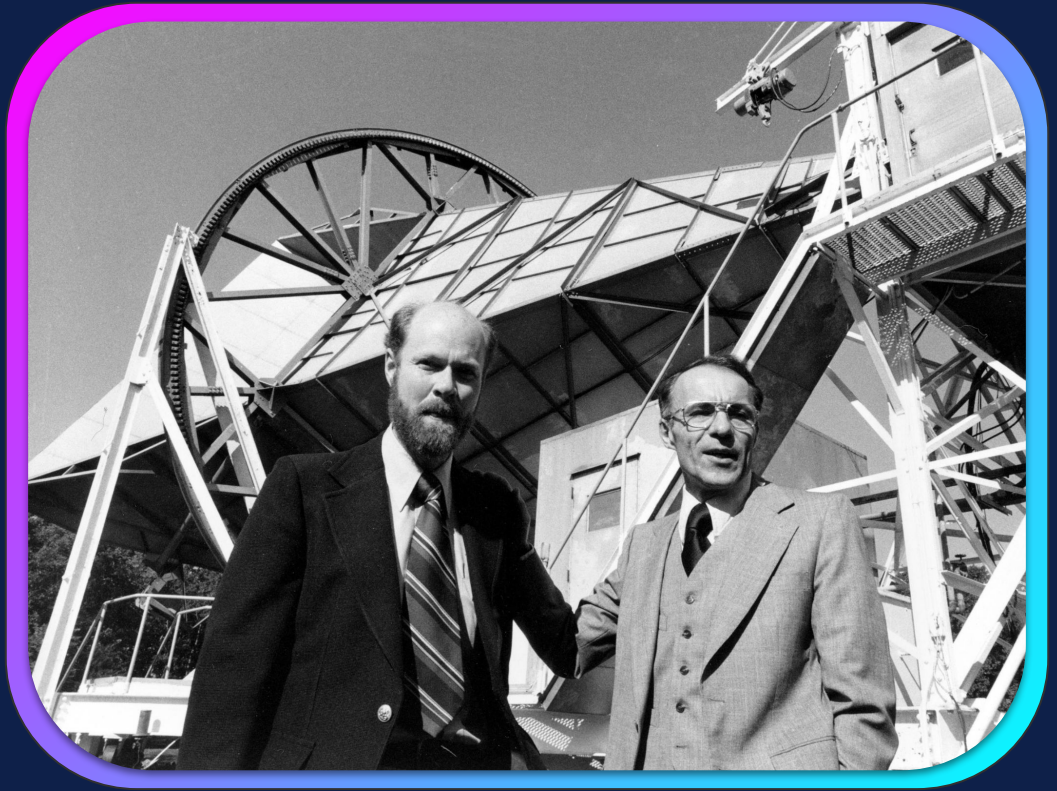
ASTRONOMÍA Y CIENCIA DE DATOS: DE LAS ESTRELLAS A LOS NÚMEROS

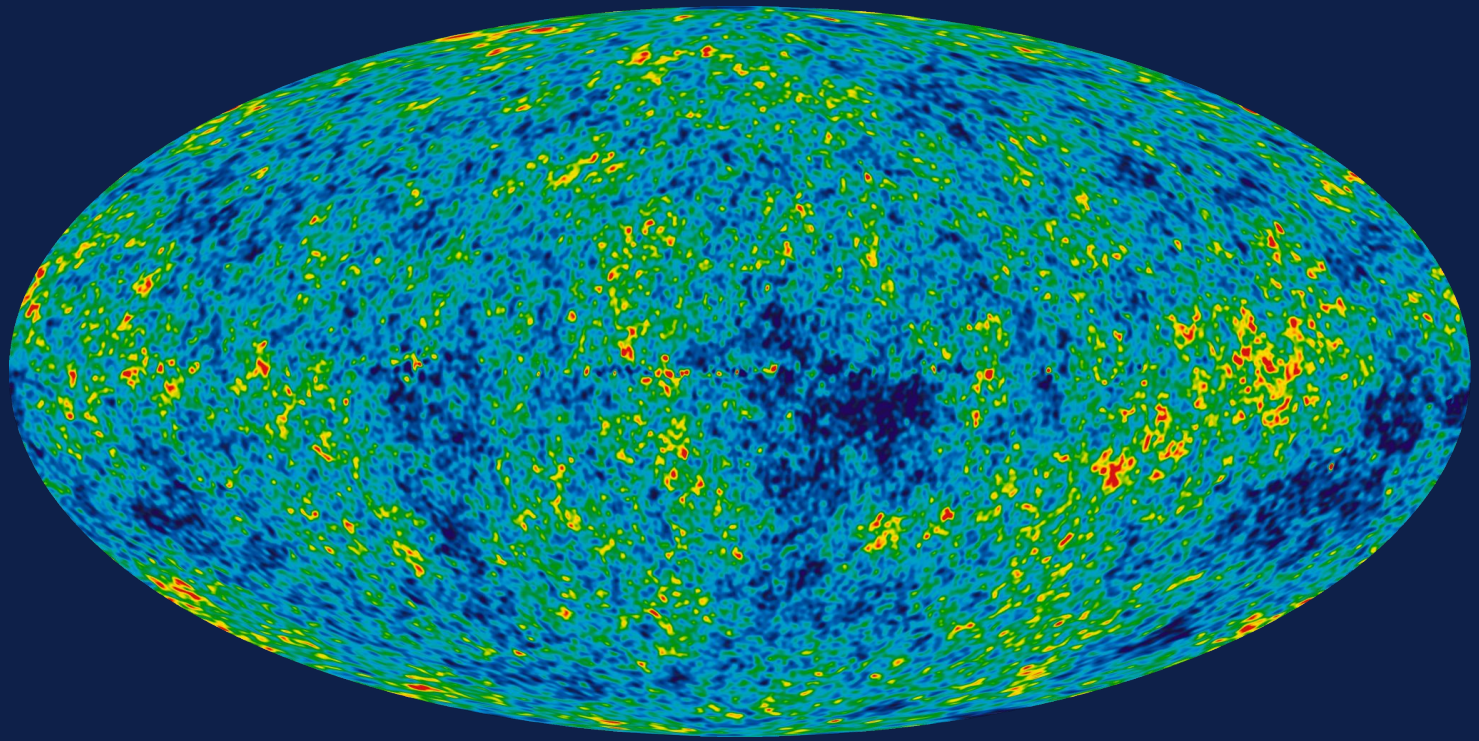
Clase 3: Herramientas Fundamentales III:
Ajuste de Modelos

Primero, una historia...

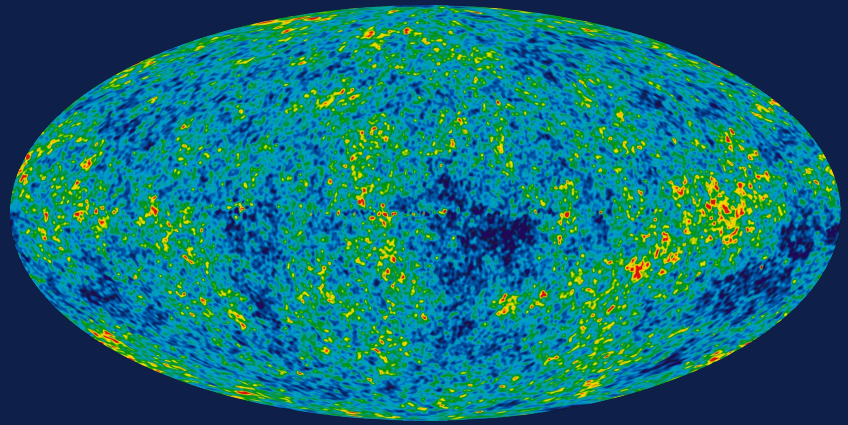
En 1967, Arno Penzias y Robert Wilson descubrieron sin querer, la radiación “más antigua” del universo...



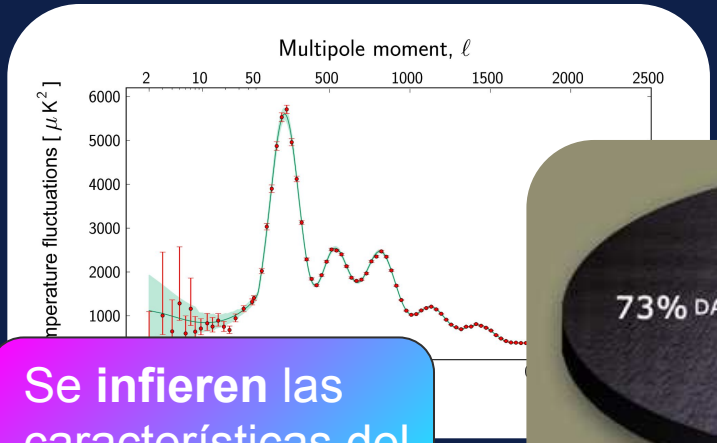
Así se ve con telescopios espaciales!



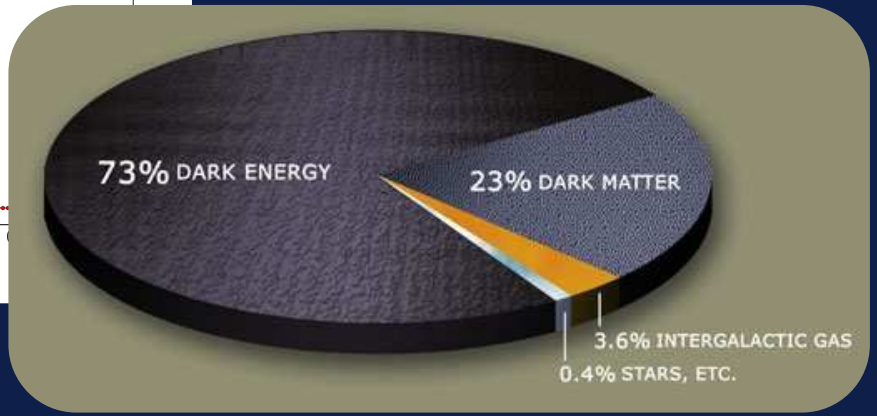
Los datos…



Se proponen **modelos**



Se **infieren** las características del universo



¿Qué es un modelo científico?

Es un marco teórico que explica **cómo** funciona la naturaleza.

Es de carácter **predictivo.**

Ejemplos:

$$\vec{F} = \frac{d\vec{p}}{dt} = m\vec{a}$$

$$i\hbar\frac{\partial}{\partial t}\left|\Psi\right\rangle = \hat{H}\left|\Psi\right\rangle$$

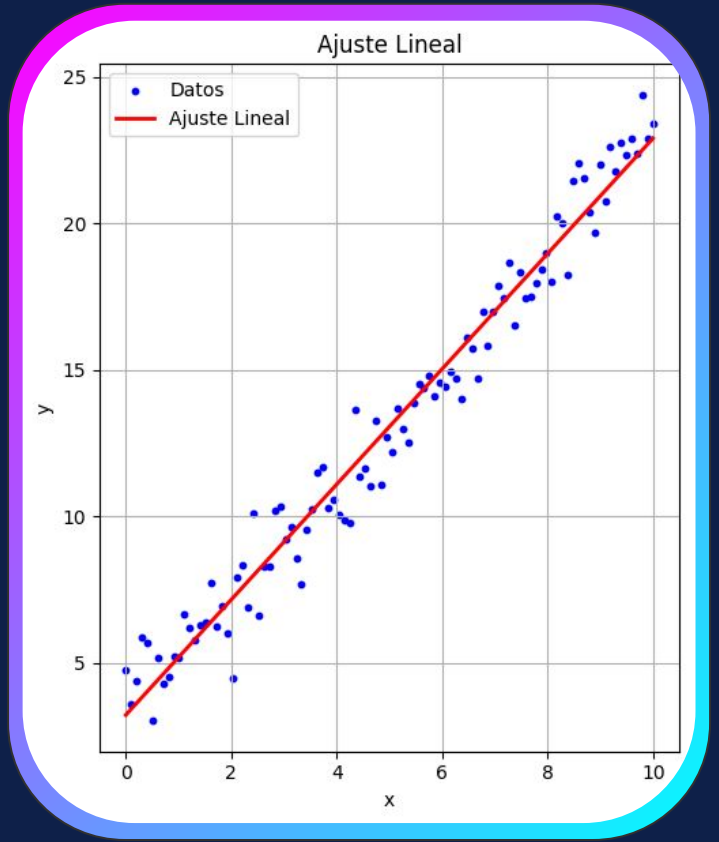
Ajuste de modelos: el problema

Supongamos que tenemos un set de datos, el modelo será alguna **función matemática que aproxime la tendencia de los datos.**

$$(x_i, y_i) \longrightarrow f(x_i)$$

$$y_i = f(x_i) + e(x_i)$$

$$e_i(x_i) \sim N(0, \sigma_i)$$



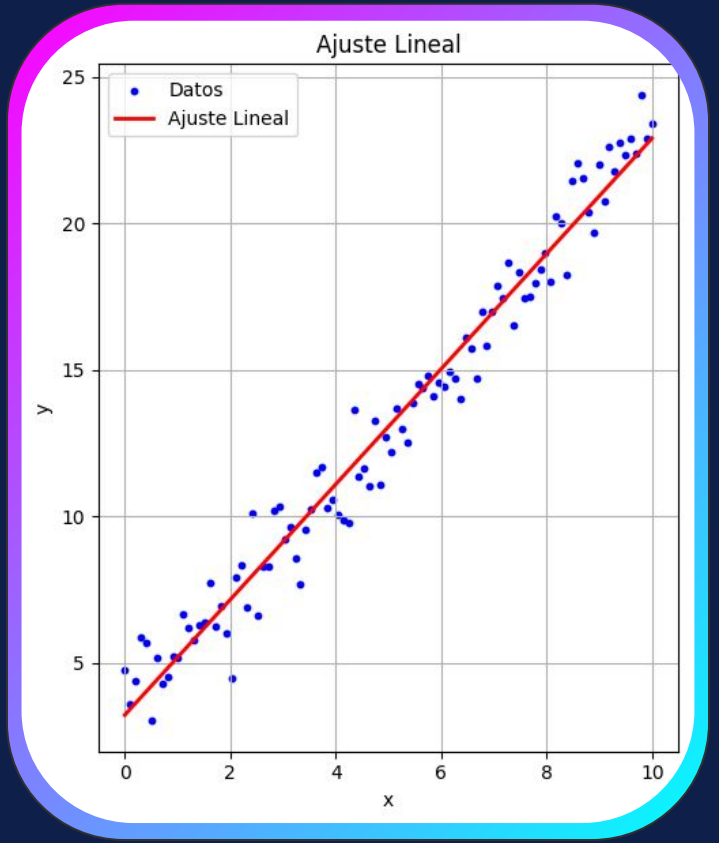
Ajuste de modelos: el problema

En general, los modelos estarán **parametrizados**, entonces la pregunta es: ¿cuál es el mejor set de parámetros que describe los datos?

$$f(x_i) \longrightarrow f(x_i, m, n) = m \cdot x_i + n$$

$$\min Error = \sum_{i=1}^{N} (y_i - f(x_i, n, m))^2$$

Notar que el Error es una función únicamente de *(n,m)* en este caso.



Encuentre la mejor línea recta

$$\min \sum_{i=1}^{N} (y_i - (mx_i))^2$$

$$m = \frac{\sum_i^N y_i x_i}{\sum_{i=1}^N x_i^2}$$

Ajuste de modelos: lineal y no-lineal

Definición

Diremos que estamos frente a un problema de ajuste de parámetros lineal **(regresión lineal)** siempre y cuando $f(x_i; \beta_j)$, sea lineal en los parámetros β_j. **En caso contrario el problema es no-lineal.**

Ejemplos: ¿Lineales o no-lineales?

$$f(x_i, \beta_1, \beta_2, \beta_3) = \beta_1 + \beta_2 \sin(x_i) + \beta_3 \sin(2x_i)$$ lineal

$$f(x_i, \beta_1, \beta_2) = \beta_1 e^{x_i/\beta_2}$$ no-lineal

Regresión lineal

Vamos a generalizar el problema, atención, se viene álgebra lineal!

Definición: problema de ajuste de parámetros lineal

$$f(x_i, \beta_1, \beta_2, \ldots, \beta_M) = f(x_i; \vec{\beta}) = \sum_{j=1}^{M} \beta_j g_j(x_i)$$

Donde $g_j(x)$ es alguna función real.
El problema a resolver es:

$$\min \sum_{i=1}^{N} (y_i - f(x_i; \vec{\beta}))^2 \quad \text{, con} \quad y_i = f(x_i; \vec{\beta}) + e_i$$

Regresión lineal

Vamos a generalizar el problema, atención, se viene álgebra lineal!

Definición: matriz de diseño

$$y_i = f(x_i; \vec{\beta}) + e_i = \sum_{j=1}^{M} \beta_j g_j(x_i) + e_i$$

De forma matricial como:

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} g_1(x_1) & g_2(x_1) & g_3(x_1) & \dots & g_M(x_1) \\ g_1(x_2) & g_2(x_2) & g_3(x_2) & \dots & g_M(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1(x_N) & g_2(x_N) & g_3(x_N) & \dots & g_M(x_N) \end{bmatrix}_{N\times M} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}_{M\times 1} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_{N\times 1}$$

Regresión lineal

Vamos a generalizar el problema, atención, se viene álgebra lineal!

Definición: matriz de diseño

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} g_1(x_1) & g_2(x_1) & g_3(x_1) & \dots & g_M(x_1) \\ g_1(x_2) & g_2(x_2) & g_3(x_2) & \dots & g_M(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1(x_N) & g_2(x_N) & g_3(x_N) & \dots & g_M(x_N) \end{bmatrix}_{N\times M} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}_{M\times 1} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_{N\times 1}$$

X se conoce como **matriz de diseño**, así de forma compacta escribimos:

$$\vec{y} = X\vec{\beta} + \vec{e}$$

Regresión lineal

Bajo todas estas definiciones el problema se puede resumir como encontrar un $\vec{\beta}$ tal que:

$$\min \sum_{i=1}^{N} (y_i - f(x_i; \vec{\beta}))^2 = \min \| \vec{e} \|^2 = \min \| \vec{y} - X\vec{\beta} \|^2$$

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

Otras métricas

¿Existen otras formas de definir el “error” a minimizar? Sí!

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

“Importan menos” aquellos datos con mayor error asociado.

Otras métricas

¿Existen otras formas de definir el “error” a minimizar? Sí!

$$\chi^2_{red} = \frac{1}{N - M} \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

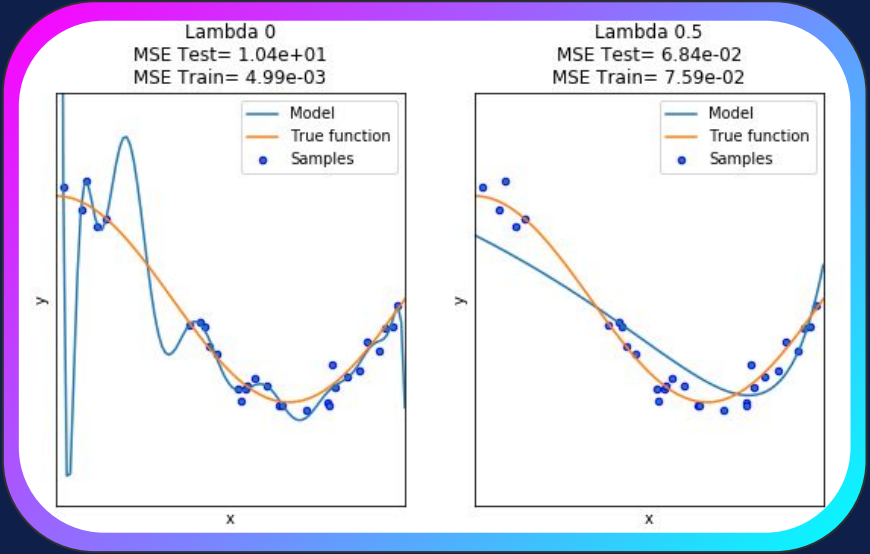
M es el número de parámetros del fit.

“Más parámetros no es necesariamente mejor”

Nota: **esta métrica sólo tiene sentido para modelos lineales!**

Ajuste lineal: MC + Regularización

¿Qué tan complejo debe ser el modelo?



En ajustes de modelos, este problema se conoce como **Overfitting** (Sobreajuste).

A veces menos es más...

Ajuste lineal: MC + Regularización

Podemos agregar una penalización a los parámetros en mínimos cuadrados:

$$J = \|\vec{y} - X\vec{\beta}\|^2 \longrightarrow J = \|\vec{y} - X\vec{\beta}\|^2 + \rho\|\vec{\beta}\|_p^p$$

El parámetro ρ nos permite "pesar" la norma de los parámetros en el ajuste.

$$\begin{aligned} \nabla_{\vec{\beta}} J &= 0 \\ -2(\vec{y} - X\vec{\beta})^\top X + 2\rho\vec{\beta}^\top &= 0 \\ -\vec{y}^\top X + \vec{\beta}^\top X^\top X + \rho\vec{\beta}^\top &= 0 \end{aligned}$$

Con un poco de álgebra lineal podemos obtener la solución con regularización

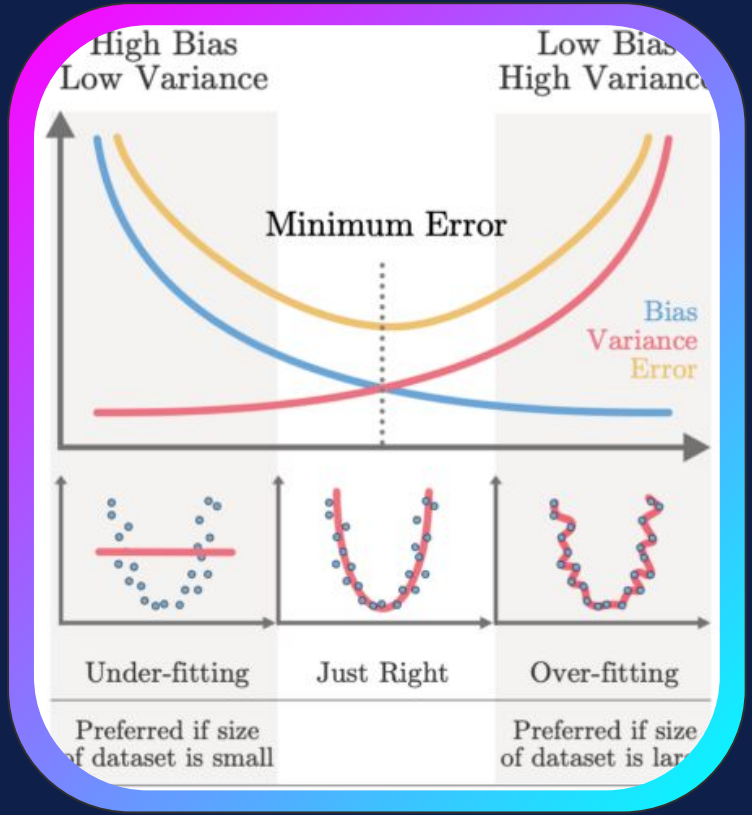
$$\vec{\beta}^\top = \vec{y}^\top X(X^\top X + \rho\delta_{ij})^{-1} \longrightarrow \vec{\beta} = (X^\top X + \rho\delta_{ij})^{-1} X^\top \vec{y}$$

Ajuste lineal: MC + Regularización

Notamos que $||||_p^p$ es la norma p. Las regularizaciones más comunes son las de **Lasso** (p=1) y **Ridge** (p=2).

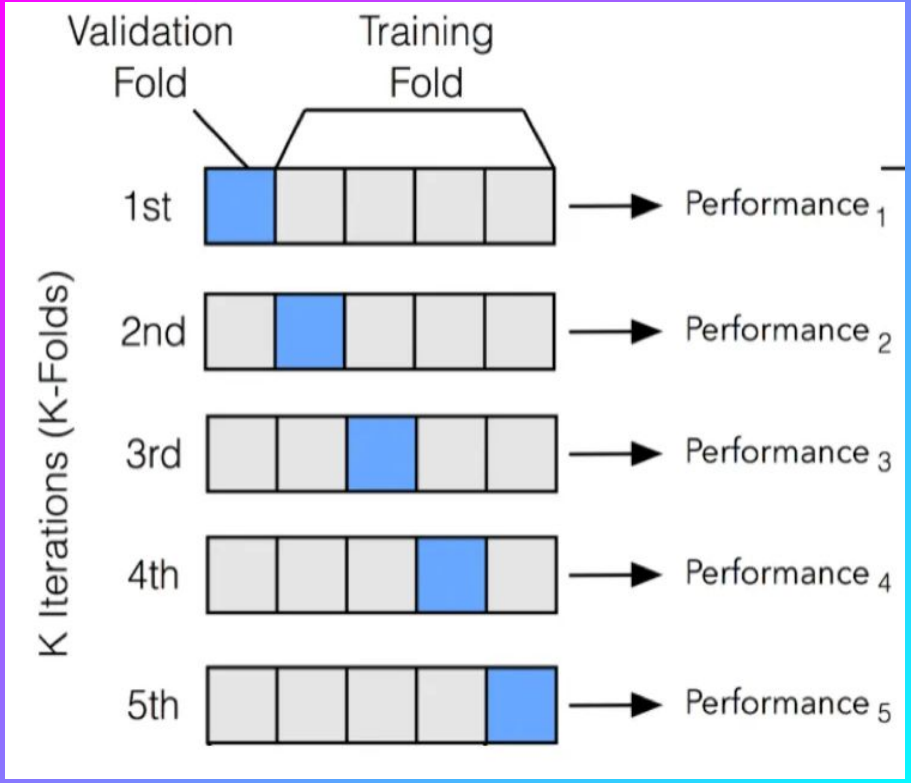
Cabe la pregunta: ¿Hay un valor óptimo de ρ ?

Lo típico es ir probando valores, aunque existen estrategias para evaluar y elegir este parámetro.



Descomposición Sesgo-Varianza

Cross-Validation



¿Cual es el problema de entrenar con el set completo de datos?

Esta estrategia nos permite evaluar el desempeño de un modelo en cuanto a su capacidad predictiva. Consiste en dividir el set de datos en entrenamiento y testeo.

- Validación cruzada exhaustiva
- Validación cruzada no exhaustiva

Cross-Validation

Validación Cruzada exhaustiva

- Leave p out (LpOCV)
- Leave One out (LOOCV)

Validación Cruzada no exhaustiva

- k-fold
- Monte Carlo CV

Nota: **Usualmente se utiliza un 75% para entrenamiento y validación, y un 25% para testeo.**

Ajuste no lineal: Método del gradiente

Los métodos de gradiente son útiles para encontrar mínimos en funciones. Se caracterizan por ser eficientes.

Lo lógico es avanzar en la dirección contraria del gradiente:

$$f(\vec{x}) \longrightarrow f(\vec{x}) - \lambda \nabla f(\vec{x})$$

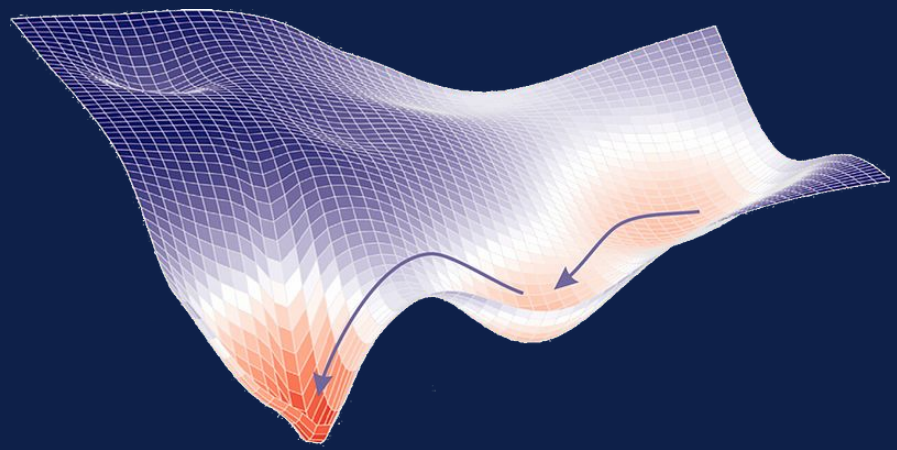
El parámetro λ se conoce como learning rate. En los algoritmos actuales es adaptativo.

Ajuste no lineal: Método del gradiente

Cuando los sets de datos son muy extensos tenemos el problema de que el gradiente convencional es muy costoso de entrenar (Matrices gigantes!). Para solucionar este problema, existen variaciones del método de gradiente:

- Gradiente descendiente estocástico: En cada iteración se avanza calculando el gradiente para un set de datos.

- Mini-batch: Se divide el set de entrenamiento en paquetes (batch). En cada iteración se realizan avances de gradiente por cada paquete.



Una perspectiva probabilística

En estadística hay 2 enfoques para tratar modelos, el enfoque frecuentista y el bayesiano:

- Enfoque Frecuentista: Ningún modelo es el real, sin embargo se ajusta a los datos mediante distintas estrategias (MC por ejemplo). Se define la probabilidad en términos de la experimentación (frecuencias).

- Enfoque bayesiano: Todos los modelos son probables de haber generado los datos. Se busca el modelo que con mayor probabilidad generó dichos datos, en términos de su distribución de probabilidad. Se utiliza el criterio de máxima verosimilitud (CMV).