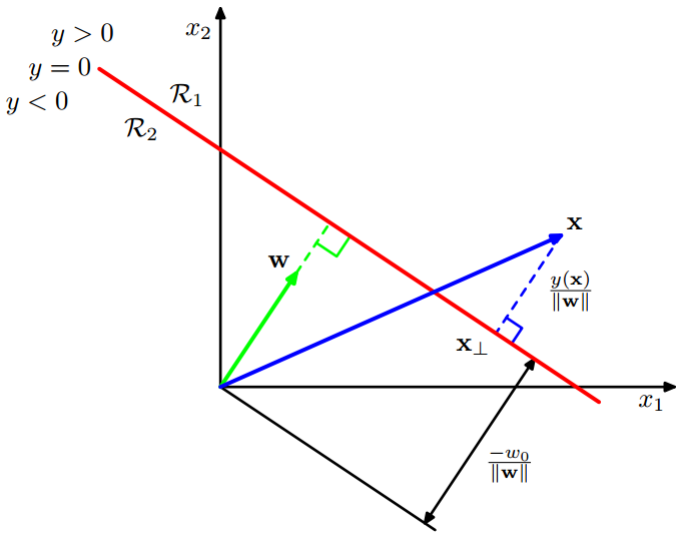

Figura 1: Geometría de la superficie de decisión con $D = 2$ definida por $y\left(\vec{m'}\right) = \vec{w'}^T\vec{m'} + w_0$, con $\mathbf{x} = \vec{m'}$ A partir de las ecuaciones de proyección y del hecho de que para los puntos en la superficie de decisión $y(\mathbf{x}) = 0$ [1].

De forma similar, si $\vec{m'}_1$ es un punto sobre la superficie de decisión, se tiene que:

$$y\left(\vec{m'}_1\right) = \vec{w'}^T\vec{m'}_1 + w_0 = 0$$

Para **simplificar la notación**, se incluye el sesgo, con un valor «tonto» $m_o = 1$, definiendo entonces $\vec{w} = [w_0, \vec{w}]^T$ y $\vec{m} = \left(m_0, \vec{m'}\right)$. Por ello, $\vec{w}, \vec{m} \in \mathbb{R}^{D+1}$. Es necesario entonces aumentar la dimensionalidad de las entradas a $D + 1$ para utilizar esta notación. Así entonces la expresión simplificada, la ecuación 2 se escribe como sigue:

$$y\left(\vec{m}\right) = \vec{w}^T\vec{m}. \tag{3}$$

1.1. Mínimos cuadrados para la clasificación en dos clases

Anteriormente se analizó el enfoque de mínimos cuadrados para el ajuste de curvas, el cual proponía una solución analítica cerrada a partir de la derivada de la función de error. Para el problema de clasificación en dos clases, la minimización del error cuadrático se hace respecto a los N pares ordenados de entrenamiento $\mathcal{D} = \{M, T\}$, con $M = \{\vec{m}_1, \ldots, \vec{m}_N\}$ y $T = \{t_1, \ldots, t_N\}$, pues recordemos que $t_j \in \{0, 1\}$ en el caso de la clasificación por dos clases $K = 2$. Generalizando para la matriz de muestras $M \in \mathbb{R}^{N\times(D+1)}$ y $\vec{w} \in \mathbb{R}^{(D+1)\times 1}$,

donde los valores de cada muestra $\vec{m}_j$ se representan en la fila j de tal matriz a la ecuación 3, se obtiene:

$$y\left(M\right) = M\,\vec{w}.$$

$$\Rightarrow y\left(M\right) = \begin{bmatrix} - & \vec{m}_1 & - \\ \vdots & \vdots & \vdots \\ - & \vec{m}_N & - \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}.$$

Así, respecto a la matriz con todos los valores conocidos de pertenencia de clase $\vec{t} = \begin{bmatrix} t_1 & \dots & t_N \end{bmatrix}^T$, con $t_j \in \{0, 1\}$, el error cuadrático en forma matricial se expresa como sigue:

$$E\left(\vec{w}\right) = \frac{1}{2}\left\|M\,\vec{w} - \vec{t}\right\|^2 = \frac{1}{2}\left\{\left(M\,\vec{w} - \vec{t}\right)^T\left(M\,\vec{w} - \vec{t}\right)\right\}. \tag{4}$$

La función gradiente respecto al vector de pesos $\vec{w}$ de la función de error 4 igualada a cero, viene dada por (según lo demostrado para el caso de la regresión):

$$\frac{\partial E(\vec{w})}{\partial \vec{w}} = M^T M\,\vec{w} - M^T\vec{t} = 0, \tag{5}$$

y despejando $\vec{w}$ para obtener el vector de pesos que logra el error mínimo:

$$\vec{w} = \left(M^T M\right)^{-1} M^T\vec{t}, \tag{6}$$

Así, la clasificación de una nueva muestra $\vec{h} \in \mathbb{R}^{D+1}$ viene dada por:

$$y\left(\vec{h}\right) = \vec{h}\,\left(\left(M^T M\right)^{-1} M^T\vec{t}\right).$$

Para simplificar la clasificación en dos clases, la superficie de decisión queda definida por $y\left(\vec{m}\right) = w_0$, por lo que si $y\left(\vec{m}\right) \geq w_0$, $\vec{m} \in \mathcal{C}_1$, de lo contrario $\vec{m} \in \mathcal{C}_2$.

Para el caso específico de la clasificación en un espacio $\mathbb{R}^2$, $(D = 2)$, se tiene que cada muestra está dada por $\vec{m} = \begin{bmatrix} m_0 = 1 & m_1 & m_2 \end{bmatrix}^T$ y el modelo está compuesto por los pesos $\vec{w} = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix}^T$, por lo que entonces el modelo está dado por:

$$y\left(\vec{m}\right) = w_0 + w_1 m_1 + w_2 m_2$$

lo cual es un plano en $\mathbb{R}^2$, para lo cual es necesario graficar la curva de nivel, para lo cual se toma el valor en $y\left(\vec{m}\right) = 0$, con lo que:

$$-w_0 = w_1 m_1 + w_2 m_2$$

y para la graficación de la curva, es necesario entonces despejar alguna de las variables:

$$\frac{-w_0 - w_1 m_1}{w_2} = m_2$$

La clasificación por mínimos cuadrados presenta una **importante debilidad**: al castigar la distancia euclidiana de las muestras respecto al modelo,

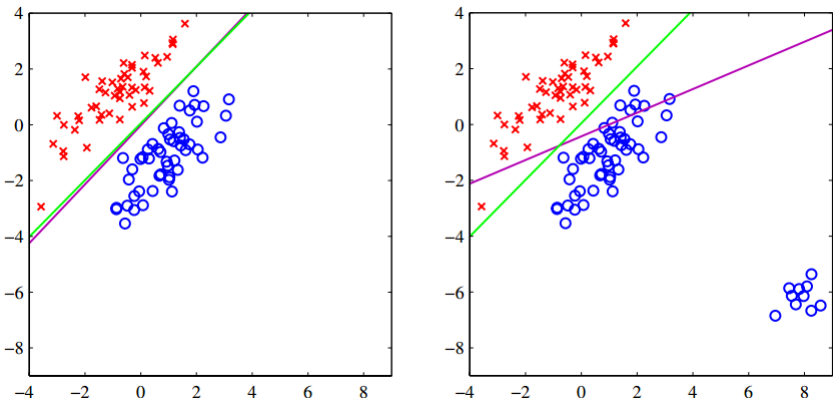

Figura 2: Sesgo de un modelo obtenido por mínimos cuadrados [1].

la exactitud de tal modelo es perjudicada cuando se presentan sesgos en las muestras, es decir, muestras muy desviadas del comportamiento usual de la clase. Incluso si estos sesgos son «muy correctos» (muestras muy alejadas de los montículos principales), el modelo es afectado, como se observa en la Figura 2.

1.2. El discriminante lineal de Fisher

El discriminante lineal de Fisher enfoca el problema de clasificación como un problema de **reducción de la dimensionalidad**. Una función lineal de cada una de las $D + 1 = D'$ dimensiones de la muestra $\vec{w}, \vec{m} \in \mathbb{R}^{D' \times 1}$ puede interpretarse como una función que proyecta tal muestra a un espacio de dimensión uno:

$$\check{m} = y\left(\vec{m}\right) = \vec{w}^T \vec{m}, \tag{7}$$

donde $\check{m} \in \mathbb{R}$ y en general el símbolo del «sombrero invertido» denota una reducción de dimensionalidad a $N = 1$.

En general, la proyección de un espacio de dimensión D a uno de menor dimensión $D_u < D'$ para un conjunto de datos implica pérdida de información, lo que dificulta muchas veces una clasificación exitosa de los datos, como se puede observar en la Figura 3, a la izquierda.

Los datos antes fácilmente separables en dos dimensiones, pueden llegar a traslaparse en mayor medida, dependiendo de la **perspectiva de la proyección**. La separabilidad y por ende la facilidad de clasificar los datos depende de la posición del plano proyectado, definido por el vector de pesos $\vec{w}$, como también se puede observar en la Figura 3.

El objetivo de la función discriminante lineal de Fisher es maximizar la distancia entre las medias proyectadas μ_1 y μ_2 de las clases C_1y C_2, con N_1 y N_2

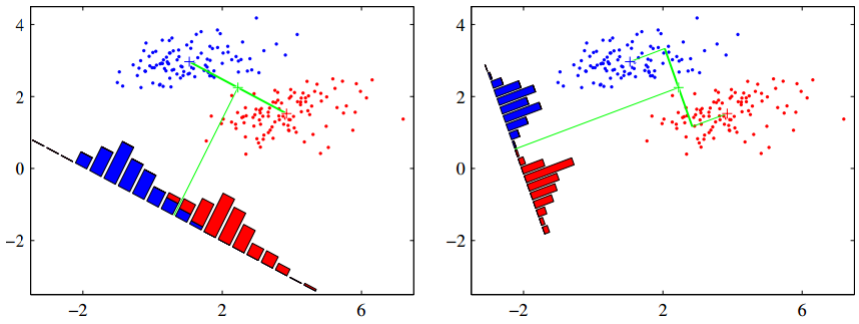

Figura 3: Proyección de un espacio de dimensión $D = 2$ a uno de dimensión $D = 1$, con diferentes perspectivas [1].

muestras respectivamente. Las medias geométricas de cada clase son determinadas a partir de las muestras de entrenamiento $\vec{m}_n$:

$$\vec{\mu}_1 = \frac{1}{N_1} \sum_{n \in C_1} \vec{m}_n, \qquad \vec{\mu}_2 = \frac{1}{N_2} \sum_{n \in C_2} \vec{m}_n, \tag{8}$$

donde $\vec{\mu}_1, \vec{\mu}_2 \in \mathbb{R}^{D' \times 1}$, y con la proyección de la diferencia de las medias dada por:

$$\check{\mu}_2 - \check{\mu}_1 = \vec{w}^T \left(\vec{\mu}_1 - \vec{\mu}_2 \right), \tag{9}$$

puesto que μ_k son también muestras de dimensionalidad N, y su proyección en general viene dada por

$$\check{\mu}_k = \vec{w}^T \vec{\mu}_k. \tag{10}$$

Basado en la ecuación 9, podemos escribir la siguiente función a maximizar:

$$J_\mu \left(\vec{w} \right) = \left(\check{\mu}_2 - \check{\mu}_1 \right)^2 = \left(\vec{w}^T \vec{\mu}_1 - \vec{w}^T \vec{\mu}_2 \right) \left(\vec{w}^T \vec{\mu}_1 - \vec{w}^T \vec{\mu}_2 \right)^T = \vec{w}^T \left(\vec{\mu}_1 - \vec{\mu}_2 \right) \left(\vec{\mu}_1 - \vec{\mu}_2 \right)^T \vec{w}, \tag{11}$$

y al usar notación matricial , la ecuación 11 equivale a:

$$J_\mu \left(\vec{w} \right) = \vec{w}^T S_B \vec{w}, \tag{12}$$

donde S_B se refiere a la matriz $D \times D$ de **covarianza inter-clase**, basada en las medias de las dos clases (recuerde que el factor a la derecha transpuesto genera un producto externo):

$$S_B = \left(\vec{\mu}_1 - \vec{\mu}_2 \right) \left(\vec{\mu}_1 - \vec{\mu}_2 \right)^T. \tag{13}$$

Sin embargo, la Figura 3 también muestra que maximizar la distancia entre las medias no es suficiente para garantizar la separabilidad de los datos en un espacio de dimensionalidad reducida, dado el aún posible traslape en la proyección de las poblaciones ocasionada por altas varianzas. La maximización J_μ según la ecuación 12 puede arribar a un conjunto de pesos $\vec{w}$ incorrecto sino se

toma en cuenta la **varianza intra-clase,** es decir, la varianza de las muestras en cada clase.

El objetivo entonces de la función de costo J es: **maximizar la varianza inter-clase** (distancia entre las medias de las poblaciones) y **minimizar la varianza intra-clase** (varianza de las muestras en la población de cada clase).

La varianza intra-clase proyectada a minimizar $\check{s}_k$ de una clase C_k, se define como la diferencia al cuadrado respecto a su media proyectada μ_k de todas las muestras proyectadas $\check{m}_n = \vec{w}^T \vec{m}_n$:

$$\check{s}_k = \sum_{n \in C_k} \left(\check{m}_n - \check{\mu}_k\right)^2 = \sum_{n \in C_k} \left(\vec{w}^T \vec{m}_n - \vec{w}^T \vec{\mu}_k\right) \left(\vec{w}^T \vec{m}_n - \vec{w}^T \vec{\mu}_k\right)^T, \tag{14}$$

$$\Rightarrow \check{s}_k = \vec{w}^T \left(\sum_{n \in C_k} \left(\vec{m}_n - \vec{\mu}_k\right) \left(\vec{m}_n - \vec{\mu}_k\right)^T \right) \vec{w}, \tag{15}$$

por lo que entonces para el caso de dos clases, la **varianza intra-clase total** viene dada por $\check{s}_W = \check{s}_1 + \check{s}_2$. En el espacio de dimensionalidad D', la ecuación 14 se expresa en términos de las **matrices de covarianza** S_i:

$$S_i = \sum_{n \in C_i} \left(\vec{m}_n - \vec{\mu}_i\right) \left(\vec{m}_n - \vec{\mu}_i\right)^T, \tag{16}$$

donde entonces $S_w = S_1 + S_2$. Para expresar la varianza proyectada $\check{s}_k^2$ en el espacio original de dimensión N, tomamos las ecuaciones 7 , 10 y 16 para obtener la siguiente equivalencia:

$$\check{s}_w = \check{s}_1 + \check{s}_2 = \vec{w}^T S_w \vec{w}.$$

La **función a maximizar** J, tanto en el espacio proyectado y el espacio original de dimensión D', viene dada por:

$$J\left(\vec{w}\right) = \frac{\left(\check{\mu}_2 - \check{\mu}_1\right)^2}{\check{s}_1 + \check{s}_2} = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_w \vec{w}}. \tag{17}$$

Un elemento importante a notar respecto a la función a maximizar 17, es el hecho de que respecto a el vector de pesos $\vec{w}$, nos interesa su **dirección u orientacion, y no su magnitud,** pues este primer atributo es el que definirá la separabilidad de los datos en la proyección a realizar. Tomando en cuenta tal aspecto, nos prestamos a derivar e igualar respecto a cero, la función J respecto a $\vec{w}$ (recordando que $\nabla_{\vec{x}} \left(\vec{x}^T A \vec{x}\right) = 2 A \vec{x}$):

$$\nabla_{\vec{w}} J = \frac{\left(2 S_B \vec{w}\right) \left(\vec{w}^T S_w \vec{w}\right) - \left(\vec{w}^T S_B \vec{w}\right) \left(2 S_w \vec{w}\right)}{\left(\vec{w}^T S_w \vec{w}\right)^2} = 0,$$

$$\Rightarrow \nabla_{\vec{w}} J = \frac{\left(S_B \vec{w}\right) \cancel{\left(\vec{w}^T S_w \vec{w}\right)}}{\left(\vec{w}^T S_w \vec{w}\right)^{\cancel{2}}} = \frac{\left(\vec{w}^T S_B \vec{w}\right) \left(S_w \vec{w}\right)}{\left(\vec{w}^T S_w \vec{w}\right)^2},$$

tomando en cuenta que las formas cuadráticas son escalares y multiplicando por la inversa de S_w:

$$S_w^{-1} S_B \vec{w} = \left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_w \vec{w}} \right) \vec{w}, \tag{18}$$

Si S_W tiene una inversa calculable S_W^{-1}, entonces multiplicando tal factor a ambos lados de la ecuación 18 se obtiene que:

$$S_w^{-1} S_B \vec{w} = \left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_w \vec{w}} \right) \vec{w}, \tag{19}$$

donde el problema de encontar $\vec{w}$ en la ecuación 18 se puede interpretar como el de calcular el **auto-vector** $\vec{w}$ para el **autovalor** λ**:**

$$\lambda = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_w \vec{w}}$$

y la matriz $A = S_w^{-1} S_B$, en la conocida ecuación para los auto-vectores y auto-valores:

$$A\vec{w} = \lambda \vec{w}.$$

Recordando el concepto de auto-vector y auto-valor: El auto-vector $\vec{w}$ de dimensión $D' \times 1$, para la matriz cuadrada A de dimensiones $D' \times D'$, corresponde a un vector que no cambia su dirección cuando se aplica la transformación especificada en la matriz A, $T(\vec{w}) = A\,\vec{w}$. El vector transformado $A\vec{w}$ es entonces paralelo al vector $\vec{w}$. Los auto-vectores de una matriz de covarianza como S_w resultan en las direcciones principales de la varianza, donde la predominancia de un auto-vector es definida por el auto-valor λ.

Continuando con el problema , tal y como se mencionó anteriormente, lo importante en el caso del vector $\vec{w}$ es su dirección, y no su magnitud, por lo que entonces podemos simplificar el problema haciendo

$$\lambda = 1.$$

Tomando en cuenta la ecuación 13, podemos entonces desarrollar la ecuación 19:

$$S_w^{-1} \left(\vec{\mu}_1 - \vec{\mu}_2 \right) \left(\vec{\mu}_1 - \vec{\mu}_2 \right)^T \vec{w} = \lambda \vec{w}, \tag{20}$$

donde el vector $\left(\mu_1 - \mu_2 \right) \left(\mu_1 - \mu_2 \right)^T \vec{w}$ tiene siempre la dirección $S_w^{-1} \left(\mu_1 - \mu_2 \right)$, puesto que $\left(\vec{\mu}_1 - \vec{\mu}_1 \right)^T \vec{w}$ es un escalar, y recordando que es posible hacer λ=1, la ecuación 20 nos permite arribar a la orientación del vector de pesos $\vec{w}$:

$$\vec{w} \propto S_w^{-1} \left(\vec{\mu}_1 - \vec{\mu}_1 \right). \tag{21}$$

Por medio de la ecuación 21 es posible encontrar un vector de pesos $\vec{w}$ que **maximiza la varianza inter-clase y minimiza la varianza intra-clase**, en la proyección de las muestras $\check{m} = y\left(\vec{m}\right) = \vec{w}^T \vec{m}$, con lo cual es relativamente sencillo a través de un análisis Gaussiano, encontrar el umbral para clasificar una muestra como perteneciente a $\mathcal{C}_1$ o a $\mathcal{C}_2$.

1.3. El perceptrón

El algoritmo del preceptrón propuesto por [2], es uno de los algoritmos de clasificación que marcó el desarrollo del reconocimiento de patrones, sirviendo de fundamento para el área de aprendizaje automático por redes neuronales. El perceptrón se basa en la abstracción matemática de una **neurona a nivel biológico**. Una neurona tiene una serie de entradas $\vec{m} = \begin{bmatrix} m_0 & \dots & m_D \end{bmatrix}^T$ las cuales son combinadas linealmente por un vector de pesos o de **enlaces sinápticos** $\vec{w} = \begin{bmatrix} w_0 & \dots & w_D \end{bmatrix}^T$ en la **función de red**:

$$\Upsilon(\vec{m}) = \sum_{i=0}^{D} m_i w_i = \vec{w}^T \vec{m}.$$

Lo cual anteriormente correspondía a la función $y = \Upsilon$. Una neurona es excitada según una **función de activación** $f(\Upsilon)$, que recibe como entrada el resultado de la función de red Υ, y puede corresponder por ejemplo a una función de tipo escalón:

$$f(\vec{m}) = \begin{cases} +1 & \Upsilon(\vec{m}) \geq 0 \\ -1 & \Upsilon(\vec{m}) < 0 \end{cases}.$$

Para el problema de clasificación en dos clases, las etiquetas de las clases $\mathcal{C}_1$ y $\mathcal{C}_2$ corresponderían a la salida de la función de excitación $t = y(\vec{m})$, por lo que entonces

$$t \in \{-1, 1\}$$

, con $t = 1$ si $t \in \mathcal{C}_1$ y $t = -1$ si $t \in \mathcal{C}_2$. El perceptrón es un algoritmo supervisado de clasificación, en el que debe estar especificado para el conjunto M de N muestras de entrenamiento

$$M = \begin{bmatrix} - & \vec{m}_1 & - \\ \vdots & \vdots & \vdots \\ - & \vec{m}_N & - \end{bmatrix},$$

sus etiquetas de pertenencia de clase correspondientes

$$\vec{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}.$$

La Figura 4 ilustra el concepto del perceptrón.

Para la determinación de los pesos $\vec{w}$, se utiliza el **criterio del perceptrón,** el cual establece que una muestra $\vec{m}_j$ de dimensionalidad N es correctamente clasificada si cumple que:

$$\Upsilon(\vec{m}_j)\ t_j = \vec{w}^T \vec{m}_j\, t_j > 0,$$

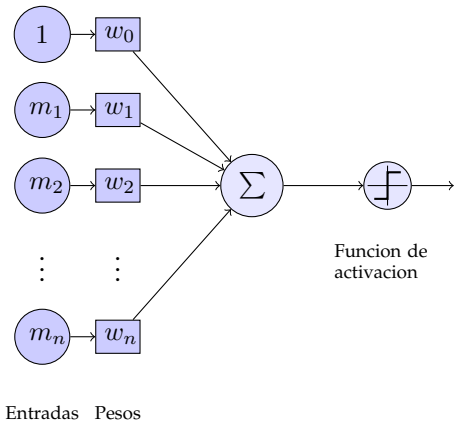

Figura 4: Diagrama conceptual de un perceptrón, abstracción del concepto biológico de neurona.

donde t_j es la etiqueta correcta de la muestra (pertenencia a la clase), previamente conocida. El criterio del perceptrón especifica que el error total toma en cuenta únicamente las muestras incorrectamente clasificadas ($\vec{w}^T \vec{m}_j \, t_j \leq 0$) las cuales están contenidas en el conjunto de muestras incorrectamente clasificadas $\mathcal{U}$, como lo muestra la función de error $E_P(\vec{w})$:

$$E_P(\vec{w}) = -\sum_{\vec{m}_j \in \mathcal{U}} \left(\vec{w}^T \vec{m}_j \, t_j\right).$$

De este modo, si todas las muestras son correctamente clasificadas, entonces $E_P(\vec{w}) = 0$.

El algoritmo de entrenamiento usa el criterio del perceptrón para evaluar el error $E_P(\vec{w})$ en P iteraciones, donde por cada iteración τ, se evalúa el error para todas las muestras $\vec{m}_j$ en las clases C_1 y C_2, y se modifica el vector de pesos $\vec{w}$ como sigue:

$$\vec{w}(\tau+1)^T = \vec{w}(\tau)^T - \alpha \, \nabla E_P(\vec{w}) = \vec{w}(\tau)^T + \alpha \begin{cases} \sum_{\vec{m}_j \in \mathcal{U}} \vec{m}_j \, t_j & \vec{m}_j \in \mathcal{U} \\ 0 & \vec{m}_j \notin \mathcal{U} \end{cases}.$$

El coeficiente α es referido como el **coeficiente de aprendizaje**, y define la tasa de cambio del vector de pesos $\vec{w}$, entre las iteraciones τ y $\tau+1$. **El teorema de la convergencia** del perceptrón establece que el vector de pesos $\vec{w}$ **converge** después de un número finito de iteraciones [2], aún cuando un cambio en tal vector involucre uno o más nuevos errores en muestras clasificadas correctamente en iteraciones anteriore únicamente si el conjunto de muestras a clasificar es **linealmente seperable**.

El proceso de entrenamiento del perceptrón puede interpretarse en el espacio vectorial como el corrimiento gradual del hiper-plano $\vec{w}$ según las muestras

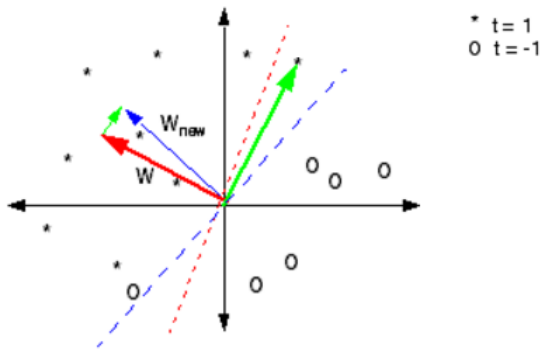

Figura 5: Interpretación en el espacio vectorial del vector de pesos **w**.

incorrectamente clasificadas, como se observa en la Figura 5. Con tal interpretación vectorial es posible apreciar el **impacto del valor inicial del vector de pesos** $\vec{w}(1)$ en los valores finales al final del entrenamiento y la cantidad de iteraciones P necesarias para lograr la convergencia del algoritmo.

El algoritmo del perceptrón tiene la ventaja respecto al enfoque de mínimos cuadrados de ser menos sensible a las muestras muy lejanas, datos aislados u *outliers*.

1.3.1. Ejemplo de ejecución del perceptrón

Tómese un ejemplo con coeficiente de aprendizaje $\alpha = 1$ en el que existen las siguientes muestras para la clase 1:

$$\begin{aligned} \vec{x}_1 &= \begin{bmatrix} 1 & 7{,}0639 & 3{,}9717 \end{bmatrix}^T \quad t_1 = -1 \\ \vec{x}_2 &= \begin{bmatrix} 1 & 10{,}4392 & 4{,}7711 \end{bmatrix}^T \quad t_2 = -1 \end{aligned}$$

y para la clase 2:

$$\begin{aligned} \vec{x}_3 &= \begin{bmatrix} 1 & 6{,}3433 & 8{,}7337 \end{bmatrix}^T \quad t_3 = 1 \\ \vec{x}_4 &= \begin{bmatrix} 1 & 1{,}0321 & 7{,}1919 \end{bmatrix}^T \quad t_4 = 1 \end{aligned}$$

Para la primer iteración, se inicializa el vector de pesos como (usualmente se inicializa aleatoriamente), incluyendo la concatenación del valor *tonto*:

$$w(1) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$$

Por lo que entonces, para **la iteración** $\tau = 2$ se recorren todas las muestras, primero evaluando la función de red:

1. Para la muestra $\vec{x}_1$:

$$\gamma(\vec{x}_1) = x_{1,0}w_0 + x_{1,1}w_1 + x_{1,2}w_2 = 1 + 7{,}0639 + 3{,}9717 = 12{,}0356$$

(Loading...)

Figura 6: Funcionamiento del algoritmo del perceptrón.

revisando la condición de preservación de los pesos:

$$\gamma\left(\vec{x}_1\right)\ t_1 > 0 \Rightarrow 12{,}0356 \cdot (-1) < 0$$

verificamos que no se cumple, por lo que se actualizan los pesos haciendo:

$$\vec{w}(\tau+1)^T = \vec{w}(\tau)^T + \alpha \vec{m}_j\, t_j$$

$$\Rightarrow \vec{w}(\tau+1)^T = \begin{bmatrix}1\\1\\1\end{bmatrix} + 1 \begin{bmatrix}1\\7{,}0639\\3.9717\end{bmatrix} (-1) = \begin{bmatrix}0\\-6{,}0639\\-2{,}9717\end{bmatrix}.$$

2. Para la muestra $\vec{x}_2$:

   $\gamma\left(\vec{x}_2\right) = x_{2,0}w_0 + x_{2,1}w_1 + x_{2,2}w_2 = 0 + (-6{,}0639)\,(10{,}4392) + (-2{,}9717)\,(4{,}7711) = -77{,}4805$

   y verificando la condición de preservación de los pesos:

   $$\gamma\left(\vec{x}_2\right)\ t_2 > 0 \Rightarrow 77{,}4805 > 0$$

   se cumple, por lo que el vector de pesos se preserva.

3. Para la muestra $\vec{x}_3$:

   $\gamma\left(\vec{x}_3\right) = x_{3,0}w_0 + x_{3,1}w_1 + x_{3,2}w_2 = 0 + (-6{,}0639)\,(6{,}3433) + (-2{,}9717)\,(8{,}7337) = -64{,}4191$

revisando la condición de preservación de los pesos:

$$\gamma\left(\vec{x}_3\right)\ t_3 > 0 \Rightarrow -64{,}4191 \cdot (1) < 0$$

observamos que no se cumple tal condición, por lo que nos disponemos a cambiar el arreglo de pesos haciendo:

$$\vec{w}(\tau+1)^T = \vec{w}(\tau+1)^T + \alpha \vec{m}_j\, t_j$$

$$\Rightarrow \vec{w}(\tau+1)^T = \begin{bmatrix} 0 \\ -6{,}0639 \\ -2{,}9717 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 6{,}3433 \\ 8{,}7337 \end{bmatrix} (1) = \begin{bmatrix} 1 \\ 0{,}2794 \\ 5{,}7621 \end{bmatrix}.$$

4. Para la muestra $\vec{x}_4$:

   $\gamma\left(\vec{x}_4\right) = x_{4,0}w_0 + x_{4,1}w_1 + x_{4,2}w_2 = 1 + 0{,}2794 \cdot 1{,}0321 + 5{,}7621 \cdot 7{,}1919 = 42{,}7288$

   revisando la condición de preservación de los pesos:

   $$\gamma\left(\vec{x}_4\right)\ t_4 > 0 \Rightarrow 42{,}7288 \cdot (1) > 0$$

   se cumple, por lo que el vector de pesos se preserva. Para esta iteración acabamos con el arreglo:

$$\Rightarrow \vec{w}(2)^T = \begin{bmatrix} 1 \\ 0{,}2794 \\ 5{,}7621 \end{bmatrix}.$$

Nos disponemos a realizar **la iteración** $\tau = 3$:

1. Para la muestra $\vec{x}_1$:

   $\gamma\left(\vec{x}_1\right) = x_{1,0}w_0 + x_{1,1}w_1 + x_{1,2}w_2 = 1 + 0{,}2794 \cdot 7{,}0639 + 5{,}7621 \cdot 3{,}9717 = 25{,}8589$

   revisando la condición de preservación de los pesos:

   $$\gamma\left(\vec{x}_1\right)\ t_1 > 0 \Rightarrow 25{,}8589\,(-1) < 0$$

   verificamos que no se cumple, por lo que se actualizan los pesos haciendo:

$$\vec{w}(\tau+1)^T = \vec{w}(\tau)^T + \alpha \vec{m}_j\, t_j$$

$$\Rightarrow \vec{w}(\tau+1)^T = \begin{bmatrix} 1 \\ 0{,}2794 \\ 5{,}7621 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 7{,}0639 \\ 3{,}9717 \end{bmatrix} (-1) = \begin{bmatrix} 0 \\ -6{,}7844 \\ 1{,}7904 \end{bmatrix}.$$

2. Para la muestra $\vec{x}_2$:

   $\gamma\left(\vec{x}_2\right) = x_{2,0}w_0 + x_{2,1}w_1 + x_{2,2}w_2 = 0 + -6{,}7844 * 10{,}4392 + 1{,}7904 * 4{,}7711 = -62{,}2824$

   y verificando la condición de preservación de los pesos:

   $$\gamma\left(\vec{x}_2\right)\ t_2 > 0 \Rightarrow (-62{,}2824)\,(-1) > 0$$

   se cumple, por lo que el vector de pesos se preserva.

3. Para la muestra $\vec{x}_3$:

   $\gamma(\vec{x}_3) = x_{3,0}w_0 + x_{3,1}w_1 + x_{3,2}w_2 = 0 + -6{,}7844 * 6{,}3433 + 1{,}7904 * 8{,}7337 = -27{,}3991$

   revisando la condición de preservación de los pesos:

   $$\gamma(\vec{x}_3)\ t_3 > 0 \Rightarrow -27{,}3991 \cdot (1) < 0$$

   observamos que no se cumple tal condición, por lo que nos disponemos a cambiar el arreglo de pesos haciendo:

   $$\vec{w}(\tau+1)^T = \vec{w}(\tau+1)^T + \alpha \vec{m}_j\, t_j$$

   $$\Rightarrow \vec{w}(\tau+1)^T = \begin{bmatrix} 0 \\ -6{,}7844 \\ 1{,}7904 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 6{,}3433 \\ 8{,}7337 \end{bmatrix} (1) = \begin{bmatrix} 1 \\ -0{,}4412 \\ 10{,}5241 \end{bmatrix}.$$

4. Para la muestra $\vec{x}_4$:

   $\gamma(\vec{x}_4) = x_{4,0}w_0 + x_{4,1}w_1 + x_{4,2}w_2 = 1 + (-0{,}4412) \cdot 1{,}0321 + (10{,}5241) \cdot 7{,}1919 = 42{,}7288$

   revisando la condición de preservación de los pesos:

   $$\gamma(\vec{x}_4)\ t_4 > 0 \Rightarrow 42{,}7288 \cdot (1) > 0$$

   se cumple, por lo que el vector de pesos se preserva. Para esta iteración acabamos con el arreglo:

   $$\Rightarrow \vec{w}(2)^T = \begin{bmatrix} 1 \\ -0{,}4412 \\ 10{,}5241 \end{bmatrix}.$$

El lector puede realizar la cuarta iteración, y confirmar que para la quinta, no el vector $\vec{w}$ no cambia.

1.4. Bayes ingenuo y regresión logística

1.4.1. Bayes Ingenuo

El problema de clasificación se puede modelar como el problema de estimar una función de densidad. Por ejemplo, el clasificador de **Bayes ingenuo** estima una función de densidad condicional, como sigue. Suponga que estamos ante el problema de clasificar imágenes vectorizadas binarias $\vec{m} \in \mathbb{R}^D$ en dos categorías, correspondientes a si simbolizan el digito 1 o 2. El objetivo del Bayes ingenuo es entonces **estimar las siguientes funciones de densidad de probabilidad condicional:**

$$p(t = 1|\overrightarrow{m})$$
$$p(t = 2|\overrightarrow{m})$$

lo cual se lee como la probabilidad de que la etiqueta sea 1 dada la muestra $\overrightarrow{m}$, y la etiqueta sea 2 dada esa muestra $\overrightarrow{m}$, respectivamente. Para hacer esto, usamos el teorema de Bayes:

$$p\left(t|\overrightarrow{m}\right) = \frac{p\left(\overrightarrow{m}|t\right)p\left(t\right)}{p\left(\overrightarrow{m}\right)} = \frac{p\left(m_0,\ldots,m_D|t\right)p\left(t\right)}{p\left(m_0,\ldots,m_D\right)}$$

lo cual implica que las probabilidades condicionales se expresan en términos de las probabilidades conjuntas. En este caso particular, cada componente m_i corresponde a un pixel en la imagen, con lo cual por ejemplo $p\left(m_1|t\right)$ se lee como la probabilidad de que el pixel 1 esté encendido dado que la muestra tiene la etiqueta t. La asunción clave del bayes ingenuo, **es la independencia de todas las variables aleatorias, en este caso, los pixeles**, por lo que la ecuación anterior se reescribe como sigue:

$$p\left(t|\overrightarrow{m}\right) = \frac{p\left(m_0|t\right)p\left(m_1|t\right)\ldots p\left(m_D|t\right)p\left(t\right)}{p\left(m_0\right)p\left(m_1\right)\ldots p\left(m_D\right)} = \frac{\prod_{i=0}^{D} p\left(m_i|t\right)p\left(t\right)}{\prod_{i=0}^{D} p\left(m_i\right)}$$

Observe que en general, la sumatoria de las probabilidades condicionales para todas las etiquetas K, para una muestra $\overrightarrow{m}$ es

$$\sum_{j}^{K} p\left(t=j|\overrightarrow{m}\right) = 1 = \sum_{j}^{K} \frac{\prod_{i=0}^{D} p\left(m_i|t=j\right)p\left(t\right)}{\prod_{i=0}^{D} p\left(m_i\right)} = 1$$

donde el denominador es el mismo para todos los términos, al ser independiente de la etiqueta, por lo que podemos prescindir de él y encontrar una relación proporcional:

$$p\left(t|\overrightarrow{m}\right) \propto \prod_{i=0}^{D} p\left(m_i|t\right)p\left(t\right)$$

Las probabilidades marginales para las etiquetas $p\left(t\right)$ se calculan, por ejemplo para el caso de la estimación del digito que aparece en la imagen, si se cuenta con $M = 10000$, como la proporción entre la cantidad de muestras de cada clase en tal conjunto de datos, por ejemplo si la mitad de las muestras son de un dígito y la otra mitad de otro, entonces $p\left(t=1\right) = p\left(t=2\right) = \frac{5000}{10000} = 0{,}5$, por lo que se dice que las clases en ese conjunto de datos están **balanceadas**. Las probabilidades condicionales $p\left(m_i|t\right)$ se calculan de acuerdo a la cantidad de veces en las que el pixel m_i está encendido, cuando su etiqueta en el conjunto de datos es t. Algunas veces, pixeles específicos puede que nunca se enciendan en un conjunto de muestras para una clase específica t, lo cuál haría que su probabilidad condicional sea nula, $p\left(m_i|t\right) = 0$, por lo cual se agrega un número bajo de ocurrencias ϵ, para evitar que al ejecutar la multiplicatoria de las probabilidades condicionales, el cálculo se anule. A esto se le conoce como **suavizado laplaciano**.

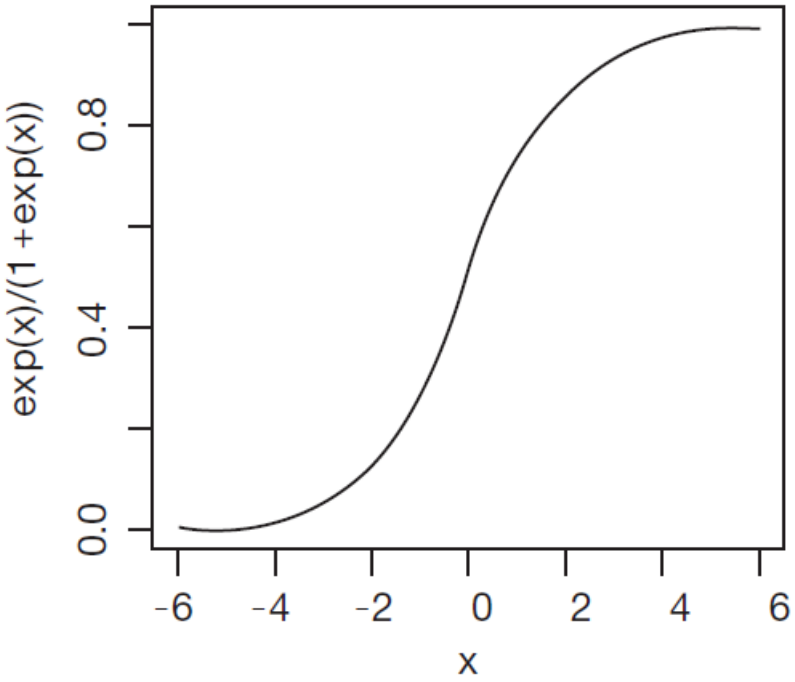

Figura 7: Función sigmoidal utilizada en la regresión logística.

1.4.2. Regresión logística

El enfoque de estimar una función de densidad cambia el problema a una regresión, donde la salida del modelo y puede tomar un valor real en el intervalo de 0 a 1, por lo que entonces $y \in \{0, 1\}$. De esta forma, si $y \geq 0{,}5 \Rightarrow t = 1$, y de lo contrario $t = 0$.

Para una muestra $\vec{m} \in \mathbb{R}^D$, la **regresión logística** implementa un modelo basado en la función sigmoidal, por lo que el mismo viene dado por:

$$y\left(\overrightarrow{m}, \overrightarrow{w}\right) = \frac{1}{1 + e^{-\vec{w}^T \overrightarrow{m}}} = \frac{1}{1 + e^{-\gamma(\overrightarrow{m})}}$$

donde el vector $\vec{w}$ corresponde al vector de parámetros o pesos a ajustar, y $\gamma\left(\overrightarrow{m}\right) = \vec{w}^T \overrightarrow{m} = \vec{w} \cdot \overrightarrow{m}$ corresponde a un producto punto, o la **función de peso neto**, y la función $y\left(\overrightarrow{m}\right)$ corresponde a la **función de activación**, en analogía con el algoritmo del perceptrón. La Figura 7 grafica tal función.

La función sigmoidal en general:

$$y\left(x\right) = \frac{1}{1 + e^{-x}}$$

es una función suave y continua acotada entre 0 y 1, por lo que $0 \leq y\left(x\right) \leq 1$,

que además, tiene la propiedad de que la primera derivada está dada por

$$\frac{\mathrm{d}}{\mathrm{d}x}\text{sigmoid}\left(x\right) = \frac{e^{-x}}{\left(1+e^{-x}\right)^2} = \frac{1}{\left(1+e^{-x}\right)}\frac{e^{-x}}{\left(1+e^{-x}\right)} \tag{22}$$

donde tomando el término derecho de tal multiplicación:

$$\frac{e^{-x}}{\left(1+e^{-x}\right)} = \frac{1+e^{-x}-1}{\left(1+e^{-x}\right)}$$
$$\Rightarrow \frac{1+e^{-x}}{\left(1+e^{-x}\right)} - \frac{1}{\left(1+e^{-x}\right)} = \left(1-\frac{1}{\left(1+e^{-x}\right)}\right),$$

por lo que entonces la ecuación 22 se puede reescribir como:

$$\frac{\mathrm{d}}{\mathrm{d}x}\text{sigmoid}\left(x\right) = \frac{e^{-x}}{\left(1+e^{-x}\right)^2} = \frac{1}{\left(1+e^{-x}\right)}\left(1-\frac{1}{\left(1+e^{-x}\right)}\right)$$

lo cual significa que:

$$\frac{\mathrm{d}}{\mathrm{d}x}\text{sigmoid}\left(x\right) = \text{sigmoid}\left(x\right)\left(1-\text{sigmoid}\left(x\right)\right) \tag{23}$$

Por lo que entonces la **derivada de forma más compacta** se escribe como:

$$\frac{d}{dz}y\left(z\right) = y\left(z\right)\left(1-y\left(z\right)\right).$$

Volviendo al problema de clasificación, este se reduce a encontrar el vector de parámetros $\vec{w}$ que **minimice el error de clasificación para un conjunto** $\mathcal{D} = \{M, T\}$. Para ello utilizaremos un enfoque de **maximización de la verosimilitud**, por lo que entonces interpretamos la probabilidad de obtener la etiqueta $t = 1$ dada una muestra $\vec{m}$ y un vector de parámetros $\vec{w}$ como:

$$p\left(t=1|\vec{m},\vec{w}\right) = y\left(\vec{m},\vec{w}\right)$$

y de obtener la etiqueta $t = 0$ con el complemento:

$$p\left(t=0|\vec{m},\vec{w}\right) = 1-y\left(\vec{m},\vec{w}\right)$$

lo anterior es posible pues la función sigmoidal cumple las propiedades básicas de una función de densidad de probabilidad. De forma más compacta, las dos anteriores ecuaciones se pueden reescribir con una **función de densidad discreta binomial**:

$$p\left(t|\vec{m},\vec{w}\right) = \left(y\left(\vec{m},\vec{w}\right)\right)^t\left(1-y\left(\vec{m},\vec{w}\right)\right)^{1-t}$$

la cual constituye la **función de verosimilitud** de que la muestra t haya sido generada por un modelo con pesos $\vec{w}$. Para una matriz con N muestras de entrenamiento:

$$M = \begin{bmatrix} - & \vec{m}_1 & - \\ \vdots & \vdots & \vdots \\ - & \vec{m}_N & - \end{bmatrix},$$

y sus respectivas etiquetas dadas en el vector $\vec{t} \in \mathbb{R}^N$:

$$\vec{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix},$$

las cuales se asumen fueron generadas de forma independiente una de otra, con lo que la función de verosimilitud para las N muestras está dada por:

$$p\left(\vec{t}|M, \vec{w}\right) = \prod_{i=1}^{N} p\left(t_i|\vec{m}_i, \vec{w}\right) = \prod_{i=1}^{N} \left(y\left(\vec{m}_i, \vec{w}\right)\right)^{t_i} \left(1 - y\left(\vec{m}_i, \vec{w}\right)\right)^{1-t_i}.$$

Como ya se vió, usualmente es más fácil maximizar el logaritmo de la función de verosimilitud, por lo que entonces lo tomamos:

$$\ln\left(p\left(\vec{t}|M, \vec{w}\right)\right) = \sum_{i=1}^{N} t_i \ln\left(y\left(\vec{m}_i, \vec{w}\right)\right) + \left(1 - t_i\right) \ln\left(1 - y\left(\vec{m}_i, \vec{w}\right)\right).$$

Recordando entonces que el **negativo de la función de verosimilitud corresponde al error respecto a los párametros del modelo** $\vec{w}$, definimos la función de error como sigue:

$$E\left(\vec{w}\right) = -\ln\left(p\left(\vec{t}|M, \vec{w}\right)\right) = -\left\{\sum_{i=1}^{N} t_i \ln\left(y\left(\vec{m}_i, \vec{w}\right)\right) + \left(1 - t_i\right) \ln\left(1 - y\left(\vec{m}_i, \vec{w}\right)\right)\right\}.$$

Para encontrar el punto óptimo, **utilizamos el método de optimización por descenso de gradiente**, por lo que entonces se calcula el vector gradiente del error de la siguiente forma:

$$\nabla_{\vec{w}} E\left(\vec{w}\right) = -\sum_{i=1}^{N} \vec{m}_i \left\{\left(\frac{t_i}{y\left(\vec{m}_i, \vec{w}\right)}\right) \nabla_{\vec{w}}\left(y\left(\vec{m}_i, \vec{w}\right)\right) - \left(\frac{1 - t_i}{1 - y\left(\vec{m}_i, \vec{w}\right)}\right) \nabla_{\vec{w}}\left(y\left(\vec{m}_i, \vec{w}\right)\right)\right\}$$

$$\Rightarrow \nabla E\left(\vec{w}\right) = -\sum_{i=1}^{N} \vec{m}_i \left\{\left(\frac{t_i}{\cancel{y\left(\vec{m}_i, \vec{w}\right)}}\right) \cancel{y\left(\vec{m}_i, \vec{w}\right)} \left(1 - y\left(\vec{m}_i, \vec{w}\right)\right) - \left(\frac{1 - t_i}{\cancel{1 - y\left(\vec{m}_i, \vec{w}\right)}}\right) y\left(\vec{m}_i, \vec{w}\right) \cancel{\left(1 - y\left(\vec{m}_i, \vec{w}\right)\right)}\right\}$$

$$\Rightarrow \nabla E\left(\vec{w}\right) = -\sum_{i=1}^{N} \vec{m}_i \left\{t_i \left(1 - y\left(\vec{m}_i, \vec{w}\right)\right) - \left(1 - t_i\right) y\left(\vec{m}_i, \vec{w}\right)\right\}$$

$$\Rightarrow \nabla E\left(\vec{w}\right) = -\sum_{i=1}^{N} \vec{m}_i \left\{t_i - y\left(\vec{m}, \vec{w}\right)\right\}.$$

De esta forma, la siguiente es la ecuación de actualización de los pesos:

$$\vec{w}(\tau + 1)^T = \vec{w}(\tau)^T - \alpha \nabla E_P\left(\vec{w}\right) = \vec{w}(\tau)^T + \alpha \sum_{i=1}^{N} \vec{m}_i \left\{t_i - y\left(\vec{m}, \vec{w}\right)\right\}.$$

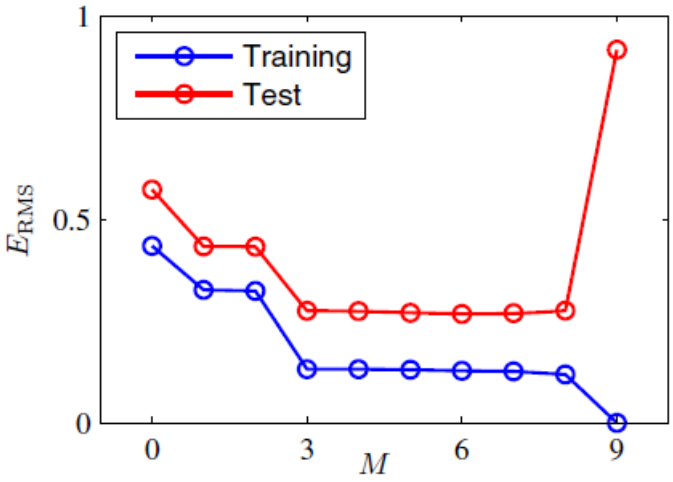

Figura 8: Error con datos de entrenamiento y datos de *validación*, tomado de [1].

Observe que en este caso se optó por minimizar el negativo de la función de verosimilitud, pero también es posible la minimización de la función de error cuadrática usando el gradiente, sin embargo es computacionalmente más eficiente la minimización del negativo de la verosimilitud.

2. Selección del modelo y la maldición de la dimensionalidad

2.1. Selección del modelo

Como se estudió en el problema de la regresión, el utilizar el error de ajuste del modelo con los datos de entrenamiento únicamente como parámetro de decisión del modelo a usar, conlleva un aumento en el error de clasificación cuando nuevos datos, con los cuales no se realizó el *entrenamiento* del modelo, se utilizan, como lo muestra la Figura 8.

Es necesario entonces realizar una evaluación del error con los datos de *validación,* los cuales no fueron utilizados para construir en el modelo. En los algoritmos supervisados de clasificación y regresión, se suele particionar el conjunto de datos etiquetados en un conjunto de entrenamiento y de validación, usualmente con una proporción de 70 % de los datos para entrenamiento y el restante 30 % para validación. Sin embargo, muchas veces se dispone de no suficiente número de datos etiquetados, por lo que se procede a realizar lo que se conoce como una *validación cruzada* de los datos. Para ello, se realizan C

Figura 9: Validación cruzada del modelo, tomado de [1].

corridas del algoritmo, y en cada una se escoge un conjunto distinto de muestras de entrenamiento y de validación, como muestra la Figura 9. Esto además es útil para algoritmos con componentes aleatorios, donde una corrida con los mismos datos puede arrojar resultados distintos.

2.2. Maldición de la dimensionalidad

Tómese el caso en que se desea realizar la clasificación de un conjunto de muestras de dos dimensiones $\vec{x} \in \mathbb{R}^2$ y $K = 3$ clases, como se muestra en la Figura 11. De esta forma, si un dato nuevo se desea clasificar, se asigna su clase según la pertenencia de la celda a una clase de las $K = 3$ clases.

Este esquema de división del espacio en celdas tiene dos problemas, en función de la dimensionalidad del espacio:

- La cantidad de celdas crece exponencialmente con la dimensión de tal espacio, como lo muestra la Figura 11.
- La cantidad de datos necesaria para determinar la pertenencia de la celda a alguna de las clases debe también crecer exponencialmente, de forma que se eviten las celdas vacías.

Las anteriores claúsulas suponen que el agregar más dimensiones a las muestras a clasificar, para nuestro clasificador sencillo por celdas, aumenta la complejidad y la cantidad necesaria de muestras, lo que se le conoce como la *maldición de la dimensionalidad.*

Referencias

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

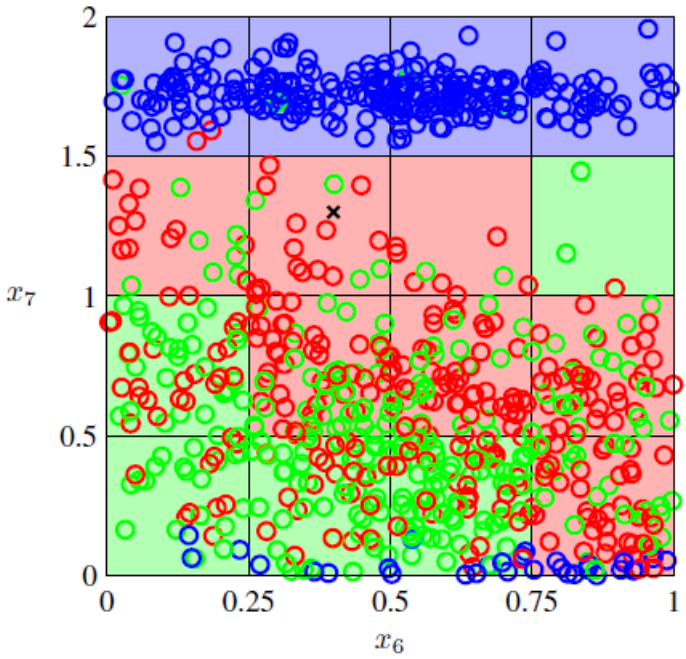

Figura 10: Clasificador por celdas, tomado de [1].

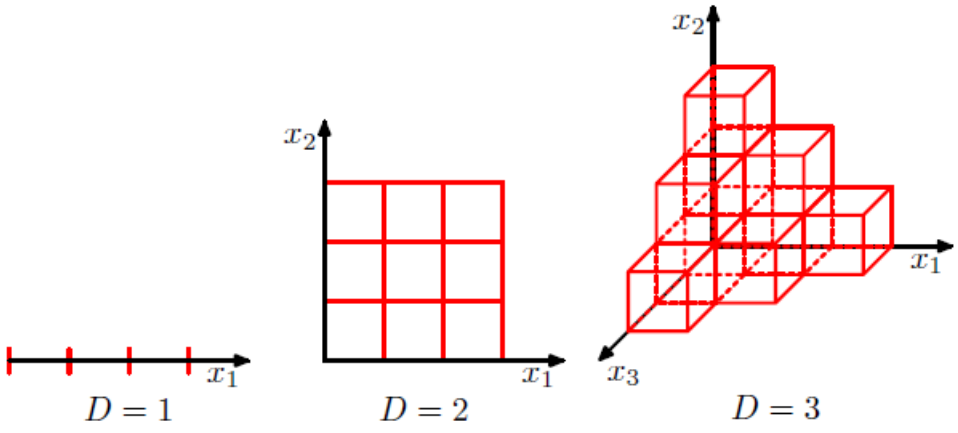

Figura 11: La maldición de la dimensionalidad, tomado de [1].