Análisis numérico: Mínimos cuadrados

1 Teoría de los Mínimos Cuadrados desde una Perspectiva Geométrica

Definición 1
Sea V un espacio vectorial sobre un campo $\mathbb{K}$ (comúnmente $\mathbb{R}$ o $\mathbb{C}$) y sea $W \subseteq V$ un subespacio vectorial. Dado un vector $v \in V$, queremos encontrar un vector $\hat{v} \in W$ tal que $\hat{v}$ sea la **proyección ortogonal** de v sobre W.

La proyección ortogonal $\hat{v}$ satisface las siguientes condiciones:

1. $\hat{v} \in W$, es decir, pertenece al subespacio.
2. $v - \hat{v} \in W^{\perp}$, donde $W^{\perp}$ es el complemento ortogonal de W, definido como:

$$W^{\perp} = \{u \in V : \langle u, w \rangle = 0 \ \forall w \in W\}$$

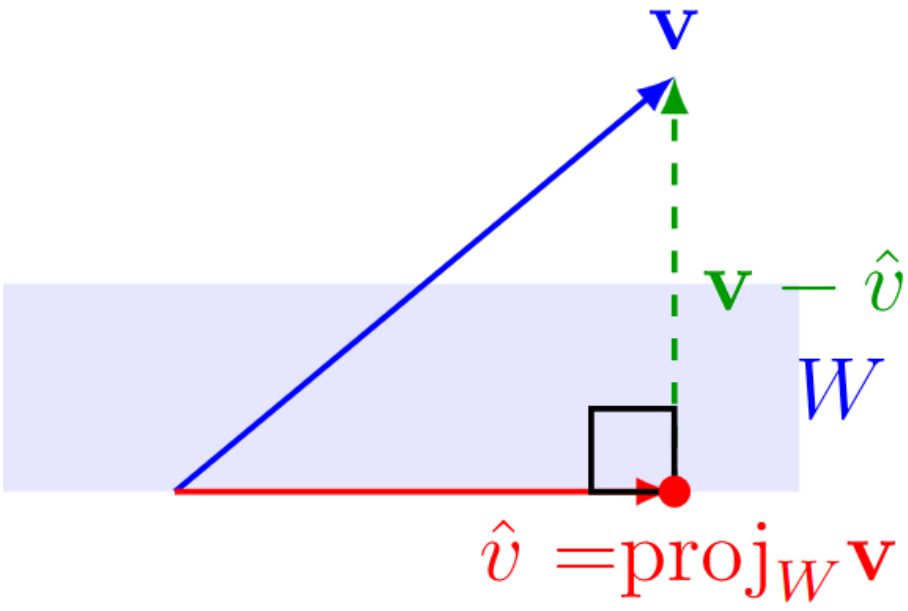

Figure 1: proyección

En otras palabras, la proyección $\hat{v}$ es el **vector en** W **más cercano a** v, y la diferencia $v - \hat{v}$ es ortogonal a todo vector en W.

Cómo hallar la proyección ortogonal con una base ortonormal
Sea $\{w_1, w_2, \ldots, w_k\}$ una base ortonormal de W. Cualquier vector $w \in W$ puede escribirse como combinación lineal de los vectores base:

$$\hat{v} = \sum_{i=1}^{k} \alpha_i w_i$$

Para determinar los coeficientes α_i, usamos la condición de ortogonalidad:

$$v - \hat{v} \in W^{\perp} \quad \Rightarrow \quad \langle v - \hat{v}, w_j \rangle = 0 \quad \forall j = 1, \ldots, k$$

Sustituyendo $\hat{v}$ en esta expresión:

$$\langle v - \sum_{i=1}^{k} \alpha_i w_i, w_j \rangle = 0$$

Expandiendo el producto interno y usando la linealidad:

$$\langle v, w_j \rangle - \sum_{i=1}^{k} \alpha_i \langle w_i, w_j \rangle = 0$$

Dado que la base es ortonormal, $\langle w_i, w_j \rangle = \delta_{ij}$ (donde δ_{ij} es el delta de Kronecker), se simplifica a:

$$\alpha_j = \langle v, w_j \rangle \quad \forall j$$

Por lo tanto, la proyección ortogonal es:

$$\hat{v} = \sum_{i=1}^{k} \langle v, w_i \rangle w_i$$

Teorema 1
Sea W un subespacio de $\mathbb{R}^n$. Entonces, dado el vector v en $\mathbb{R}^n$, el vector en W más cercano a v es $\text{proj}_W v$. Esto es, para w en W, $\|v - w\|$ es mínima cuando $w = \text{proj}_W v$.

Demostración
Sea w cualquier vector en W. Entonces:

$$v - w = (v - \text{proj}_W v) + (\text{proj}_W v - w).$$

Como w y $\text{proj}_W v$ están ambos en W, $\text{proj}_W v - w$ está en W. Por definición 1, $v - \text{proj}_W v$ es ortogonal a cada vector en W, de modo que

$$\|v - w\|^2 = (v - w) \cdot (v - w) = ((v - \text{proj}_W v) + (\text{proj}_W v - w)) \cdot ((v - \text{proj}_W v) + (\text{proj}_W v - w))$$
$$= \|v - \text{proj}_W v\|^2 + \|\text{proj}_W v - w\|^2.$$

Si $w \neq \text{proj}_W v$, entonces $\|\text{proj}_W v - w\|^2$ es positivo, y

$$\|v - w\|^2 > \|v - \text{proj}_W v\|^2.$$

Se sigue, entonces, que $\text{proj}_W v$ es el vector en W que minimiza $\|v - w\|^2$, y por lo tanto, minimiza $\|v - w\|$.

Nota adicional
Para minimizar $\|v - w\|$, es suficiente proyectar v en el subespacio W. Recordando que la proyección ortogonal asegura que el residuo $r = v - \text{proj}_W v$ es perpendicular al espacio proyectado.

$$r \perp W \quad \text{y} \quad v = \text{proj}_W v + r.$$

Esto confirma que $\text{proj}_W v$ es la mejor aproximación de v en el espacio W.

Representación de un conjunto de datos en el Espacio Vectorial

La teoría de los **Mínimos Cuadrados** se basa en encontrar la mejor aproximación de un conjunto de datos observados mediante un modelo matemático, minimizando el error cuadrático entre los valores observados y los valores predichos por dicho modelo. Desde una **perspectiva geométrica**, este método puede entenderse en términos de **proyecciones** en espacios vectoriales.

Supongamos que tenemos un conjunto de datos con n observaciones, cada una asociada con una variable dependiente y_i y asociadas p variables independientes representadas por $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$.

El conjunto de datos dependiente se puede pensar como

$$y = \begin{bmatrix} y_1, y_2, \ldots, y_n \end{bmatrix}^T \in \mathbb{R}^n$$

y cada conjunto de datos de la i-ésima variable independiente como

$$X_i = \begin{bmatrix} x_{i,1}, x_{i,2}, \ldots, x_{i,n} \end{bmatrix}^T \in \mathbb{R}^n.$$

El objetivo es ajustar un modelo lineal de la forma:

$$\hat{y} = X_1\beta_1 + \cdots + X_p\beta_p = \begin{bmatrix} X_1 & \ldots & X_p \end{bmatrix}_{n \times p} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p \times 1} = X\beta.$$

Donde:

- X es la matriz de diseño ($n \times p$) que contiene las variables independientes.
- β es el vector de parámetros ($p \times 1$) a estimar.
- $\hat{y}$ es el vector de valores predichos ($n \times 1$).

Ahora, del álgebra lineal se sabe que $X\hat{\beta} = y$ es consistente si y sólo si y pertenece al espacio columna de X. De forma equivalente, $X\hat{\beta} = y$ es inconsistente si y sólo si y no está en el espacio columna de X. Los sistemas inconsistentes aparecen en muchas situaciones, por lo que debemos saber cómo tratarlos. Nuestro método consiste en modificar el problema de manera que no sea indispensable que la ecuación matricial $X\hat{\beta} = y$ se satisfaga. En cambio, buscamos un vector β en $\mathbb{R}^n$, tal que $X\beta$ sea tan cercano a y como sea posible.

Si W es el espacio generado por las columnas de X, el teorema 1 implica que el vector en W más cercano a y es $\text{proj}_W y$. Es decir, $\|y - w\|$, para $w \in W$, se minimiza cuando $w = \text{proj}_W y$. En consecuencia, si encontramos β tal que $X\beta = \text{proj}_W y$, estamos seguros de que $\|y - X\beta\|$ será lo más pequeña posible.

Como se indica en la demostración del teorema 1, $y - \text{proj}_W y = y - X\beta$ es ortogonal a todo vector en W (Vea la figura 1). Esto implica que $y - X\beta$ es ortogonal a cada columna de X. En términos de una ecuación matricial, tenemos que

$$\vec{0} = \begin{bmatrix} \langle X_1, y - X\beta \rangle \\ \vdots \\ \langle X_p, y - X\beta \rangle \end{bmatrix}_{p\times 1} = X^T(y - X\beta)$$

o, de manera equivalente,

$$X^T X\beta = X^T y.$$

Por lo tanto, β es una solución para

$$X^T X\hat{\beta} = X^T y. \tag{1}$$
$$(X^T X)_{p\times p}\hat{\beta}_{p\times 1} = (X^T y)_{p\times 1}.$$

Cualquier solución de (1) es una **solución por mínimos cuadrados** del sistema lineal $X\hat{\beta} = y$. (**Cuidado:** en general, $X\beta \neq y$.) La ecuación (1) es el **sistema normal** de ecuaciones asociadas con $X\hat{\beta} = y$, o simplemente, el sistema normal. Observe que si $X\hat{\beta} = y$ es consistente, una solución para este sistema es una solución por mínimos cuadrados.

En particular, si X es no singular, una solución por mínimos cuadrados para $X\hat{\beta} = y$ es la solución usual, $\hat{\beta} = X^{-1}y$

Teorema 2

Si X es una matriz de $n \times p$ tal que $\text{rango}(X) = p$, entonces $X^T X$ es no singular y el sistema lineal $X\beta = y$ tiene una única solución por mínimos cuadrados, dada por $\beta = (X^T X)^{-1} X^T y$. Es decir, el sistema normal de ecuaciones tiene una única solución.

Demostración

Si X tiene rango p, las columnas de X son linealmente independientes. La matriz $X^T X$ es no singular si el sistema lineal $X^T X\beta = 0$ sólo tiene la solución nula. Al multiplicar, por la izquierda, ambos lados de $X^T X\beta = 0$ por β^T, obtenemos

$$0 = \beta^T X^T X\beta = (X\beta)^T(X\beta) = (X\beta) \cdot (X\beta).$$

De acuerdo con el teorema 4.3 de la sección 4.2 (Kolman and Hill, 1984) , resulta entonces que $X\beta = 0$. Pero esto implica que tenemos una combinación lineal de las columnas linealmente independientes de X igual a cero, por lo cual $\beta = 0$. En consecuencia, $X^T X$ es no singular y la ecuación (1) tiene la solución única

$$\beta = (X^T X)^{-1} X^T y.$$

Interpretación Geométrica del Residuo

Una vez que se proyecta y sobre $\text{Col}(X) = W$, el vector y se descompone como la suma de dos componentes ortogonales:

$$y = \hat{y} + r = \hat{y} + (y - \text{proj}_W y) = \hat{y} + (y - X\beta)$$

- $\hat{y} \in \text{Col}(X)$: Es la proyección de y, es decir, los valores ajustados por el modelo.
- $y - X\beta = r \perp \text{Col}(X)$: Es el residuo, que es perpendicular al espacio generado por las variables independientes.

Interpretación Visual resumen
Si representamos geométricamente los datos:

- El vector y apunta a un punto en el espacio de n dimensiones.
- El subespacio $\text{Col}(X)$ es un plano (o hiperplano) dentro de este espacio.
- La solución de mínimos cuadrados corresponde al pie de la perpendicular que va desde y al plano $\text{Col}(X)$.

Ventajas del Enfoque Geométrico
El enfoque geométrico permite:

1. **Visualizar la naturaleza del ajuste**: Ver cómo el modelo proyecta y en el subespacio de las variables independientes.
2. **Interpretar los residuos**: Entender que representan la "parte" de y que no puede ser explicada por X.

2 Perspectiva algebraica

Mínimos cuadrados lineales Supongamos que tenemos un conjunto de n datos dependientes y_i donde cada y_i se determina en función de x_i, tal que obtenemos a su vez un conjunto de datos x_i para i = 1, . . . , n; supongamos que se sospecha que la relación entre ambos conjuntos es lineal. Veamos, entonces, cómo se realiza una regresión cuadrática lineal desde sus principios algebraicos. Dado que los datos dependientes pueden estar sesgados, tener errores de aproximación o estar influenciados por otros aspectos que los hagan imprecisos, tenemos que

$$y \approx a_1 \cdot x + a_0$$

donde a_1 y a_0 son constantes por ahora desconocidas. De este modo, el error entre la aproximación lineal y el valor obtenido empíricamente está dado por la ecuación

$$E_i = (y_i - (a_1 \cdot x_i + a_0))^2$$

Donde se eleva al cuadrado para evitar errores negativos (recordemos que para este método no se usa el error en valor absoluto pues debemos derivar esta ecuación, como veremos a continuación, y la función valor absoluto no es derivable en todo su dominio). Ahora bien, lo que buscamos es que nuestra aproximación sea tal que nos dé una línea recta tal que esté lo menos distanciada del total de datos y_i como sea posible. En otras palabras, que el error total, dado por la ecuación:

$$E_T = \sum_{i=1}^{n} (y_i - (a_1 \cdot x_i + a_0))^2$$

sea lo más bajo posible. Notemos, ahora, que tanto los puntos y_i como los x_i ya están determinados, y el mínimo error lo podemos obtener al escoger a_1 y a_0 adecuados, lo cual podemos hacer al minimizar la función de error total. Usando el criterio de la primera derivada, obtenemos que la función E_T será lo más pequeña posible para a_0 y a_1 tales que

$$\frac{\partial E_T}{\partial a_0} = 0, \quad \frac{\partial E_T}{\partial a_1} = 0$$

Al diferenciar la función E_T con relación a a_0 tenemos que, al usar la regla de la cadena:

$$\frac{\partial E_T}{\partial a_0} = \sum_{i=1}^{n} 2 \cdot (y_i - (a_1 \cdot x_i + a_0)) \cdot (-1)$$

$$= 2 \sum_{i=1}^{n} ((a_1 \cdot x_i + a_0) - y_i) = 0$$

Esto es:

$$\sum_{i=1}^{n} y_i = a_1 \cdot \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a_0 = a_1 \cdot \sum_{i=1}^{n} x_i + n \cdot a_0 \qquad (*)$$

Por otra parte, al derivar E_T respecto a a_1 y haciendo uso de la regla de la cadena, tenemos que:

$$\frac{\partial E_T}{\partial a_1} = \sum_{i=1}^{n} 2 \cdot (y_i - (a_1 \cdot x_i + a_0)) \cdot (-x_i) = 2 \sum_{i=1}^{n} y_i \cdot (-x_i) - (a_1 \cdot (x_i)^2 + a_0 \cdot (-x_i) = 0$$

Esto es:

$$\sum_{i=1}^{n} y_i \cdot x_i = a_1 \sum_{i=1}^{n} (x_i)^2 + a_0 \sum_{i=1}^{n} x_i \qquad (**)$$

Así, tenemos dos ecuaciones de dos incógnitas, tal que a partir de (*) y (**) podemos obtener a_0 y a_1:

$$a_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$a_1 = \frac{n \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Mínimos cuadrados polinomiales Ahora supongamos que tenemos un conjunto de n datos dependientes y_i donde cada y_i se determina en función de un dato x_i, tal que obtenemos a su vez un conjunto de datos x_i para i = 1, . . . , n; supongamos que se sospecha que la relación entre ambos conjuntos es polinomial, esto es, que

$$y \approx P(x_i) = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x_i^m$$

Donde $a_0, \cdots, a_m$ son constantes por ahora desconocidas y m es el grado del polinomio, no mayor al número de datos conocidos para que al final no resulten más incógnitas que datos conocidos. Como en el caso anterior, dado que los datos dependientes pueden estar sesgados, tener errores de aproximación o estar influenciados por otros aspectos que los hagan imprecisos, tenemos que el error entre la aproximación y el valor obtenido empíricamente está dado por

$$E_i = (y_i - P(x_i))^2$$

Aquí, como en el caso anterior, se eleva al cuadrado para evitar errores negativos y la dificultad del valor absoluto. Ahora bien, este caso es análogo al anterior en el sentido en que queremos minimizar la función de error, solo que ahora tendremos que determinar m constantes distintas con el proceso que sigue. Primero obtenemos la función del error total, dada por

$$\begin{aligned} E_T &= \sum_{i=1}^{n} (y_i - P(x_i))^2 = \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} y_i \cdot P(x_i) + \sum_{i=1}^{n} (P(x_i))^2 \\ &= \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} y_i \cdot \left(\sum_{j=0}^{m} a_j x_i^j \right) + \sum_{i=1}^{n} \left(\sum_{j=0}^{m} a_j x_i^j \right)^2 \\ &= \sum_{i=1}^{n} y_i^2 - 2 \sum_{j=0}^{m} a_j \cdot \left(\sum_{i=1}^{n} y_i x_i^j \right) + \sum_{i=1}^{n} \left(\sum_{j=0}^{m} a_j x_i^j \right) \left(\sum_{j=0}^{m} a_j x_i^j \right) \\ &= \sum_{i=1}^{n} y_i^2 - 2 \sum_{j=0}^{m} a_j \cdot \left(\sum_{i=1}^{n} y_i x_i^j \right) + \sum_{j=0}^{m} \sum_{k=0}^{m} a_j a_k \left(\sum_{i=1}^{n} x_i^{j+k} \right) \end{aligned}$$

Ahora bien, para minimizar este error en función de todas las constantes debemos tener que:

$$\frac{\partial E_T}{\partial a_j'} = 0 \quad \text{para } j' = 0, 1, \ldots, m$$

Al diferenciar esta función de error para un j' particular, obtenemos que:

$$\frac{\partial E_T}{\partial a_{j'}} = \frac{\partial}{\partial a_{j'}} \left(\sum_{i=1}^{n} y_i^2 \right)$$

$$-2 \frac{\partial}{\partial a_{j'}} \left(a_0 \sum_{i=1}^{n} y_i x_i^j + \cdots + a_{j'} \sum_{i=1}^{n} y_i x_i^j + \cdots + a_m \sum_{i=1}^{n} y_i x_i^j \right)$$

$$+\frac{\partial}{\partial a_{j'}}\left(a_0\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}+\cdots+a_{j'}\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}+\cdots+a_m\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}\right)$$

Por facilidad, distribuimos la derivada en las tres partes que componen a la función del error total, y al separar los términos notamos que la primera sumatoria no depende de $a_{j'}$, por lo cual:

$$\frac{\partial}{\partial a_{j'}}\left(\sum_{i=1}^{n}y_i^2\right)=0$$

Ahora bien, la expansión de la segunda sumatoria inicial nos permite fácilmente notar que solo la sumatoria cuyo coeficiente es $a_{j'}$ depende de $a_{j'}$, tal que:

$$\begin{aligned}&-2\frac{\partial}{\partial a_{j'}}\left(a_0\sum_{i=1}^{n}y_ix_i^j+\cdots+a_{j'}\sum_{i=1}^{n}y_ix_i^j+\cdots+a_m\sum_{i=1}^{n}y_ix_i^j\right)\\&=-2\frac{\partial}{\partial a_{j'}}\left(a_{j'}\sum_{i=1}^{n}y_ix_i^j\right)\\&=-2\sum_{i=1}^{n}x_i^jy_i\end{aligned}$$

Finalmente con la expansión de una de las sumatorias que componen la parte final de la función de error total, se puede notar que un término que depende de $a_{j'}$ es el que lo tiene como coeficiente y, por otra parte, al expandir las sumas

$$\sum_{k=0}^{m}a_k$$

y multiplicarlas con los respectivos a_i, obtendremos otro término igual al primero que mencionamos. Así:

$$\begin{aligned}&\frac{\partial}{\partial a_{j'}}\left(a_0\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}+\cdots+a_{j'}\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}+\cdots+a_m\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}\right)\\&=\frac{\partial}{\partial a_{j'}}\left(2a_{j'}\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}\right)\\&=2\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}\end{aligned}$$

Así, tenemos que

$$\frac{\partial E_T}{\partial a_{j'}}=-2\sum_{i=1}^{n}x_i^jy_i+2\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}$$

De esta forma, como

$$\frac{\partial E_T}{\partial a_{j'}}=0$$

entonces

$$\sum_{i=1}^{n}x_i^jy_i=\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^{j+k}$$

para $j=0,1,\ldots,m$.

Ahora bien, para $j=0$ tenemos que:

$$\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^k=a_0\sum_{i=1}^{n}x_i^0+a_1\sum_{i=1}^{n}x_i^1+\cdots+a_m\sum_{i=1}^{n}x_i^m=\sum_{i=1}^{n}x_i^0y_i$$

Similarmente, para $j=1$ tenemos que:

$$\sum_{k=0}^{m}a_k\sum_{i=1}^{n}x_i^k=a_0\sum_{i=1}^{n}x_i^1+a_1\sum_{i=1}^{n}x_i^2+\cdots+a_m\sum_{i=1}^{n}x_i^{m+1}=\sum_{i=1}^{n}x_i^1y_i$$

Así sucesivamente hasta llegar a $j = m$ tal que:

$$\sum_{k=0}^{m} a_k \sum_{i=1}^{n} x_i^{m+k} = a_0 \sum_{i=1}^{n} x_i^m + a_1 \sum_{i=1}^{n} x_i^{m+1} + \cdots + a_m \sum_{i=1}^{n} x_i^{2m} = \sum_{i=1}^{n} x_i^m y_i$$

Esto es, tenemos un sistema de ecuaciones que puede representarse como en la matriz a continuación

$$\begin{bmatrix} \sum_{i=1}^n x_i^0 & \sum_{i=1}^n x_i^1 & \sum_{i=1}^n x_i^2 & \cdots & \sum_{i=1}^n x_i^m \\ \sum_{i=1}^n x_i^1 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \cdots & \sum_{i=1}^n x_i^{m+1} \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 & \cdots & \sum_{i=1}^n x_i^{m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_i^m & \sum_{i=1}^n x_i^{m+1} & \sum_{i=1}^n x_i^{m+2} & \cdots & \sum_{i=1}^n x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \cdot x_i^0 \\ \sum_{i=1}^n y_i \cdot x_i^1 \\ \sum_{i=1}^n y_i \cdot x_i^2 \\ \vdots \\ \sum_{i=1}^n y_i \cdot x_i^m \end{bmatrix}$$

Esto es, un sistema de ecuaciones con $m+1$ incógnitas y $m+1$ ecuaciones, que tendrá solución con los métodos conocidos de álgebra lineal bajo las hipótesis para que un sistema tenga solución única.

3 Data Science

La Data Science o ciencia de datos, es la ciencia que combina métodos de las matemáticas, la estadística, la informática y la computación utilizados para analizar y evaluar diferentes fenómenos para obtener resultados y conclusiones. Esto significa que es un área multidisciplinaria, que bebe de diferentes perspectivas científicas.

El principal objetivo de la data science es obtener resultados, es decir, aproximaciones, estimaciones, pronósticos, clasificaciones o, más modernamente, productos visuales y de texto. Para esto, requiere datos relacionados al dominio y al alcance, los cuales se están investigando y se utilizan una serie de métodos y algoritmos que permiten observar relaciones matemáticas o estadísticas. Entre estos métodos mencionados se encuentra mínimos cuadrados. Este se puede usar para lograr entender tendencias, observar y dar sentido a cantidades extensas de datos. También tiene un uso subyacente a través de un área del aprendizaje de máquina, conocido como deep learning o aprendizaje profundo. En este se habla de las redes neuronales, que pueden usar como método sucesivo bases de los mínimos cuadrados.

Se observará su versión más sencilla, con un único perceptron. Suponga que se tienen datos x_i, $i = 1, 2, \ldots, n$. Se organizan de la siguiente manera:

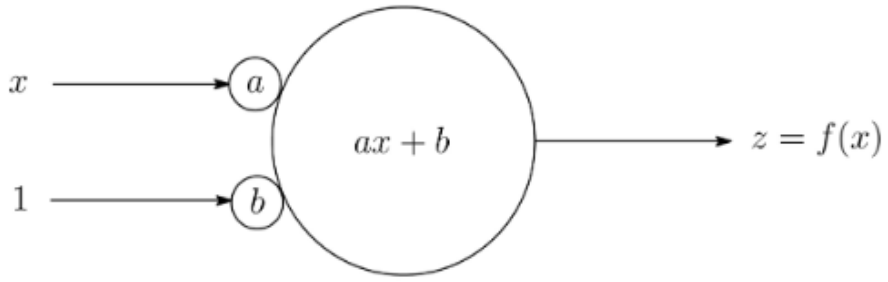

Figure 2: Perceptrón simple

Datos de entrada:$\{(x_1, 1), (x_2, 1), \ldots, (x_n, 1)\}$
Señal guía:$\{y_1, y_2, \ldots, y_n\}$

Sea b el bias, que afecta los valores de salida desplazándolos sobre el eje de las abscisas. Suele ser iniciado como 0 y se deben ir aproximando a su valor óptimo al iterarse.

En deep learning la función de error suele llamarse L por Loss function, que se observa a continuación.

$$L(a, b) = E(a, b) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 = \frac{1}{2} \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Se busca minimizar L eligiendo correctamente a y b. Sin embargo, a diferencia del procedimiento usual en estadística en donde se hallan los mínimos estacionarios, en este caso se utiliza el método del descenso por gradiente, o alguna de sus variaciones. Para $t = 0, 1, 2, \ldots$

$$\begin{bmatrix} a(0) \\ b(0) \end{bmatrix} \to \begin{bmatrix} a(1) \\ b(1) \end{bmatrix} \to \cdots \to \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} \to \begin{bmatrix} a(t+1) \\ b(t+1) \end{bmatrix} \to \ldots$$

Con a y b

$$a(t+1) = a(t) - \epsilon \frac{\partial L}{\partial a(t)}$$
$$b(t+1) = b(t) - \epsilon \frac{\partial L}{\partial b(t)}$$

L siendo la loss function y ϵ siendo la tasa de aprendizaje o learning rate, que se conoce que es un número entre 0 y 1 y es matemáticamente complejo de elegir correctamente. Expandiendo las expresiones a manera matricial, obtenemos

$$\begin{bmatrix} a(t+1) \\ b(t+1) \end{bmatrix} = \begin{bmatrix} 1 - \epsilon \sum_{i=1}^{n} x_i^2 & -\epsilon \sum_{i=1}^{n} x_i \\ -\epsilon \sum_{i=1}^{n} x_i & 1 - n\epsilon \end{bmatrix} \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} + \epsilon \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}.$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \epsilon \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} + \epsilon \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}.$$

Que se puede resumir como

$$x(t) = \begin{bmatrix} a(t) \\ b(t) \end{bmatrix}, \; A = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}, \; f = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

Con lo que terminamos con una ecuación simple

$$x(t+1) = (I - \epsilon A)x(t) + \epsilon f$$

Donde I es la matriz identidad. Podemos ver que A es invertible siempre y cuando exista $x_i \neq x_j$, para $i, j \in 1, 2, \ldots, n$, lo que nos permite escapar de la solución trivial y hallar

$$x(t) = (I - \epsilon A)x(t-1) + \epsilon f$$

Repitiendo la sustitución

$$x(t) = (I - \epsilon A)x(0) + \sum_{k=0}^{t-1} (I - \epsilon A)^k \epsilon f$$

Se busca expandir la serie. Sea $M = I - \epsilon A$ y $S = \sum_{k=0}^{t-1} (I - \epsilon A)^k$. Expandimos la serie. Veamos que:

$$S = I + M + M^2 + \cdots + M^{t-1}$$

Multiplicamos ambos lados por $I - M$

$$(I - M)S = (I - M)(I + M + M^2 + \cdots + M^{t-1}).$$

Los valores del lado derecho se cancelan telescópicamente

$$(I - M)S = I - M^t.$$

Separamos la suma

$$S = (I - M^t)(I - M)^{-1}.$$

Sustituimos M y la suma de nuevo

$$\sum_{k=0}^{t-1}(I-\epsilon A)^k = (I-(I-\epsilon A)^t)(I-(I-\epsilon A))^{-1}.$$

$$\sum_{k=0}^{t-1}(I-\epsilon A)^k = (I-(I-\epsilon A)^t)(\epsilon A)^{-1}.$$

Esta expresión la reemplazamos en $x(t)$

$$x(t) = (I-\epsilon A)x(0) + (I-(I-\epsilon A)^t)A^{-1}f$$

Que son valores que se pueden conocer algebraicamente.

Medición del error

En la práctica se utilizan múltiples métodos estadísticos para lograr calcular la cercanía de las regresiones a los valores verdaderos. Cada uno tiene su base matemática y su debida interpretación, como se observará a continuación.

Mean Squared Error (MSE) También conocido en español como error cuadrático medio. Calcula la calidad de la estimación con el valor medio del cuadrado de los errores. Siendo Y_i uno de los n valores observados y $\hat{Y}_i$ los valores estimados, con $i = 1, 2, \ldots, n$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

Es fácil observar que el MSE será un número real no negativo, dado que se deriva de la distancia euclidiana. Mientras menor sea el resultado, se considera que la regresión es más precisa.

R^2 (R-squared) También conocido como el coeficiente de determinación, es la proporción de la variación que se puede predecir de la variable dependiente usando la variable independiente.

Sea un dataset con n valores, marcados como y_i, $i = 1, 2, \ldots, n$, cada uno asociado a un valor estimado f_i, $i = 1, 2, \ldots, n$, se definen los residuales como $e_i = y_i - f_i$.

Sea $\overline{y}$ la media de la data observada, con

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

La variabilidad del dataset puede ser medida con:

1. La suma de cuadrados de los residuales

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

2. La suma total de los cuadrados, que es proporcional a la varianza de los datos:

$$SS_{tot} = \sum_i (y_i - \overline{y})^2$$

La definición del coeficiente de determinación es

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

El mejor de los casos es en el que los valores estimados son iguales a los valores observados, por lo tanto $SS_{res} = 0$ y $R^2 = 1$.

4 Referencias

- Burden, R. L., and Faires, J. D. (2017). Análisis numérico. Cengage Learning.
- Fujii, K. (2018) Least Squares Method from the View Point of Deep Learning. Advances in Pure Mathematics, 8, 485-493
- Kolman, B., and Hill, D. R. (1984). Álgebra lineal (1.ª ed.). Pearson Educación.
- Trefethen, L. N., and Bau, D. III. (1997). Numerical Linear Algebra. SIAM.