Modelos Lineales para Regresión I

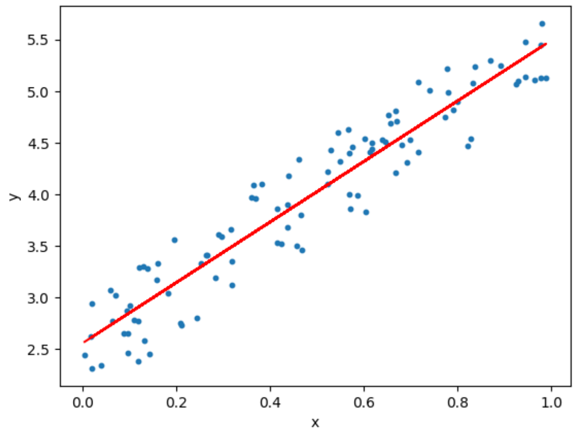
Jonnatan Arias Garcia
jonnatan.arias@utp.edu.co
jariasg@uniquindio.edu.co

David Cardenas peña - dcardenasp@utp.edu.co
Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Regresión

Conjunto de técnicas que se utilizan para predecir un valor numérico basado en la relación entre una o más variables de entrada.

El propósito principal es encontrar la función que mejor describe la relación entre las variables y luego utilizar esa función para hacer predicciones sobre valores futuros.



Definiciones:

Variable Dependiente (Objetivo):

1. La variable dependiente es la que se quiere predecir o explicar en un modelo estadístico o de machine learning. También se conoce como la variable de respuesta o salida.
2. En términos más simples, es la variable que estamos tratando de entender o predecir.

Variable Independiente (Características):

1. Las variables independientes son aquellas que se utilizan para predecir o explicar la variabilidad en la variable dependiente.
2. También se llaman variables predictoras, características o variables explicativas.

Ejemplo de Regresión Lineal:

1. **Variable Dependiente:** Precio de una casa.
2. **Variables Independientes:** Número de habitaciones, área total, ubicación, etc.
3. **Explicación:** Se podría utilizar una regresión lineal para predecir el precio de una casa en función de características como el número de habitaciones, el área total y la ubicación.

Mas ejemplos:

Predicción de Ingresos:

1. **Variable Dependiente:** Ingresos mensuales.
2. **Variables Independientes:** Nivel educativo, años de experiencia laboral, industria, ubicación geográfica, etc.
3. **Explicación:** Un modelo de regresión podría ayudar a predecir los ingresos mensuales de una persona basándose en su nivel educativo, años de experiencia laboral, industria en la que trabaja y su ubicación geográfica.

Estimación de Producción Agrícola:

1. **Variable Dependiente:** Cantidad de cosecha.
2. **Variables Independientes:** Clima, tipo de suelo, cantidad de agua, tipo de cultivo, entre otros.
3. **Explicación:** Un modelo de regresión podría prever la cantidad de cosecha de un cultivo basándose en factores como el clima, el tipo de suelo, la cantidad de agua y el tipo de cultivo plantado.

Pronóstico de Ventas Minoristas:

1. **Variable Dependiente:** Ventas diarias o mensuales.
2. **Variables Independientes:** Publicidad, promociones, días festivos, temporada del año, etc.
3. **Explicación:** Un modelo de regresión podría predecir las ventas minoristas en función de variables como la inversión en publicidad, la implementación de promociones, la presencia de días festivos y la estacionalidad.

Valoración de Bienes Raíces:

1. **Variable Dependiente:** Valor de la propiedad.
2. **Variables Independientes:** Tamaño de la propiedad, número de habitaciones, ubicación, características especiales, etc.
3. **Explicación:** Utilizando un modelo de regresión, se podría estimar el valor de una propiedad basándose en su tamaño, número de habitaciones, ubicación y otras características especiales.

Tiempo de Respuesta en Servicios en Línea:

1. **Variable Dependiente:** Tiempo de respuesta del sistema.
2. **Variables Independientes:** Carga del servidor, número de usuarios concurrentes, tipo de solicitud, etc.
3. **Explicación:** Un modelo de regresión podría predecir el tiempo de respuesta de un sistema en línea considerando variables como la carga del servidor, el número de usuarios concurrentes y el tipo de solicitud realizada.

I. Modelo Lineal

Secuencia de regresión Lineal

1. **Conjunto de Datos:**

   Entrenamiento 70-80%

   Prueba 30-20%

2. **Modelo de Base Lineal**

3. **Entrenamiento del Modelo:**

   Ajuste de parámetros **w**

   Máxima Verosimilitud

4. **Métricas de Evaluación**

Modelo de base lineal

- **Regresión lineal.** El modelo más simple de regresión lineal consiste de una combinación lineal de las variables de entrada

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D.$$

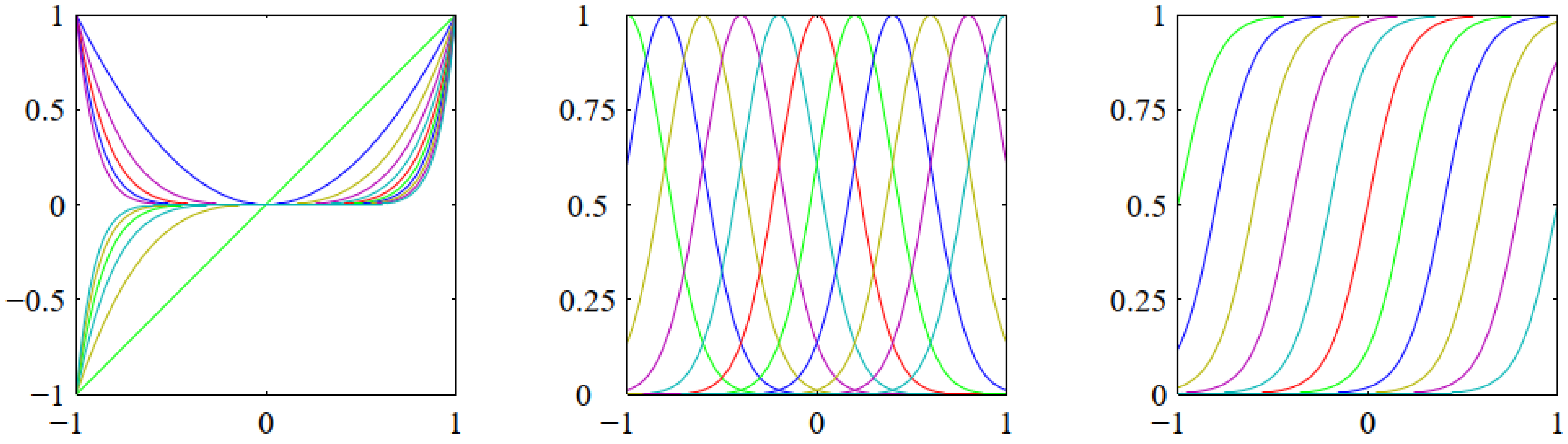
- El modelo anterior se puede extender para combinaciones lineales de funciones no lineales fijas de las variables de entrada,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}),$$

  donde $\phi_i(\mathbf{x})$ son funciones base, M es el número de parámetros del modelo, y w_0 es el desplazamiento.

- Igualmente, $\mathbf{w} = [w_0 \cdots w_{M-1}]^\top$, $\boldsymbol{\phi}(\mathbf{x}) = [\phi_0(\mathbf{x}) \cdots \phi_{M-1}(\mathbf{x})]^\top$.

Ejemplos funciones base



Polinomial: $\phi_i(x) = x^i$.

Exponencial: $\phi_i(x) = \exp\left\{ -\frac{(x-\mu_i)^2}{2s^2} \right\}$

Sigmoidal: $\phi_i(x) = \sigma\left(\frac{x-\mu_i}{s}\right),\ \sigma(a) = 1/(1+\exp(-a))$.

II. Máxima Verosimilitud

Como calcular en valor de **w**

Máxima verosimilitud (I)

- Supongamos t dado como

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

donde $\epsilon \sim \mathcal{N}(0, \beta^{-1})$.

- La incertidumbre en t está dada como

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

- Consideremos un conjunto de datos (de entrenamiento)

$$\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\},$$
$$\mathbf{t} = \{t_1, \cdots, t_N\}$$

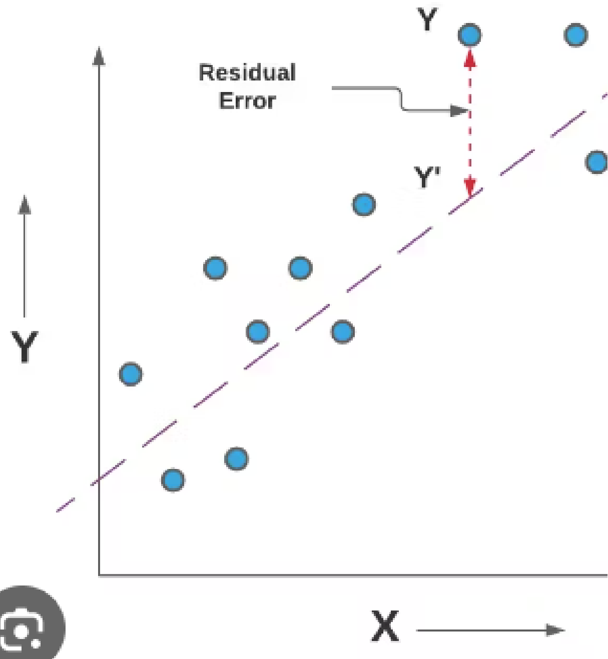
Máxima verosimilitud (II)

Independiente e idénticamente distribuidas

- Suponiendo que los datos son iid

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}).$$

- Tomando el logaritmo de la verosimilitud se tiene



$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}),$$

Termino de normalización	Termino de precisión	Termino de Error cuadrático

donde

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - y(\mathbf{x}_n)\}^2.$$

- Cuáles son los **w**, y el parámetro β que mejor explican los datos.

Máxima verosimilitud (III)

- Maximizar la verosimilitud es equivalente a minimizar $-\beta E_D(\mathbf{w})$.
- De nuevo,

$$
\begin{aligned}
E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 \\
&= \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^\top(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}),
\end{aligned}
$$

donde

$$
\boldsymbol{\Phi} = \begin{bmatrix} \phi(\mathbf{x}_1)^\top \\ \phi(\mathbf{x}_2)^\top \\ \vdots \\ \phi(\mathbf{x}_N)^\top \end{bmatrix} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & & & \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}
$$

Máxima verosimilitud (IV)

- La verosimilitud logarítmica está dada entonces como

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^\top(\mathbf{t} - \mathbf{\Phi}\mathbf{w}),$$

- Se tiene entonces,

$$\begin{aligned}\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)}{\partial \mathbf{w}} &= -\frac{\beta}{2}\frac{\partial}{\partial \mathbf{w}}\left[(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^\top(\mathbf{t} - \mathbf{\Phi}\mathbf{w})\right] \\ &= -\frac{\beta}{2}\frac{\partial}{\partial \mathbf{w}}\left[\mathbf{t}^\top\mathbf{t} - 2\mathbf{t}^\top\mathbf{\Phi}\mathbf{w} + \mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{w}\right]\end{aligned}$$

- Recordemos las siguientes derivadas

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^\top\mathbf{x}) = \mathbf{a}, \qquad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$

Máxima verosimilitud (V)

- Esto significa que

$$\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)}{\partial \mathbf{w}} = \beta \left[\boldsymbol{\Phi}^\top \mathbf{t} - \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{w} \right].$$

- La solución de máxima verosimilitud para **w** está dada como

$$\boxed{\mathbf{w}_{ML} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t},}$$

Llegamos a una expresión para calcular nuestros **w**
Dada la pseudoinversa
$A^+ = (A^T A)^{-1} A^T$

donde $\boldsymbol{\Phi}^\dagger \equiv (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top$ es la pseudo-inversa Moore-Penrose.

- La solución de máxima verosimilitud para β se obtiene de

$$\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2} (\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^\top (\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}).$$

- Y así,

$$\frac{1}{\beta_{ML}} = \frac{1}{N} (\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{ML})^\top (\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \{t_n - \mathbf{w}_{ML}^\top \phi(\mathbf{x}_n)\}^2.$$

III. Métricas de evaluación

El objetivo es minimizar esta función para mejorar la precisión del modelo

Métricas de evaluación regresión Lineal

MSE (Error Cuadrático Medio)
MAE (Error Absoluto Medio)
RMSE
R2 (El coeficiente de determinación)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

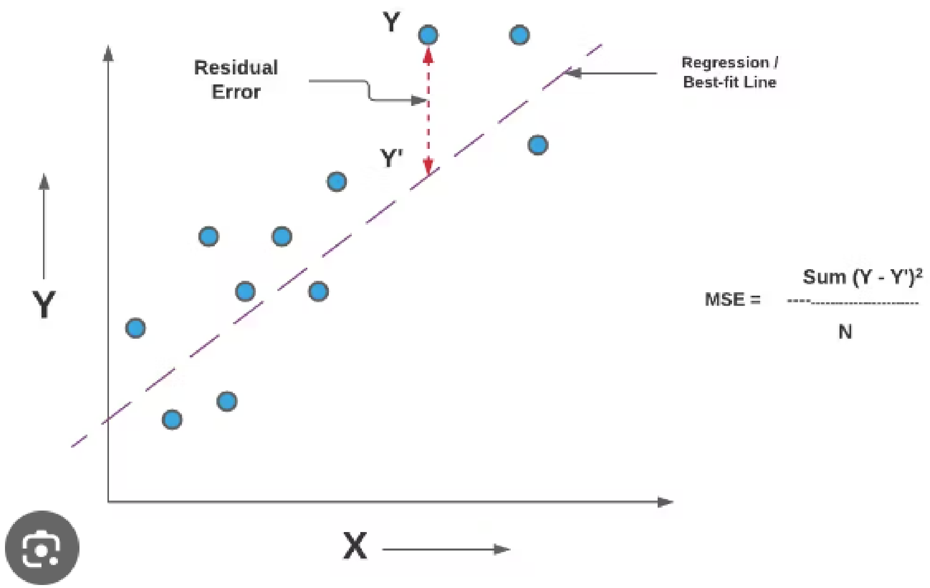
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

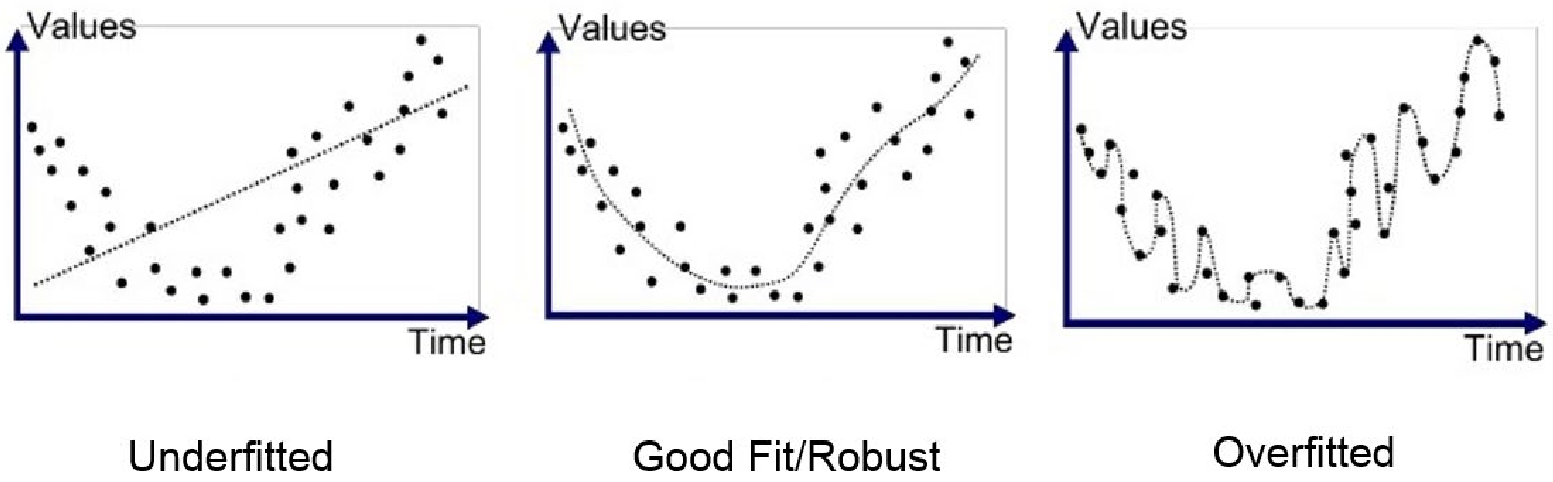
$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE}$$

Interpretación

•MSE penaliza más los errores grandes, mientras que MAE trata todos los errores por igual.
•RMSE: medida de la magnitud promedio de los errores en la misma unidad que la variable objetivo
•R2: medida de la proporción de la variabilidad en la variable dependiente(cercano a 1 indica que el modelo explica una gran proporción de la variabilidad)





IV. Regularización

La regularización se utiliza para evitar el sobreajuste del modelo y mejorar su generalización

Definición

- Controlar el sobre entrenamiento.
- La función de error toma la forma

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w},$$

donde $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w}$.

- El valor de $\mathbf{w}$ que minimiza $E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$ está dado por

$$\mathbf{w} = (\lambda \mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{t}.$$

Alternativas de regularización

- En general, la función de error toma la forma

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q.$$

- El caso $q = 2$ es el regularizador cuadrático anterior.
- El caso $q = 1$ se conoce como la regresión **lasso**.