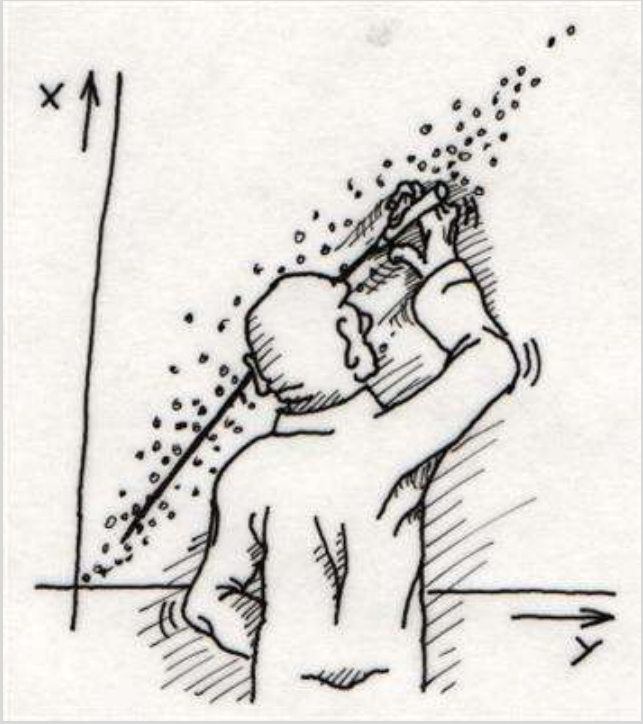


DigitalHouse > Coding School

DATA SCIENCE

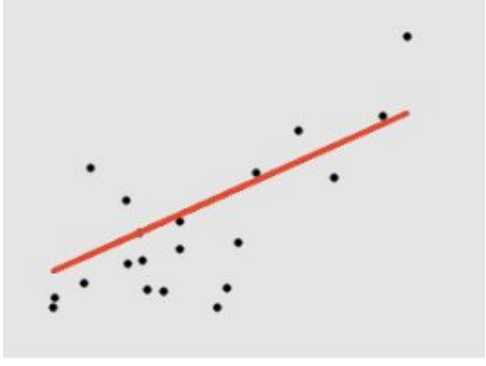
MÓDULO 3

Normalización

OBJETIVOS DE LA CLASE

1. **Entender el concepto y utilidad de la normalización de datos**

2. **Usar el módulo de preprocesamiento de scikit-learn para normalizar datos**



REPASO

La **regresión lineal simple** intenta predecir una respuesta cuantitativa **Y** en base a una única variable predictora **X.** Asume que hay aproximadamente una relación lineal entre X e Y.

$$Y \approx \beta_0 + \beta_1 X.$$

β_0 y β_1 son dos constantes que representan el intercepto y la pendiente en el modelo lineal. Juntos, β_0 y β_1 son conocidos como los **parámetros** del modelo.
Una vez que hemos usado nuestro set de entrenamiento para producir los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ para los coeficientes del modelo, podemos predecir futuras valores de la variable **Y.**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

donde $\hat{y}$ indica una **predicción de Y en base a X.** Aquí usamos un símbolo ^ para denotar el valor estimado para un parámetro o coeficiente desconocido, o para denotar el valor predicho de la respuesta.

En lugar de ajustar un modelo distinto de regresión simple para cada predictor, una mejor aproximación es extender el modelo de regresión simple para que puede incluir **múltiples predictores**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Dados estimadores de los coeficientes de pendiente podemos pronosticar la variable de respuesta para una observación con valores dados de los predictores como:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

Elegimos los valores para los estimadores de los coeficientes que minimizan la suma de residuos al cuadrado:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

Bajo ciertas condiciones, conocidas como los **supuestos de Gauss - Markov**, los coeficientes de la regresión son lineales, insesgados y tienen varianza mínima.

1. El modelo es **lineal en los parámetros**.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

2. Los estimadores de los parámetros poblacionales se estiman a partir de una **muestra aleatoria**.

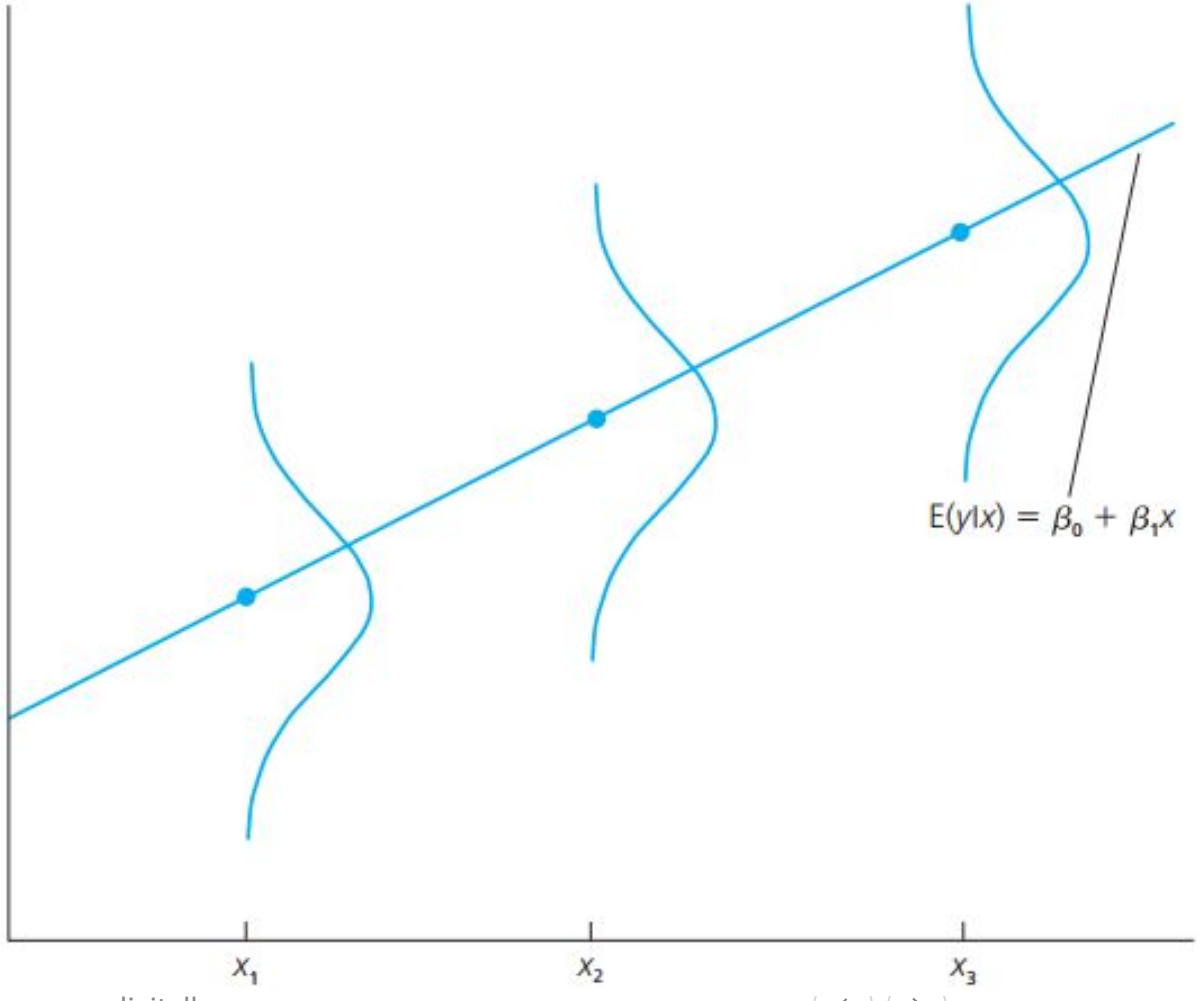
3. **No hay colinealidad perfecta** entre las variables explicativas.

4. El valor **esperado del error es 0** para cualquier valor de la variable explicativa.
5. Para cualquier valor de la variable explicativa, el error tienen la misma varianza (**homocedasticidad**).
6. El error es independiente de las variables explicativas y se distribuye normalmente.

$$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$$

7. No existe autocorrelación entre los errores de dos observaciones diferentes condicionadas a X.

$$Cov(\varepsilon_i, \varepsilon_h | X) = 0$$



Homocedasticidad + Media Condicional igual a 0.

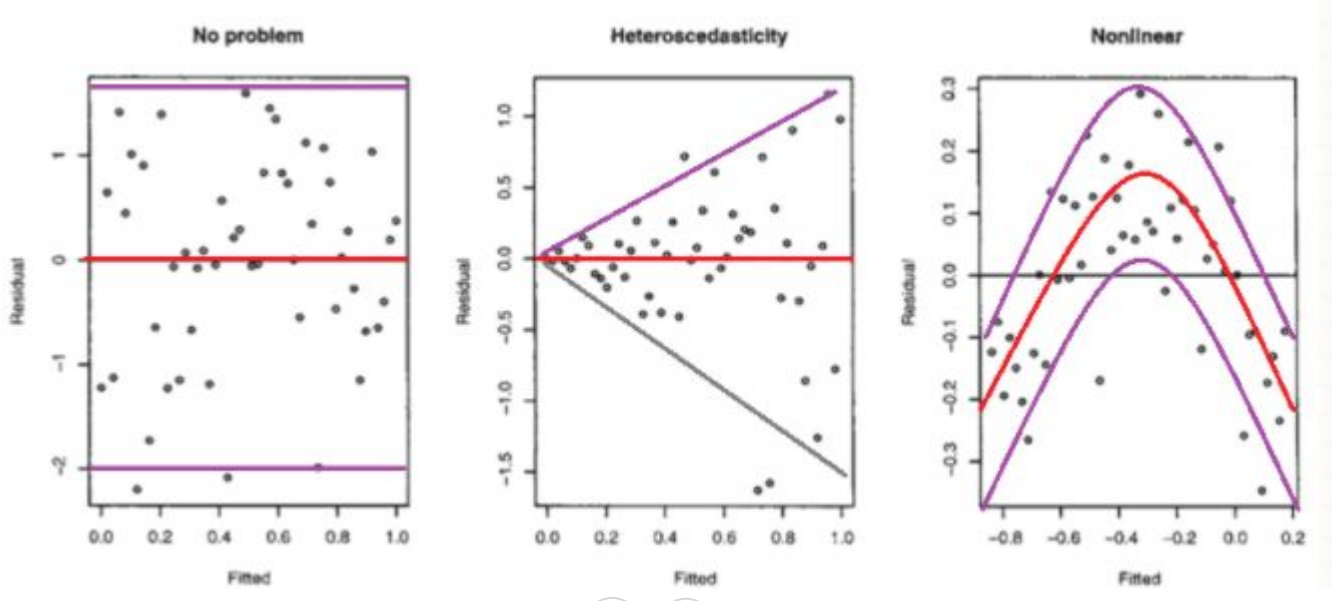
La consecuencia de estos dos supuestos, es que para cada valor de X, los errores se distribuyen con media cero y desvío estándar constante (σ)

$$E(y|x) = \beta_0 + \beta_1 x.$$

$$Var(y|x) = \sigma^2.$$

Una buena forma de evaluar el cumplimiento de los supuestos sobre los residuos es graficar los mismos contra los valores predichos.

- En el primer caso los residuos tienen la misma media y varianza para todos los valores de ŷ.
- En el segundo caso, la varianza de los residuos aumenta con ŷ
- En el último caso la varianza parece constante, pero los residuos tienden a ser negativos para valores muy bajos o muy altos de ŷ y positivos para valores intermedios.



¿Existe evidencias para afirmar que hay relación entre X e Y?

Los errores estándar de los estimadores de los coeficientes también pueden ser usados para realizar tests de hipótesis.

El test de significación individual tiene las siguientes hipótesis:

H_0: No hay relación entre X e Y

$$H_0 : \beta_1 = 0$$

versus la **hipótesis alternativa:**

H_a: Hay alguna relación entre X e Y

$$H_a : \beta_1 \neq 0,$$

- Recordemos la función que se minimiza en la estimación de mínimos cuadrados:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- La función que se minimiza en Regresión Ridge es:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- La función que se minimiza en LASSO es:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

NORMALIZACIÓN

¿Por qué normalizar?

- Manejo de cantidades en **diferentes unidades o escalas**
- Muchos **algoritmos** de machine learning toman la normalización como **requerimiento**
- Existen distintas razones por las cuales un algoritmo de ML requiere estandarizar los datos.

Muchos algoritmos de ML se basan en el cálculo de medidas de distancia que se calculan entre todos los puntos en base a distintos features.

La medida de distancia que viene implementada por default, es la distancia euclídea que requiere matemáticamente que todos los features sean numéricos.

Además, para no favorecer a ningún feature en particular a la hora de explicar la distancia, tenemos que estandarizar y deshacernos de las unidades.

¿Cómo normalizar?

Existen algunas formas típicas de normalizar:

- La ***estandarización***: $x_{norm} = (x - \mu) / \sigma$
- La normalización ***min-max:*** $x_{norm} = (x - min) / (max - min)$

La elección entre min-max y estandarización depende del objetivo del método:

- **Min-max**: Tiene sentido en los casos donde importa que los features tengan las mismas unidades pero no necesariamente la misma varianza

- **Estandarización:** Tiene sentido donde se necesita que los features tengan las mismas unidades y también la misma varianza, como por ejemplo en componentes principales.

Práctica Guiada

Normalización

Conclusión

Usamos la normalización para:

- Manejo de cantidades en **diferentes unidades o escalas**
- Muchos **algoritmos** de machine learning toman la normalización como **requerimiento**
- Puede **aumentar la velocidad de convergencia** usando el método de gradiente

Existen diferentes métodos de normalización, como la estandarización, min-max y L1 y L2.