Maestría en Inteligencia Artificial Aplicada

Modelo de Regresión Lineal Simple

Inteligencia Artificial y Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales

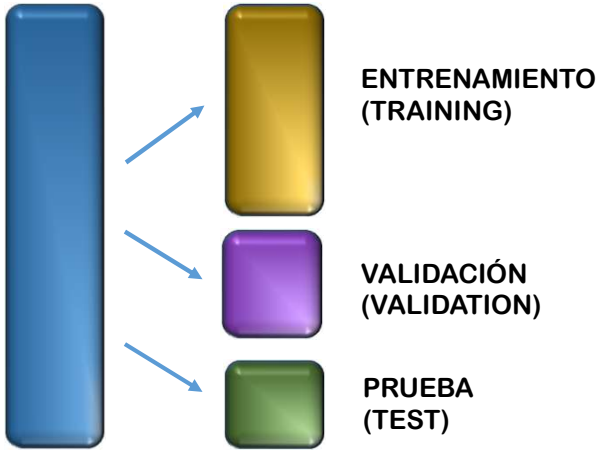
ITESM

Campus Guadalajara

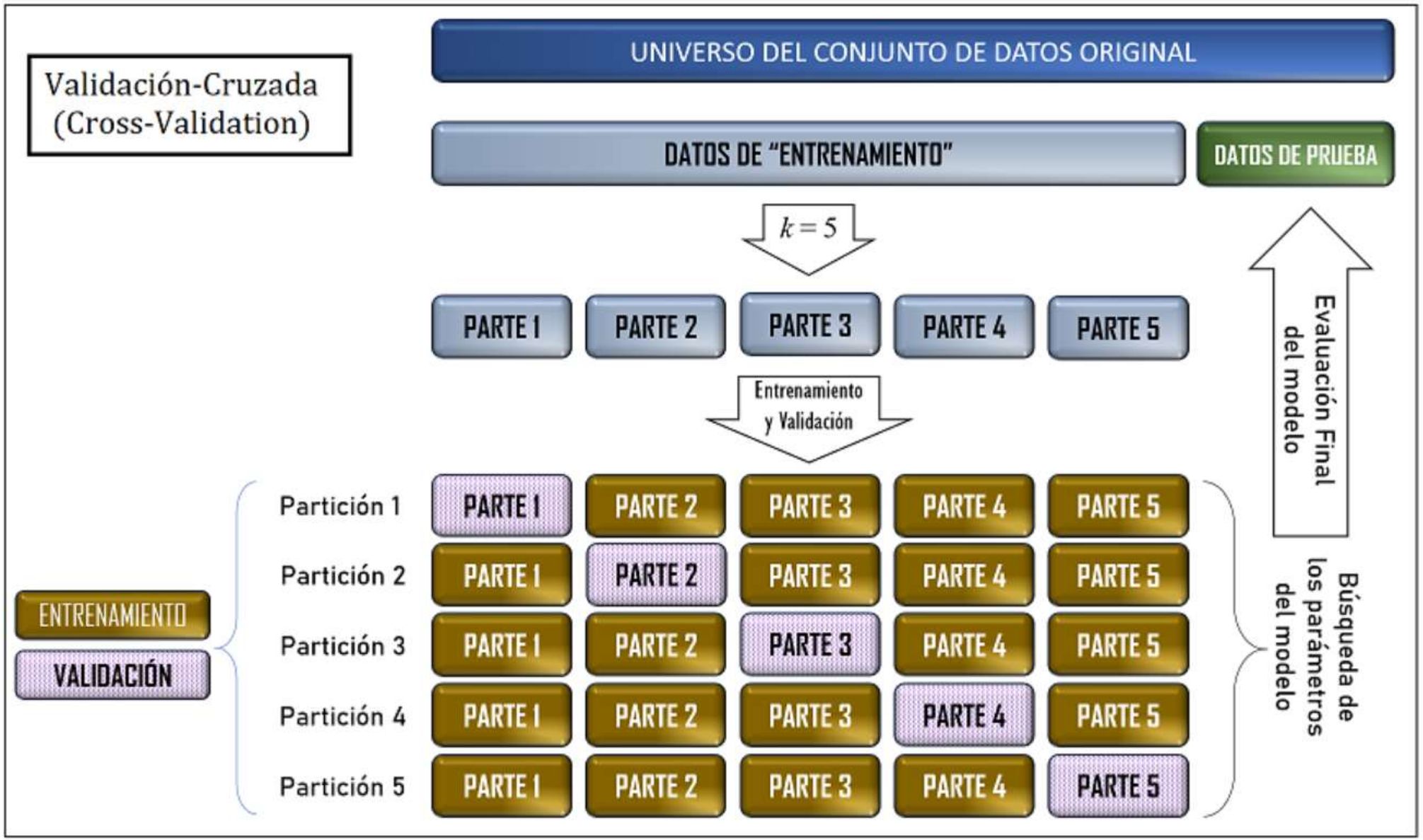
Partición Entrenamiento-Validación-Prueba en Aprendizaje Supervisado

Actualmente se propone una partición de los datos en tres conjuntos para la generación de un modelo de aprendizaje automático bajo un el criterio supervisado. Es decir, se cuenta con un conjunto de datos de entrada a partir de los cuales se desea predecir el valor de la variable de salida con valores conocidos.

- **Datos de Entrenamiento:** para generar los parámetros del modelo en la etapa de aprendizaje (también llamado entrenamiento).
- **Datos de Validación:** para medir el desempeño parcial del modelo obtenido con los datos de entrenamiento y a partir del cual proponer ajustes a los hiperparámetros que permitan mejorar dicho desempeño en un proceso iterativo del aprendizaje.
- **Datos de Prueba:** para evaluar el desempeño final del modelo.



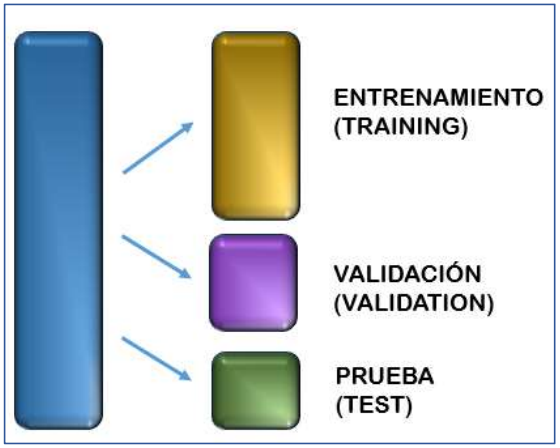
Cuando no se cuente con una buena cantidad de datos de entrada que me permita generar estos tres conjuntos con la suficiente información para obtener un modelo de aprendizaje aceptable, se podrá aplicar el método de Validación Cruzada que veremos más adelante. Igualmente varios autores siguen usando el conjunto de prueba y validación como el mismo.


Validación-Cruzada
(Cross-Validation)
UNIVERSO DEL CONJUNTO DE DATOS ORIGINAL
DATOS DE "ENTRENAMIENTO"
DATOS DE PRUEBA
k = 5
PARTE 1
PARTE 2
PARTE 3
PARTE 4
PARTE 5
Entrenamiento
y Validación
Evaluación Final
del modelo
Búsqueda de
los parámetros
del modelo
ENTRENAMIENTO
VALIDACIÓN
Partición 1
Partición 2
Partición 3
Partición 4
Partición 5

Validación-Cruzada (Cross-Validation)

- La técnica de validación-cruzada es una técnica que permite seleccionar entre varios modelos, el mejor de ellos. La manera particular en que se realiza el proceso de entrenamiento tiene como objetivo el garantizar la independencia del modelo de la partición misma.

- Existe una variedad de formas en las cuales se puede implementar la validación-cruzada, a lo largo del curso estaremos estudiando algunas de ellas. En la diapositiva siguiente se encuentra un diagrama ejemplificando el modelo básico a partir del cual se obtienen los demás.



Pasos para seleccionar el mejor modelo mediante partición tripartita

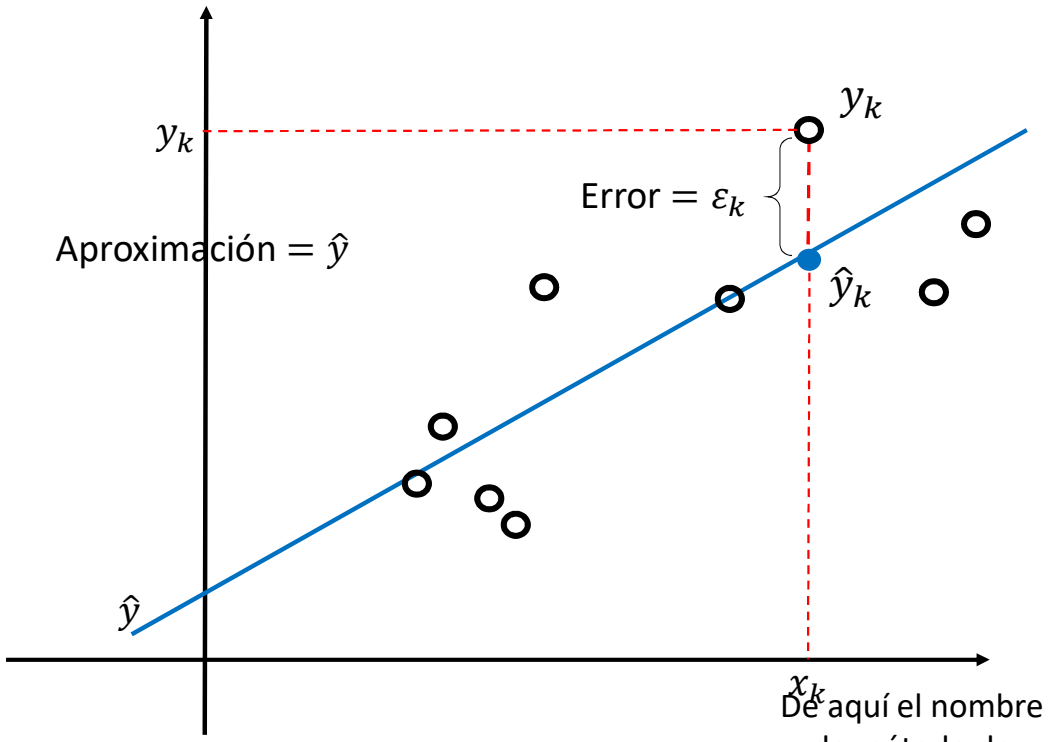
Cuando se utiliza la técnica de validación-cruzada en aprendizaje supervisado para la búsqueda del mejor modelo entre varias opciones (las diferentes opciones pueden ser modelos de aprendizaje supervisado diferentes, o bien un mismo modelo con diferentes hiperparámetros) y partiendo de los conjuntos de entrenamiento, validación y prueba, se suele proceder de la manera siguiente:

1. El conjunto de entrenamiento se utiliza para generar los diferentes modelos candidatos.
2. Para cada modelo candidato generado se obtiene su desempeño con el conjunto de validación. Estos dos primeros pasos se llevan mediante validación-cruzada.
3. Después de llevar a cabo validación-cruzada con la búsqueda de las mejores opciones para todos los modelos candidatos, se selecciona el mejor de ellos.
4. El mejor candidato seleccionado durante el proceso de validación-cruzada se entrena finalmente con los mismos parámetros que lo hicieron el mejor de ellos, pero en el "nuevo conjunto de entrenamiento" que consiste al unir los conjuntos de entrenamiento y validación iniciales. Este nuevo proceso de entrenamiento ya no implica la validación-cruzada.
5. El modelo generado en el paso anterior se evalúa en el conjunto de prueba para obtener su desempeño final.

modelo propuesto:

$y = a + bx$

Regresión Lineal mediante el método de mínimos cuadrados



De aquí el nombre de método de mínimos cuadrados.

y_k : valor observado real, dado x_k.

$\hat{y}_k$: predicción del valor observado y_k.

$\hat{y}_k = \hat{a} + \hat{b}x_k$

Error o Residuo de cada dato x_k:

$$\varepsilon_k = y_k - \hat{y}_k$$

$$SSE = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} \left(y_k - \hat{a} - \hat{b}x_k\right)^2$$

Así, el objetivo será encontrar los parámetros a y b de forma tal que la suma de todos estos errores sea la mínima posible.

Función de Costo y el Problema de Optimización

El problema de encontrar las constantes a y b lo podemos ver como un problema de optimización, en particular de minimizar la función de costo $J(a, b)$, donde:

$$J(a,b) = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - a - bx_k)^2$$

Es decir,

$$\min_{a,b} J(a,b) = \min_{a,b} \sum_{k=1}^{n} (y_k - a - bx_k)^2$$

Obtengamos ahora los valores de los parámetros a y b:

Consideremos entonces la función de costo J:

$$J(a,b) = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - a - bx_k)^2$$

Recordemos de nuestro curso de Cálculo que para minimizar el valor de J, donde los parámetros a y b son nuestras variables a determinar, requerimos resolver el siguiente sistema de ecuaciones obtenido de las primeras derivadas parciales:

$$\begin{cases} \dfrac{\partial L}{\partial a} = -2\displaystyle\sum_{k=1}^{n} (y_k - a - bx_k) = 0 \\ \dfrac{\partial L}{\partial b} = -2\displaystyle\sum_{k=1}^{n} (y_k - a - bx_k)x_k = 0 \end{cases}$$

Simplificando y resolviendo el sistema de ecuaciones 2×2 resultante, llegamos a las llamadas ecuaciones normales de mínimos cuadrados:

$$\begin{cases} na + b\displaystyle\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k \\ a\displaystyle\sum_{k=1}^{n} x_k + b\sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k \end{cases}$$

Y al resolverlas, obtenemos que el modelo óptimo de regresión lineal $\hat{y} = \hat{a} + \hat{b}x$ dado como:

$$\hat{b} = \frac{\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}}{\sum_{k=1}^{n} x_k^2 - n(\bar{x})^2}$$

Pendiente

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Ordenada en el origen

donde $\bar{x}$ y $\bar{y}$ son los valores promedios de $\{x_k\}$ y $\{y_k\}$, respectivamente.

Conceptos estadísticos: Varianza y Covarianza

Varianza

Simplificando algebraicamente se puede obtener la siguiente igualdad:

$$\sum_{k=1}^{n}(x_k-\bar{x})^2=\sum_{k=1}^{n}x_k^2-n(\bar{x})^2$$

que al dividir entre $n-1$, obtenemos lo que en Estadística se conoce como la varianza insesgada de la variable aleatoria X, cuyos datos muestrales están dados por el conjunto $\{x_k\}_{k=1}^{n}$, y denotado $Var[X]$, o bien, σ_X^2:

$$Var[X]=\frac{1}{n-1}\left\{\sum_{k=1}^{n}x_k^2-n(\bar{x})^2\right\}$$

Recordemos del curso de Estadística que si σ_X^2 es la varianza de una variable aleatoria X, entonces σ_X es llama desviación estándar de X.

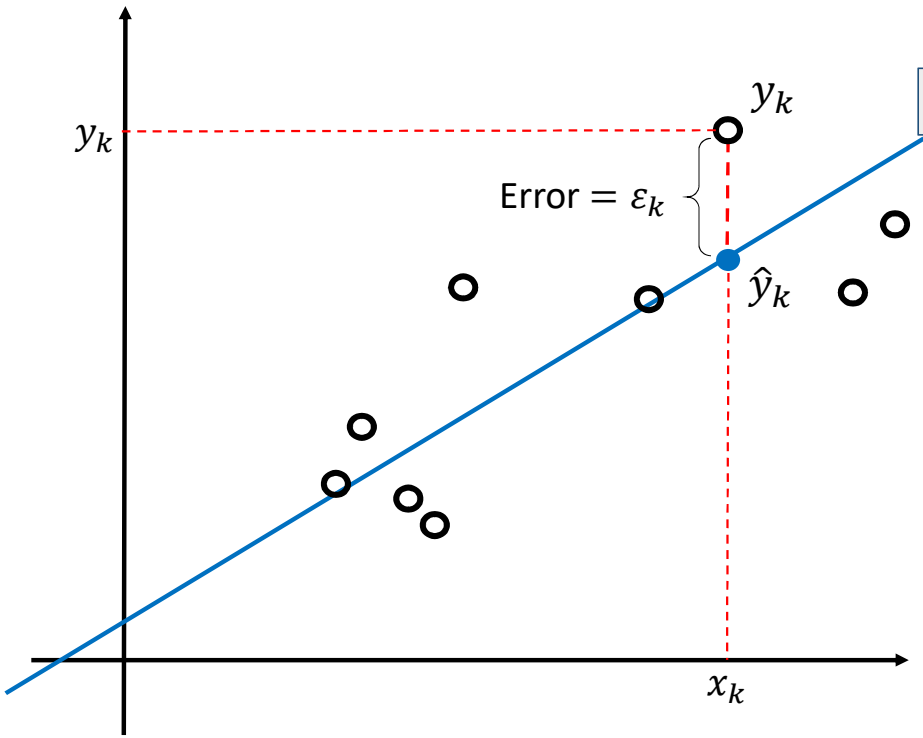
Covarianza

Análogamente, algebraicamente se muestra que:

$$\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})=\sum_{k=1}^{n}x_ky_k-n\bar{x}\bar{y}$$

Y nuevamente al dividir entre $n-1$, obtenemos lo que en Estadística es conocido como la covarianza insesgada de las variables aleatorias X y Y, y la denotamos $Cov[x,y]$, o bien σ_{XY}:

$$Cov[x,y]=\frac{1}{n-1}\left\{\sum_{k=1}^{n}x_ky_k-n\bar{x}\bar{y}\right\}$$



Error Cuadrático Medio (ECM)
Mean Squared Error (MSE)

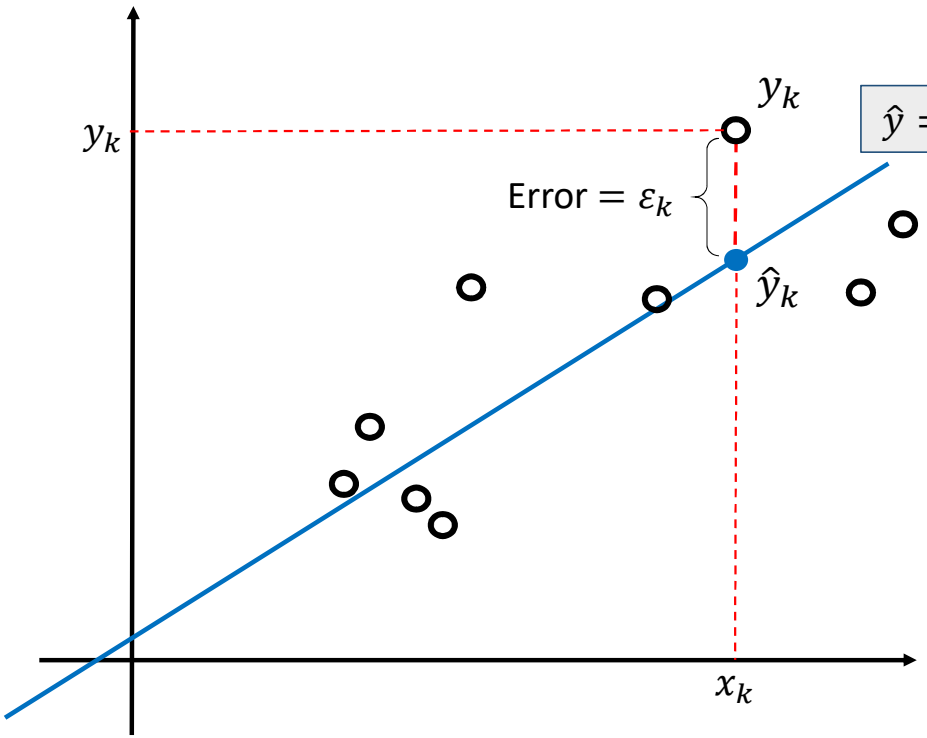
Definimos ahora un concepto ampliamente utilizado en análisis estadísticos y que por ahora nos dará una buena medida del desempeño del modelo de regresión lineal: el Error Cuadrático Medio, MSE por sus siglas en inglés:

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{k=1}^{n} \varepsilon_k^2 = \frac{1}{n}\sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

Así, vemos que el MSE es el valor promedio de la suma del cuadrado de los residuos SSE.

En general, MSE se puede utilizar como una media de la eficiencia del modelo, así como una medida comparativa con respecto a otros modelos.

Es el equivalente a la varianza de un conjunto de datos.



Raíz del Error Cuadrático Medio
Root Mean Squared Error (RMS)

También se puede usar la raíz cuadrada del error cuadrático medio, es decir:

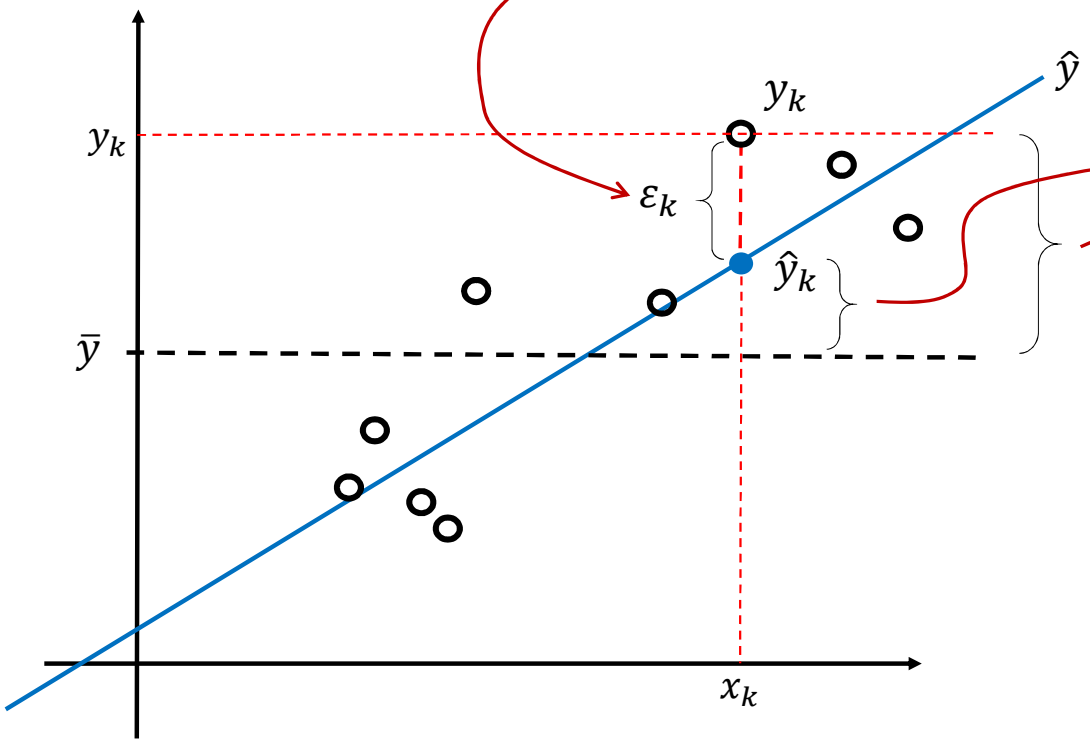
$$E_{RMS} \equiv RMS = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}$$

La ventaja de este error es que está en las mismas unidades y escala que los datos originales.

También se denota como $RMSE$.

Es el equivalente de la desviación estándar de un conjunto de datos.

Sumas de Cuadrados



$$SSE = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

$$SSR = \sum_{k=1}^{n} (\bar{y} - \hat{y}_k)^2$$

$$S_{yy} = \sum_{k=1}^{n} (y_k - \bar{y})^2$$

$$S_{yy} = SSE + SSR$$

SSE : Suma de cuadrados de los errores/residuos.
SSR: Suma de cuadrados de regresión.
Syy : Suma de cuadrados de la variabilidad total.

Métricas para la medición de errores : modelos de Regresión

Suma del Cuadrado de los Errores

$$SSE = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

Suma de Cuadrados de la Variabilidad Total

$$S_{yy} = \sum_{k=1}^{n} (y_k - \bar{y})^2$$

Error Cuadrático Medio

$$MSE = \frac{SSE}{n} = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

$$R^2 = \rho^2 = 1 - \frac{SSE}{S_{yy}}$$

$$R_{ajustada} = \bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

Raíz del Error Cuadrático Medio

$$RMSE \equiv RMS = \sqrt{MSE}$$

Errores o Residuos

$$\varepsilon_k = y_k - \hat{y}_k$$

Promedio de los Errores Absolutos

$$MAE = \frac{1}{n} \sum_{k=1}^{n} |y_k - \hat{y}_k|$$

Promedio de los Errores Porcentuales Absolutos

$$MAPE = \frac{100\%}{n} \sum_{k=1}^{n} \left| \frac{y_k - \hat{y}_k}{y_k} \right|$$

y_k : valores reales de salida
$\hat{y}_k$: valores pronosticados
$\bar{y}$: valor promedio del conjunto de entrenamiento (Train)
n : total de datos de la muestra (Train, Val o Test)
k : total de variables de entrada

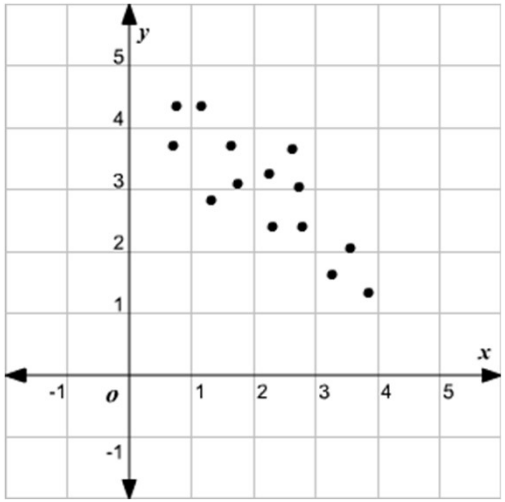
Coeficiente de Correlación de Pearson

Ya vimos que el error cuadrático medio MSE, es una medida que nos dice qué tan bueno puede ser el modelo de regresión lineal obtenido a partir de un conjunto de datos de entrada muestrales $\{(x_k, y_k)\}_{k=1}^{n}$.

Se desea tener ahora otra medida que además nos hable sobre la relación en la cual se encuentran las variables involucradas.

Es decir se requiere una medida que nos diga qué tan bien están correlacionadas las variables aleatorias X y Y.

Existen varias formas de medir dicha correlación, siendo el coeficiente de correlación de Pearson uno de los más conocidos y utilizado.



Definimos el coeficiente de correlación de Pearson de dos variables aleatorias X y Y como:

$$\rho = \frac{cov[X,Y]}{\sqrt{Var[X]\,Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$$

- Se tiene que $-1 \leq \rho \leq +1$.

Para una primera lectura del valor de $|\rho|$, podemos considerar los siguientes criterios:

- Correlación nula o casi nula: valores entre 0 y 0.1
- Correlación débil: valores entre 0.1 y 0.3
- Correlación moderada: valores entre 0.3 y 0.6
- Correlación fuerte (datos relacionados con estudios de personas): arriba de 0.6
- Correlación fuerte (datos relacionados con estudios de ingeniería): arriba de 0.75

Estos intervalos pueden variar dependiendo del área de estudio o investigadores.

Conceptos estadísticos: Varianza y Covarianza

Varianza

Simplificando algebraicamente se puede obtener la siguiente igualdad:

$$\sum_{k=1}^{n}(x_k-\bar{x})^2=\sum_{k=1}^{n}x_k^2-n(\bar{x})^2$$

que al dividir entre $n-1$, obtenemos lo que en Estadística se conoce como la varianza insesgada de la variable aleatoria X, cuyos datos muestrales están dados por el conjunto $\{x_k\}_{k=1}^{n}$, y denotado $Var[X]$, o bien, σ_X^2:

$$Var[X]=\frac{1}{n-1}\left\{\sum_{k=1}^{n}x_k^2-n(\bar{x})^2\right\}$$

Recordemos del curso de Estadística que si σ_X^2 es la varianza de una variable aleatoria X, entonces σ_X es llama desviación estándar de X.

Covarianza

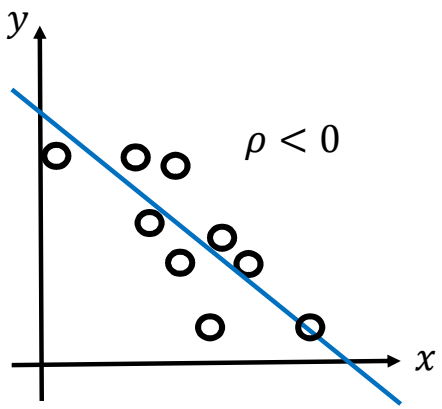
Análogamente, algebraicamente se muestra que:

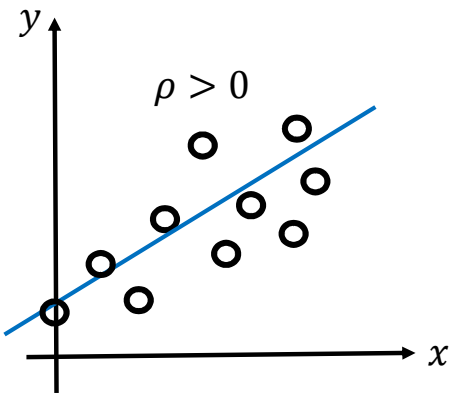
$$\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})=\sum_{k=1}^{n}x_ky_k-n\bar{x}\bar{y}$$

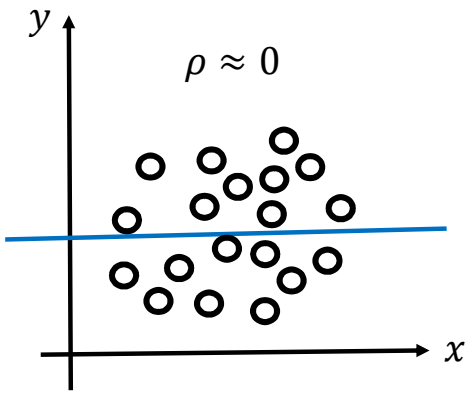
Y nuevamente al dividir entre $n-1$, obtenemos lo que en Estadística es conocido como la covarianza insesgada de las variables aleatorias X y Y, y la denotamos $Cov[x,y]$, o bien σ_{XY}:

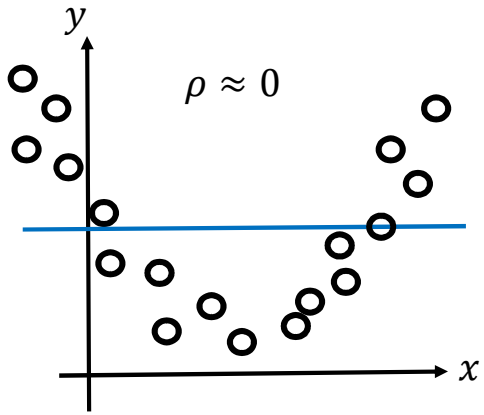
$$Cov[x,y]=\frac{1}{n-1}\left\{\sum_{k=1}^{n}x_ky_k-n\bar{x}\bar{y}\right\}$$

Interpretación geométrica de la correlación de la variable dependiente y con la variable independiente x, dado un conjunto de datos bidimensionales que los aproximan $\{(x_k, y_k)\}_{k=1}^{n}$.









Coeficiente de Determinación

Aunque existen diferentes definiciones, la más común para el coeficiente de determinación es la del cuadrado del coeficiente de correlación de Pearson, y la denotamos como R^2:

$$R^2 = \rho^2 = \frac{(cov[X,Y])^2}{Var[X]\,Var[Y]} = 1 - \frac{SSE}{S_{yy}}$$

El coeficiente de determinación es un estadístico cuyo valor nos dice la proporción del modelo que es explicado por la variables consideradas.

Es decir, es la proporción de la variabilidad de la variable dependiente que es explicada por la variable independiente, o variables independientes para el caso del modelo multilineal que se verá más adelante.

Así, en el caso del modelo lineal simple, el coeficiente de correlación de Pearson también se puede obtener como: $\rho = \sqrt{1 - \frac{SSE}{S_{yy}}}$.

Maestría en Inteligencia Artificial Aplicada

Modelo de Regresión Lineal Múltiple

Inteligencia Artificial y Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales

ITESM

Campus Guadalajara

Regresión Lineal Múltiple

La hipótesis general del modelo de **regresión lineal múltiple** podemos representarla como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

donde cada variable independiente x_k representa una característica del problema que trata de explicar el comportamiento de la variable/característica dependiente y_k.

La ordenada en el origen β_0 representa el valor esperado de y cuando todas las características independientes son iguales a cero.

Además, se incluye un error o residuo ε en la hipótesis que nos recuerda que la relación en general no será perfecta.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

Representación Matricial

Datos de entrenamiento:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_m x_{1m} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_m x_{2m} + \varepsilon_2 \\ &\vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \quad\quad\quad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_m x_{nm} + \varepsilon_n \end{aligned}$$

matricialmente

$$Y = X\beta + \varepsilon$$

Homogeneizando x_0

Vector/Factor x_1

Vector/Factor x_2

Vector/Factor x_m

1° dato de entrenamiento

n-ésimo dato de entrenamiento

$$X_{n\times(m+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

Matriz de los valores de entrada de los datos de entrenamiento

donde

$$Y_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Vector de las salidas de los datos de entrenamiento

$$\beta_{(m+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

Vector de los parámetros del modelo a determinar

$$\varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Vector de los errores o residuos

Regresión Lineal Múltiple
Función de Costo / Problema de Optimización

Así, el problema de encontrar las constantes a y b nos lleva a un problema de optimización, en particular el de minimizar la función de costo $J(\beta_0, \beta_1, \dots, \beta_m)$ donde:

$$J(\beta_0, \beta_1, \dots, \beta_m) = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - \beta_0 - \beta_1 x_{k1} - \beta_2 x_{k2} - \cdots - \beta_m x_{km})^2 = \sum_{k=1}^{n} (Y_k - X_k \beta)^T (Y_k - X_k \beta)$$

Es decir,

$$\min_{\beta_k} J(\beta_0, \beta_1, \dots, \beta_m) = \min_{\beta_k} \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

donde y_k son los valores reales observados y $\hat{y}_k$ son las predicciones del modelo:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_m x_m$$

Modelo Lineal

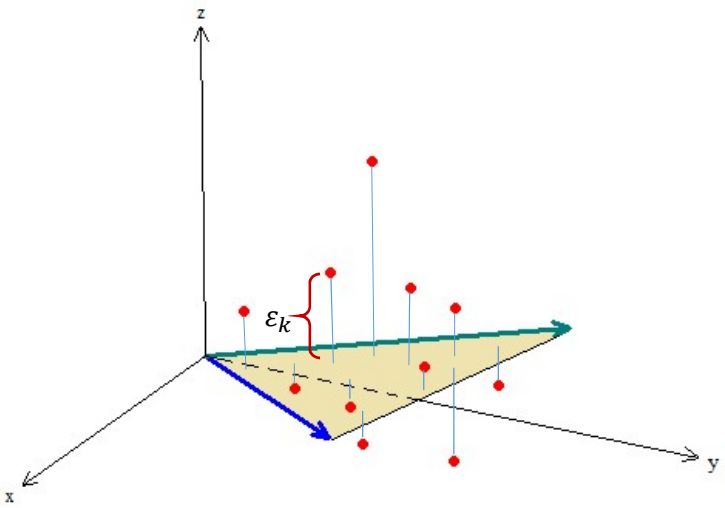


Figura: Buscar el mejor plano que aproxime los puntos de entrada dados.

Sí existe una solución analítca del problema dada por la llamada matriz pseudoinversa de Moore-Penrose.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Sin embargo cuando se tiene una gran cantidad de registros y variables, la solución analítica en general no es viable debido al costo computacional y los errores numéricos involucrados en el proceso, por lo que se prefiere resolver este problema mediante métodos numéricos iterativos, como el de mínimos cuadrados.

Regresión Lineal Múltiple mediante el método de mínimos cuadrados

En resumen, dado un conjunto de datos muestrales

$$\{(\vec{x}_k, y_k)\}_{k=1}^{N}$$

se desea encontrar el hiperplano que tenga el mejor ajuste en dichos datos. Donde cada $\vec{x}_k$ es un vector de entrada de M coordenadas cuyo valor real observado es y_k.

Entonces, se desean encontrar los mejores coeficientes $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_M$ del hiperplano de la forma:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_M x_M$$

que mejor aproxime los datos muestrales mediante sus valores de predicción y de acuerdo a la siguiente función de costo.

Función de Costo o Error: diferencia de cuadrados

Definimos una **función de costo** o **función error** mediante la suma de los cuadrados de las diferencias entre lo observado y lo estimado:

$$J = \sum_{k=1}^{N} (y_k - \hat{y}_k)^2$$

Problema de Optimización mediante mínimos cuadrados

El objetivo es minimizar el valor de dicha función mediante el ajuste de sus coeficientes por el método de mínimos cuadrados, es decir,

$$\min_{\{\beta_k\}} \{ J \} = \min_{\{\beta_k\}} \sum_{k=1}^{N} (y_k - \beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M)^2$$

Algunos autores consideran la mitad de este valor J como la función error, o inclusive el promedio. Sin embargo, en cualquier caso las conclusiones son equivalentes.

También pueden usarse las siguientes expresiones como funciones de costo:

Error Cuadrático Medio (Mean Squared Error-MSE)

Como una mejor medida del error, este puede promediarse para obtener el error cuadrático medio:

$$E_{MSE} \equiv MSE = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2$$

O bien

$$MSE = \frac{SSE}{N}$$

Este error es muy usado en diferentes tipos de pruebas y análisis estadísticos que requeriremos posteriormente.

Dicho promedio permitirá comparar errores entre muestras de diferente tamaño.

Sería el equivalente a la varianza.

Raíz del Error Cuadrático Medio (Root Mean Squared Error-RMS)

Inclusive nos será útil también en ocasiones la raíz cuadrada del error cuadrático medio, es decir,

$$E_{RMS} \equiv RMS = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}$$

La ventaja de este error es que está en las mismas unidades y escala que los datos originales.

NOTA: También se denota como $RMSE$.

Sería el equivalente a la desviación estándar.

Coeficiente de determinación R^2:

$$R^2 = 1 - \frac{SSE}{S_{yy}} = 1 - \frac{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{N}\left(y_j - \bar{y}\right)^2}$$

- El coeficiente de determinación nos habla del porcentaje de variabilidad de la variable dependiente que queda explicada por el modelo a través de las variables independientes utilizadas.
- Existen en la literatura varias definiciones diferentes de R^2, no siempre equivalentes, pero que buscan el mismo objetivo.
- En el caso del modelo lineal simple, R^2 es equivalente al cuadrado del coeficiente de correlación de Pearson.
- Observa que R^2 nos habla de qué tanto es mejor nuestro modelo de regresión usando $\hat{y}$ como predicción, a que si usáramos simplemente el promedio $\bar{y}$.

Coeficiente de Determinación Ajustado

En el caso de modelos con $k+1$, factores, donde consideramos k variables linealmente independientes, el coeficiente de determinación R^2 estará incrementando su valor aún cuando esto no signifique un mejor ajuste del modelo. Es decir, podemos estar agregando variables que en realidad no son estadísticamente significativas, y queremos por lo tanto evitar sobreestimar la importancia de las variables independientes.

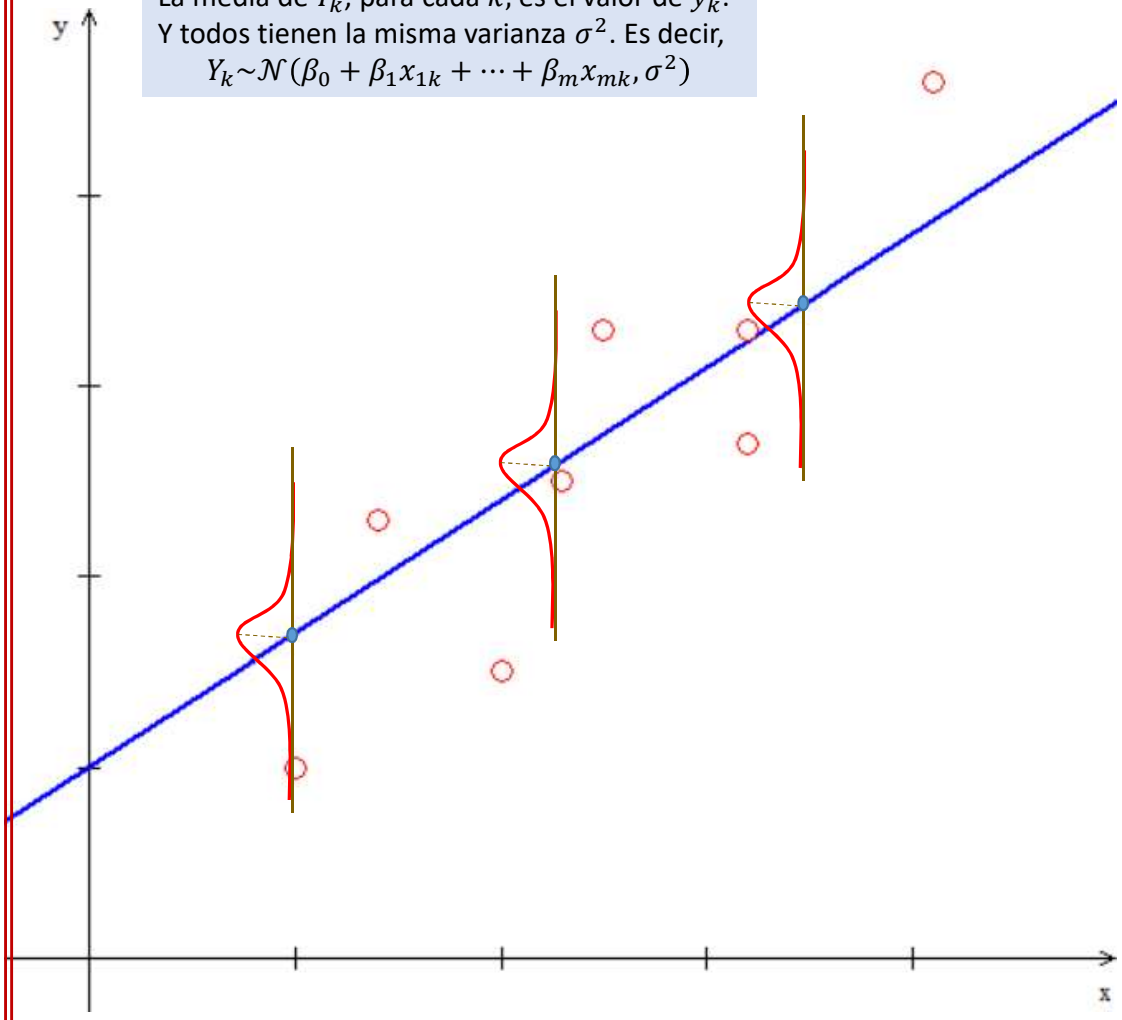
Por ello, se define el coeficiente de determinación ajustado, denotado $\bar{R}^2$ y el cual considera el total de variables independientes utilizadas en el modelo.

El **Coeficiente de determinación ajustado** $\bar{R}^2$, para una muestra de tamaño N y cuyo modelo tiene k variables independientes, se define como:

$$R_{ajustada} = \bar{R}^2 = 1 - \frac{N-1}{N-k-1}(1-R^2)$$

Observa que tener k variables independientes implica tener $(k+1)$ factores, pues se incluye además a la variable dependiente como factor.

La media de Y_k, para cada k, es el valor de $\hat{y}_k$.
Y todos tienen la misma varianza σ^2. Es decir,
$Y_k \sim \mathcal{N}(\beta_0 + \beta_1 x_{1k} + \cdots + \beta_m x_{mk}, \sigma^2)$

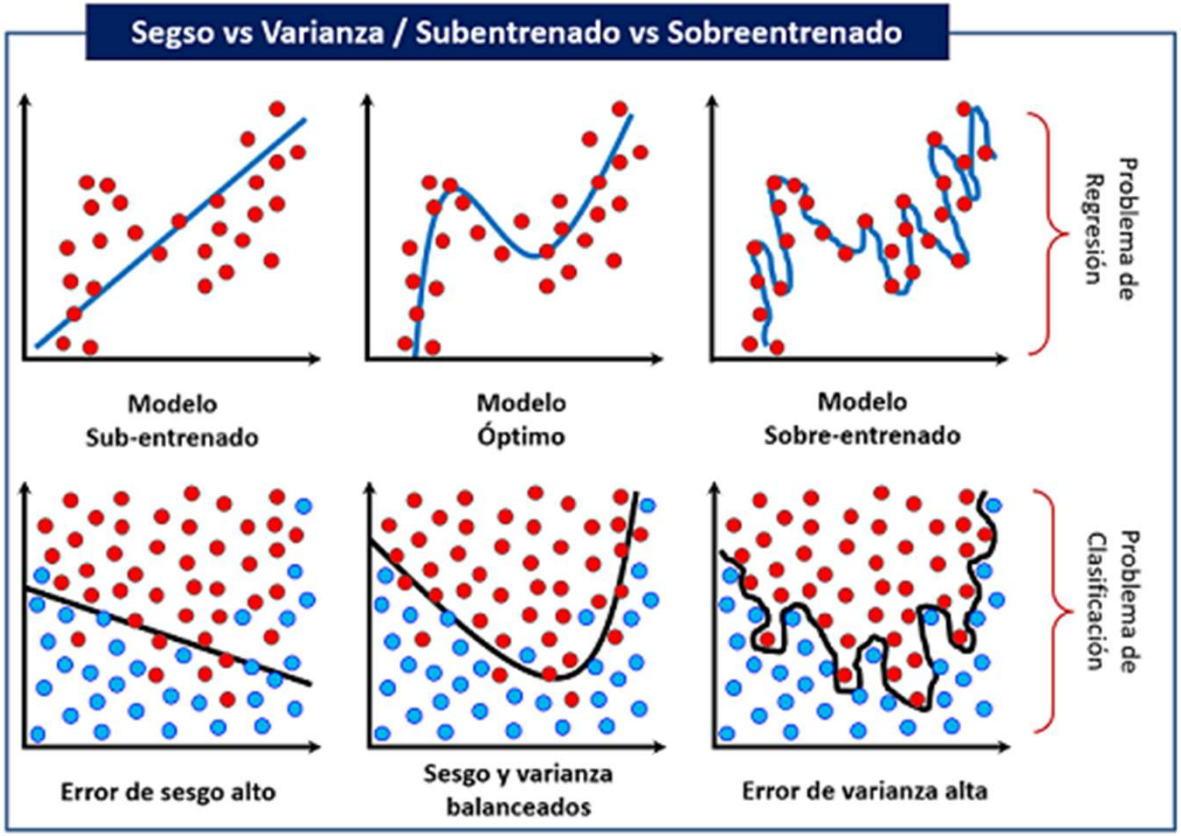


Para aplicar el modelo de regresión lineal entre dos variables aleatorias X y Y, estrictamente se deben cumplir los siguientes supuestos:

- **Linealidad:** La relación entre ambas variables debe ser lineal.
- **Normalidad:** Los valores de Y están distribuidos normalmente para cada valor de X, es decir, los residuos.
- **Homocedasticidad:** La variación alrededor de la línea de regresión tiene un mismo valor constante σ^2 para todos los valores de X.
- **Independencia de los residuos:** Los residuos o errores $\varepsilon = y - \hat{y}$ deben ser independientes para cada valor de X.

Sub-entrenamiento (underfitting)
vs
Ajuste óptimo
vs
Sobre-entrenamiento (overfitting)

Recuerda que incrementar la complejidad del modelo puede ayudar a mejorar la predicción de la variable de salida y evitar el subentrenamiento, pero a suvez debemos cuidar no caer ahora en un modelo sobreentrenado.



Prueba F global para el Modelo de Regresión

Dado un conjunto muestral de N datos, se desea probar la significancia de la relación de regresión entre la variable y, y las k variables independientes x_1, x_2, …, x_k del modelo lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Llamado usualmente ANOVA para la significación de la regresión en RLM.

Paso 1:

Hipótesis Nula:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Hipótesis Alternativa:

H_1 : al menos uno de β_1, β_2,…, β_k es diferente de cero.

Paso 2:

Calcular el estadístico de prueba F global:

$$F_{modelo} = \frac{SSR/k}{SSE/(N-k-1)}$$

Paso 3:

Calcular el valor crítico $F_{[\alpha;\ k,\ N-k-1]}$, de la distribución F de Fisher para un nivel de significancia α con k grados de libertad para el numerador y $N-k-1$ grados de libertad para el denominador.

Paso 4:

La prueba es significativa con un nivel de significancia α, si se cumple alguna de las siguientes condiciones:

- $F_{modelo} > F_{[\alpha;\ k,\ N-k-1]}$
- $valor\ p < \alpha$

Prueba t de significancia para una variable independiente x_j

La prueba F global anterior solo nos dice que al menos una de las variables independientes es significativa en un modelo de regresión, la siguiente prueba nos dirá cuál o cuáles de ellas lo serían, lo cual podría simplificar el modelo.

Dado un conjunto muestral de N datos, se desea probar la significancia de la relación de regresión entre la variable dependiente y, y una de las k variables independientes del modelo lineal, por ejemplo x_j, donde $j \in \{1,2,\dots,k\}$, :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Paso 1: Para una $j \in \{1,2,\dots,k\}$

Hipótesis Nula:

$$H_0 : \beta_j = 0$$

Hipótesis Alternativa:

$$H_1 : \beta_j \neq 0$$

Paso 2: Calcular el estadístico de prueba t:

$$t_0 = \frac{b_j}{s_{b_j}}$$

donde s_{b_j} es el error estándar de la estimación b_j.

Paso 3:

Calcular el valor crítico $t_{[\alpha/2;\ N-k-1]}$, de la distribución $t\text{-}Student$ para un nivel de significancia α con $N - k - 1$ grados de libertad.

Paso 4:

La prueba es significativa con un nivel de significancia α, si se cumple alguna de las siguientes condiciones:

- $|t_0| > t_{[\alpha/2;\ N-k-1]}$
- $valor\ p < \alpha$

- La estimación puntual de σ_{b_j} se denomina **error estándar de la estimación** b_j, y la denotaremos como s_{b_j}. La fórmula para calcularla es:

$$s_{b_j} = \sqrt{\frac{SSE \cdot c_{jj}}{N - k - 1}}$$

donde $SSE = \sum_{j=1}^{N} \left(y_j - \hat{y}_j\right)^2$ es la suma de los cuadrados de los residuos; b_j es el valor de la estimación puntual de β_j obtenido mediante el modelo de regresión de mínimos cuadrados.

Además, c_{jj} es el elemento diagonal j-$ésimo$ de la matriz $(X^T X)^{-1}$, donde X es la matriz de los valores de entrada de los N datos de entrenamiento:

$$X_{N\times(m+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad (X^T X)^{-1} = \begin{bmatrix} c_{00} & & & \\ & c_{11} & & \\ & & \ddots & \\ & & & c_{mm} \end{bmatrix}$$

- En la prueba de significancia para una variable independiente, hay que tomar en cuenta que dicha prueba de significancia dependerá de qué otras variables independientes se están considerando en el modelo.

- Se supone que la población de todos los valores posibles de la estimación puntual de mínimos cuadrados β_j está normalmente distribuida con media β_j y desviación estándar σ_{b_j}. Esto se puede lograr con la transformación de las variables originales para que tengan una distribución aproximadamente gaussiana.

- A medida que es más pequeño el nivel de significancia α mediante el cual se rechaza H_0, será más fuerte la evidencia de que la variable independiente x_j está relacionada significativamente con la variable dependiente y en el modelo de regresión.

- Generalmente se consideran valores para el nivel de significancia α de 0.05 y 0.01, pero pueden obviamente usarse otros valores.

Modelos no-lineales en las variables de entrada: Polinomial

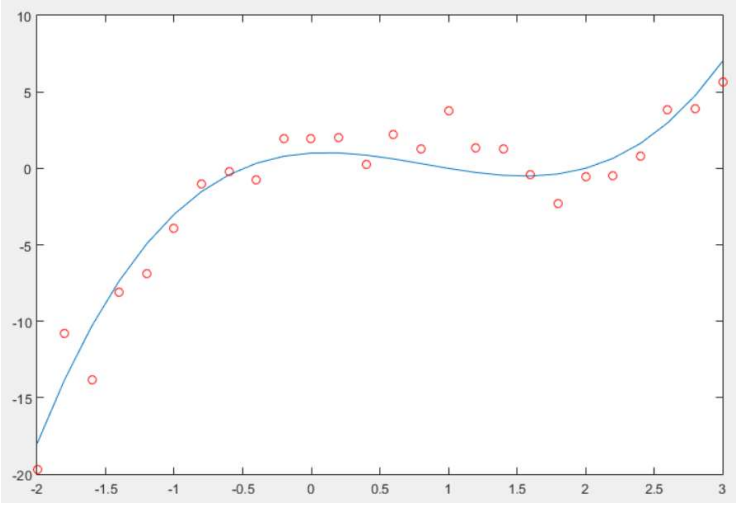
Supongamos que tenemos un conjunto de datos muestrales de entrada de dos dimensiones:

$$\{(x_k, y_k)\}_{k=1}^{N}$$

Para ejemplificar, los mostramos como círculos rojos en la imagen que se ilustra.

Supongamos además que teóricamente estos datos muestrales tienen un comportamiento polinomial de grado m, es decir:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m$$



Nuestro objetivo está en la siguiente dirección:

Objetivo:

Dado un nuevo valor de entrada no conocido $\hat{x}$, se desea pronosticar el valor de su imagen $\hat{y}$.

O bien, se desea determinar la función polinomial de grado m que mejor ajusta los datos muestrales.

Podemos generar modelos polinomiales agregando la relación no lineal entre columnas deseadas:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

Modelo lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}^2 \\ 1 & x_{21} & x_{22} & x_{21}^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 \end{bmatrix}$$

Modelo cuadrático en las variables de entrada:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{11}^3 \\ 1 & x_{21} & x_{22} & x_{21}^2 & x_{21}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 & x_{n1}^3 \end{bmatrix}$$

Modelo cúbico en las variables de entrada:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1^3$$

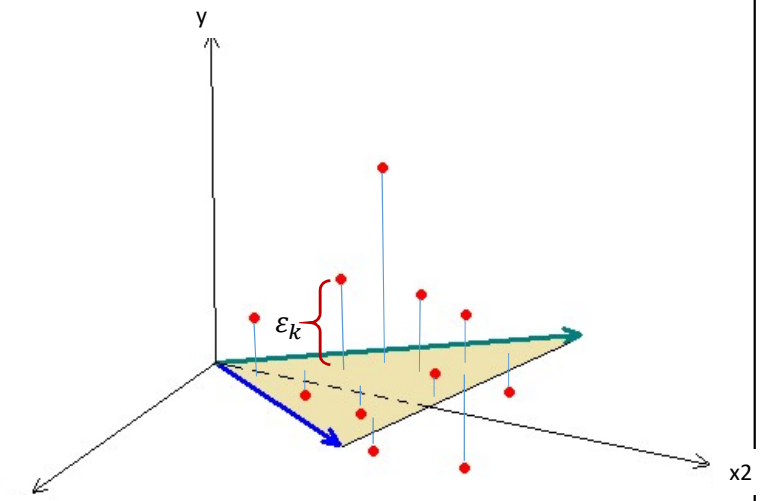
Maestría en Inteligencia Artificial Aplicada

Método del Gradiente Descendente

Inteligencia Artificial y Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales
Tecnológico de Monterrey

Método de Optimización mediante una Función de Costo



En el caso del Modelo de Regresión Lineal:
Se desea encontrar el mejor hiperplano que aproxime los puntos dados.

Dado el conjunto de **datos de entrada**:

$$\{(x_{k1}, x_{k2}, \dots, x_{km}, y_k)\}_{k=1}^{N}$$

deseamos encontrar los parámetros $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$ del modelo:

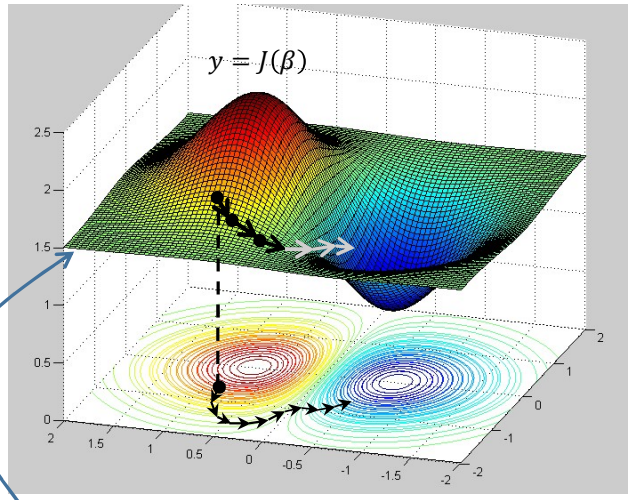
$$\hat{y} = h_\beta(X)$$

donde $X_k = (1, x_{k1}, x_{k2}, \dots, x_{km})$ y que minimice la **función de costo** J, es decir:

$$\min_\beta J(\beta)$$

La función de costo nos estará diciendo qué tan bien el modelo J aproxima los datos dados y será una medida para saber cómo ir mejorándolo.

Método del Gradiente Descendente



Superficie que representa la función de costo.

Denotemos como β el vector (x_1, x_2).

Se desean encontrar las coordenadas de β, que minimicen el valor de la función de costo $J(\beta)$.

En cada paso se desea estar más cerca del mínimo de $J(\beta)$ mediante la siguiente fórmula iterativa:

$$\beta_{(new)} = \beta_{(old)} + \Delta\beta\Big|_{\beta_{(old)}}$$

Para alcanzar el mínimo de la función de costo J, nos apoyaremos en la dirección negativa de su gradiente:

$$\beta_{(new)} = \beta_{(old)} - \eta \cdot \nabla J(\beta)\Big|_{\beta=\beta(old)}$$

donde el escalar $\eta > 0$ es el llamado tamaño de paso (*learning rate*).

Método Iterativo del Gradiente Descendente para Minimizar la función de Costo

Al aplicar un método numérico iterativo para mejorar la función de costo $J(\beta)$ obtenido en un punto $\beta_{(old)}$, se deberá obtener el vector de incremento $\Delta\beta$, tal que:

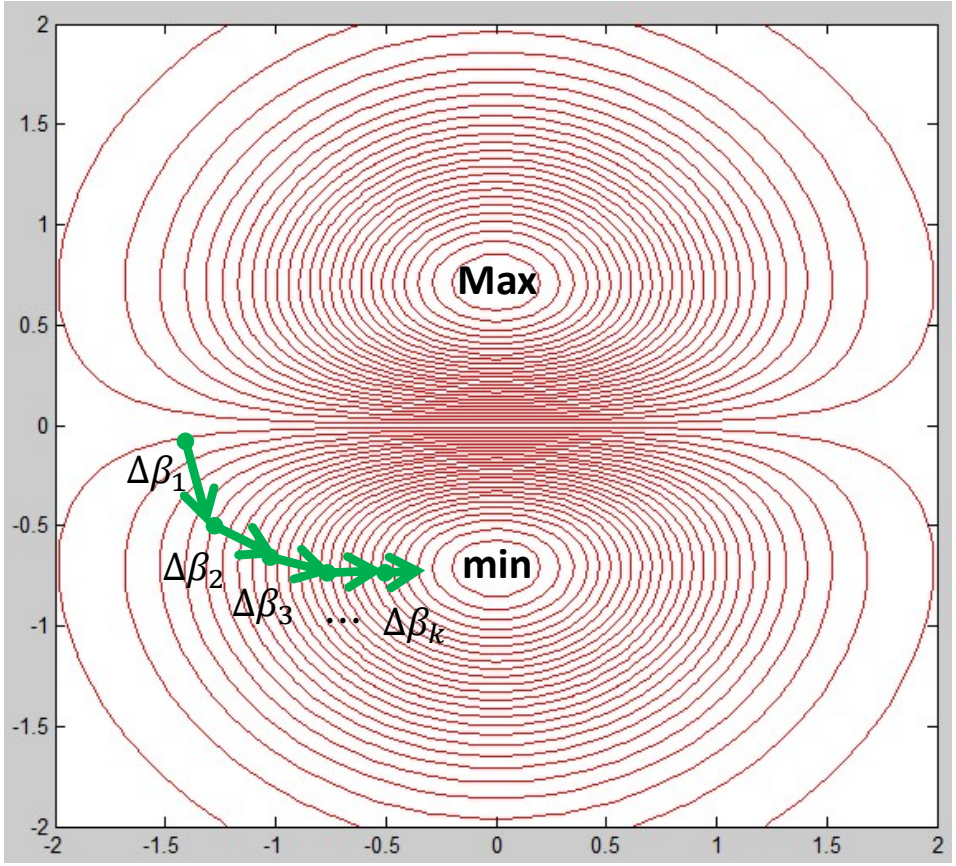
$$\beta_{(new)} = \beta_{(old)} + \Delta\beta\Big|_{\beta_{(old)}}$$

resulte en un mejor valor de $J(\beta)$.
Como es un problema de minimización, podemos utilizar el negativo del vector gradiente, $-\nabla J(\beta)$, para obtener el llamado método del **Gradiente Descendente**:

$$\beta_{(new)} = \beta_{(old)} - \lambda \nabla J(\beta)\Big|_{\beta=\beta(old)}$$

donde λ es el llamado **tamaño de paso (learning rate)**.

- Si $\lambda \gg 0$ el algoritmo puede alejarse del óptimo.
- Si $\lambda \approx 0$ el algoritmo puede ser muy lento.
- λ puede ser dinámico y variar con las iteraciones.



Función de Costo en los Algoritmos de Optimización de modelos de Regresión

- **Funciones de Costo (Loss function):**
  - **Error cuadrático medio – Mean Squared Error – MSE**:

$$\mathcal{L}oss = MSE = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2$$

También se le llama el error L2, o euclidiano.

y_k : valores reales
$\hat{y}_k$: predicciones del modelo usado

$$err = errors\_k = residuos\_k = y_k - \hat{y}_k$$

Existen otras definiciones para la función de costo en los **modelos de regresión**, entre las más comunes:

$$\mathcal{L}oss = \mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2$$

Simplifica el álgebra al calcular las derivadas.

Raíz del Error cuadrático medio – Root Mean Squared Error – RMSE:

$$\mathcal{L} = RMSE = \sqrt{\frac{1}{2N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}$$

Queda en las mismas unidades que los datos originales, pero complica el álgebra en general y en particular cálculo de las derivadas parciales.

Error absoluto medio / L1 / MAE:
(mean absolute error)

$$\mathcal{L} = MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k|$$