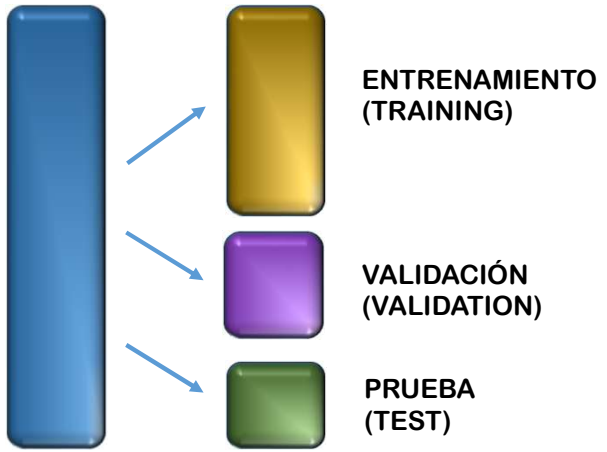
Partición Entrenamiento-Validación-Prueba en Aprendizaje Supervisado

Usualmente se propone una partición de los datos iniciales (universo) en tres conjuntos para la generación de un modelo de aprendizaje automático bajo el criterio de aprendizaje supervisado.

- **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 los pesos del modelo, en un proceso iterativo llamado proceso o etapa de aprendizaje.
- **Datos de Prueba:** para evaluar el desempeño final del modelo.



Maestría en Inteligencia Artificial Aplicada

Modelo de Regresión Lineal Simple

Inteligencia Artificial y Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales

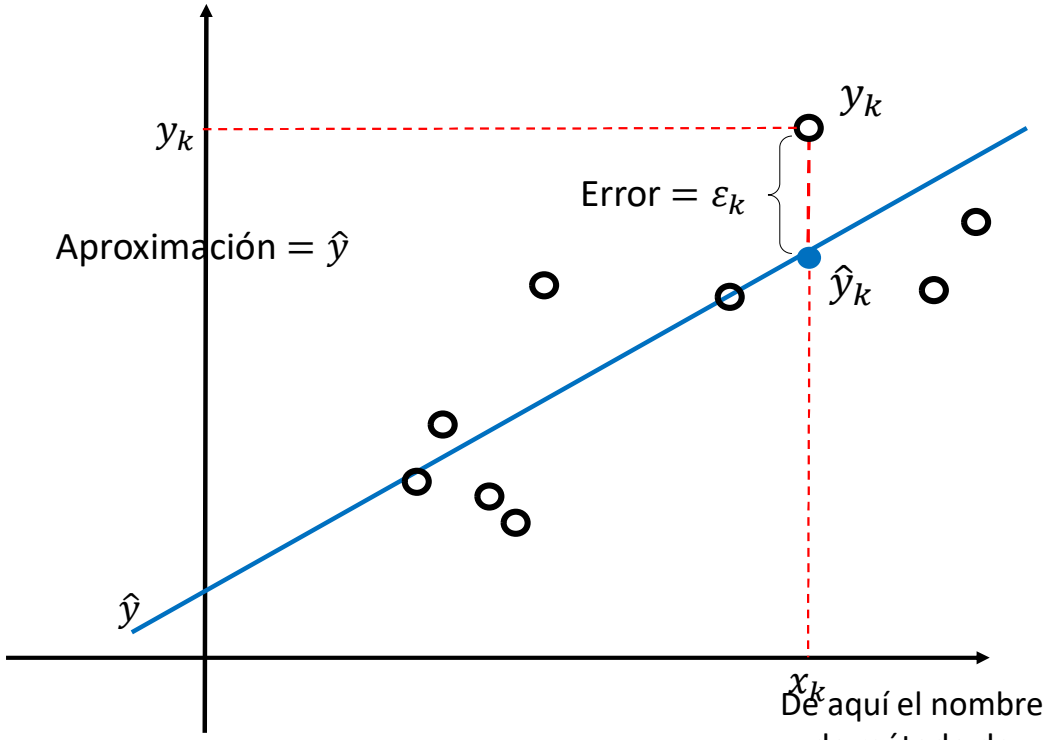
ITESM

Campus Guadalajara

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.

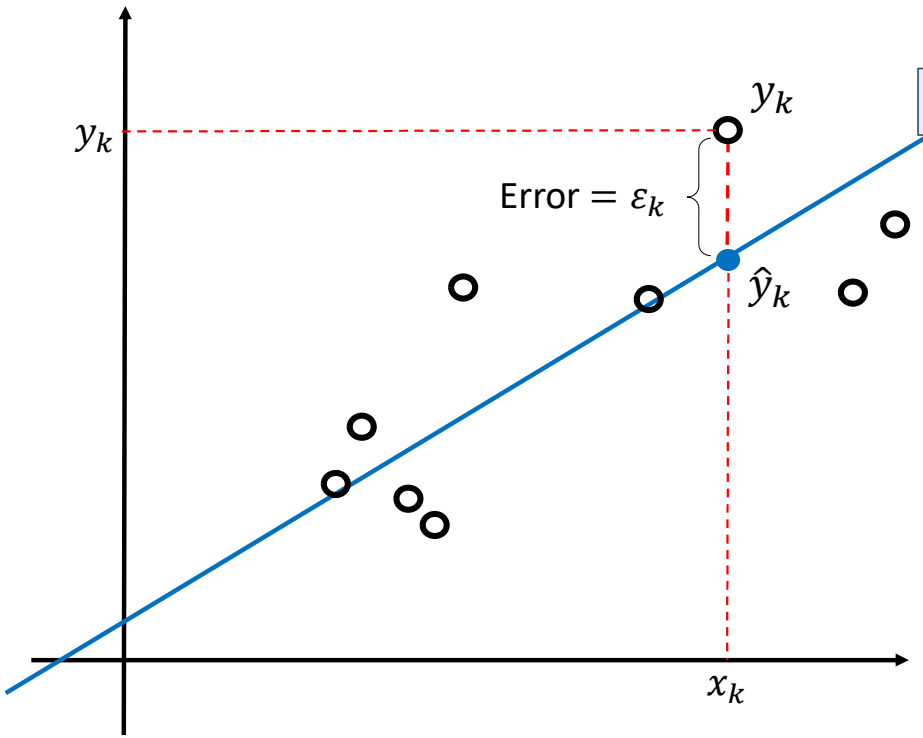
$\hat{y}_k = \hat{a} + \hat{b}x_k$

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.



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

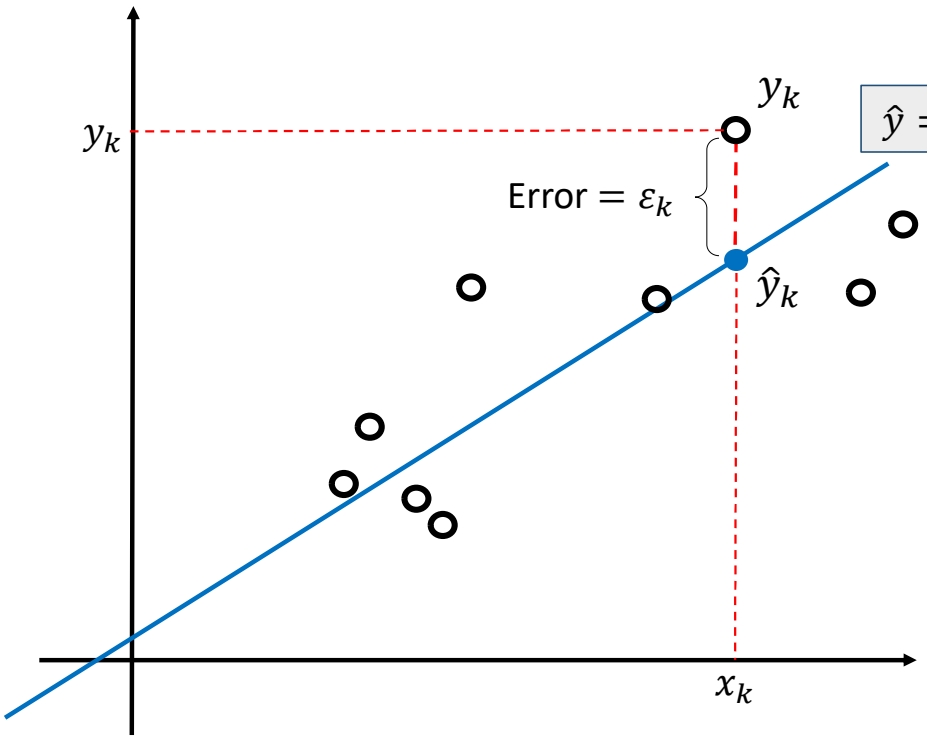
Definimos ahora un concepto ampliamente utilizado en análisis estadísticos y que por ahora nos dará una buena medida del desempeño del modelo de regresión lineal: el Error Cuadrático Medio, MSE por sus siglas en inglés:

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

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 (RMSE)

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

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

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

También se puede denotar como RMS.

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

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

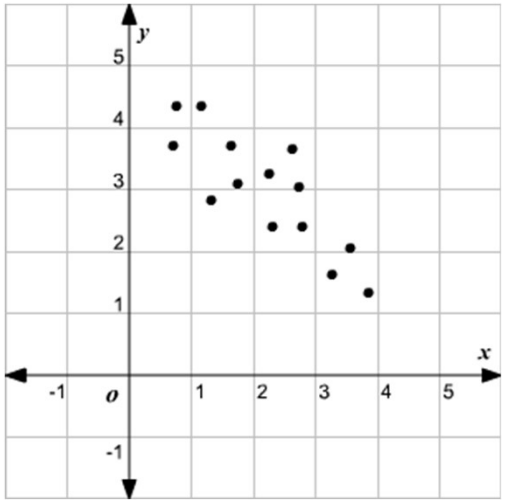
Coeficiente de Correlación de Pearson

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:

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

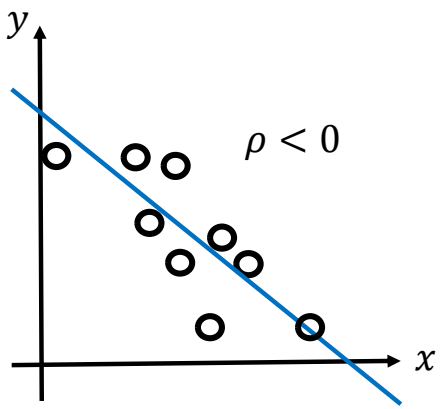
- Se tiene que $-1 \leq \rho \leq +1$.

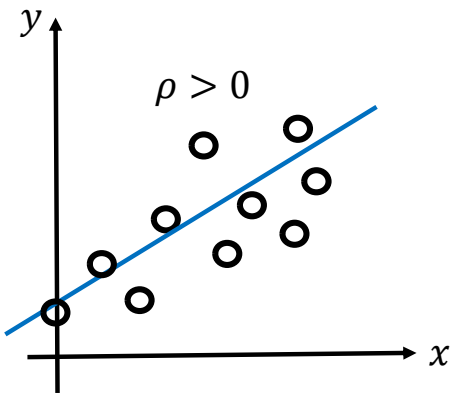
Para una primera lectura del valor de $|\rho|$, podemos considerar los siguientes criterios:

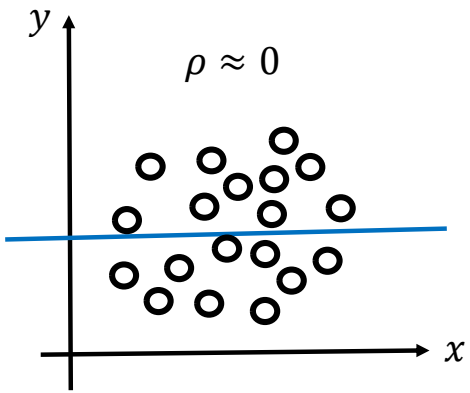
- Correlación nula o casi nula: valores entre 0 y 0.1
- Correlación débil: valores entre 0.1 y 0.3
- Correlación moderada: valores entre 0.3 y 0.6
- Correlación fuerte (datos relacionados con estudios de personas): arriba de 0.6
- Correlación fuerte (datos relacionados con estudios de ingeniería): arriba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

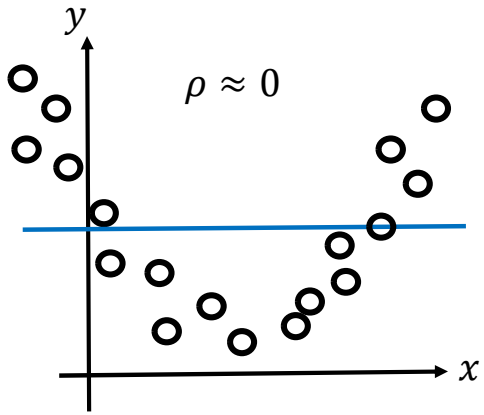
Estos intervalos pueden variar dependiendo del área de estudio o investigadores.

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.

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.

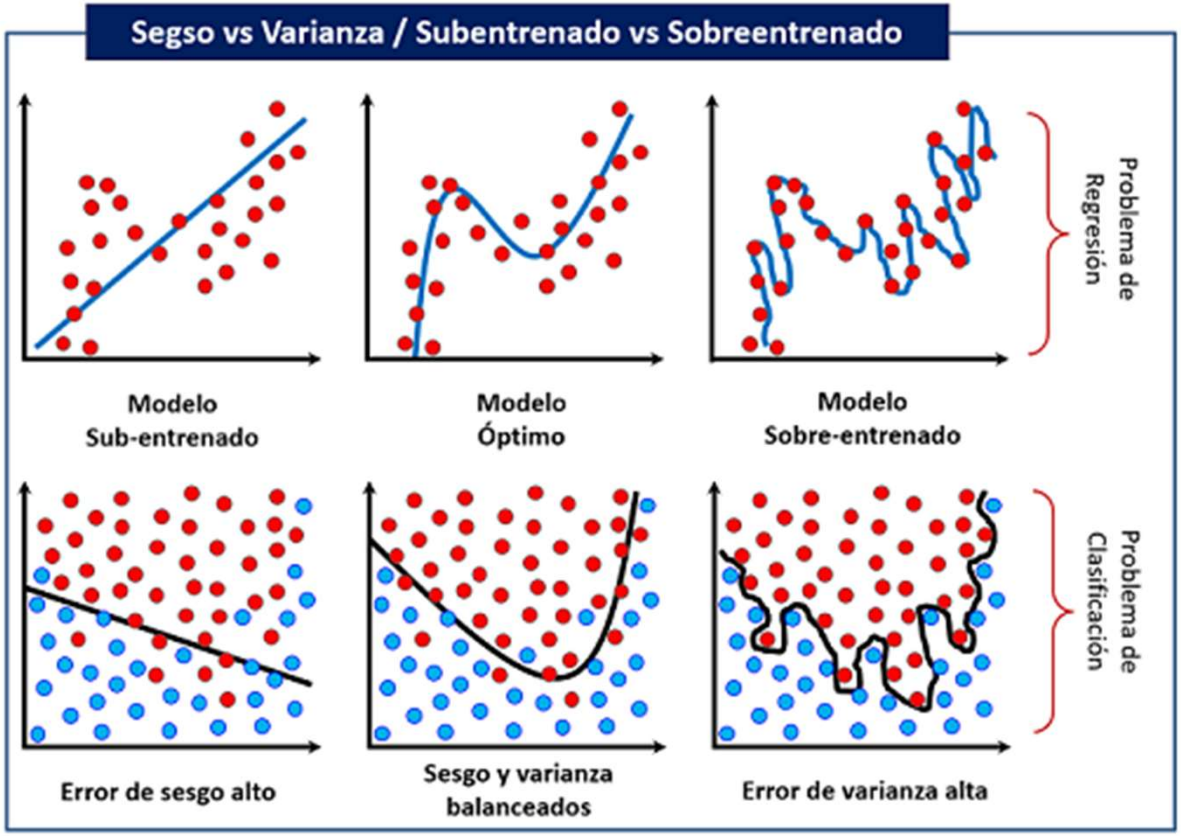
El **Coeficiente de determinación ajustado** $\bar{R}^2$, para una muestra de tamaño N y cuyo modelo tiene k variables independientes, se define como:

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

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

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.



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:

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