Técnicas Selectas de Machine Learning y Aplicaciones

Presentado: Juan Manuel Hurtado Restrepo

Supervisor: José Ángel González Prieto

Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid

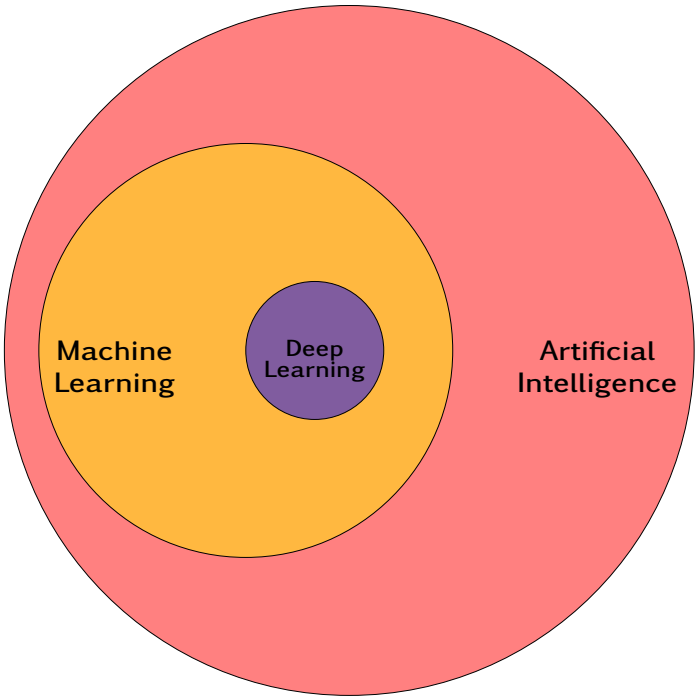
Defensa pública del TFG, 10 de julio de 2025

Índice

Índice

Machine learning e IA

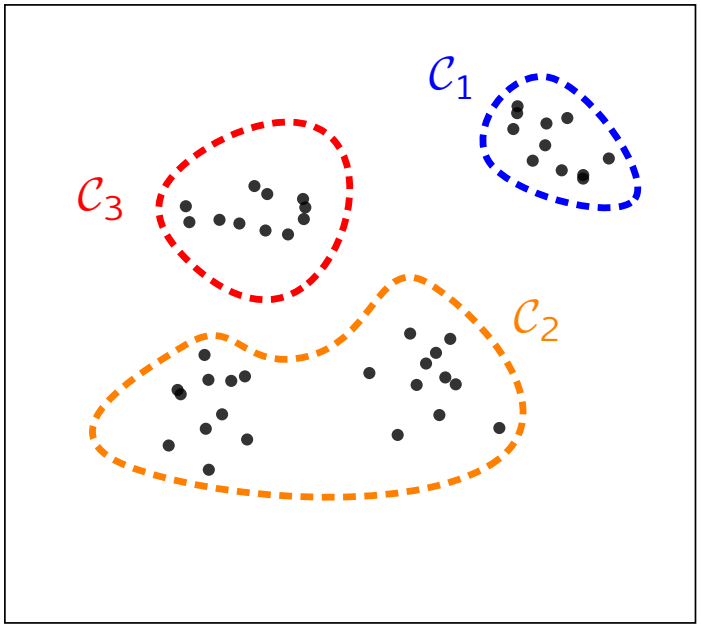
Aprendizaje supervisado y no supervisado



- El aprendizaje supervisado se basa en un conjunto de datos, de los que se conocen los correspondientes resultados esperados.

El objetivo es entonces entrenar un sistema de aprendizaje automático con la información anterior, para predecir los resultados cuando se presenten nuevas entradas.

- El enfoque de aprendizaje no supervisado:



Problema de regresión y problema de clasifición

Problema de regresión

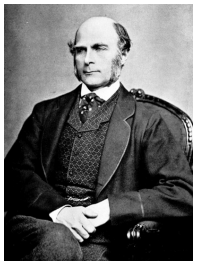


Figura: Francis Galton.

Del modelo de Galton originado tras el análisis de 205 familias:

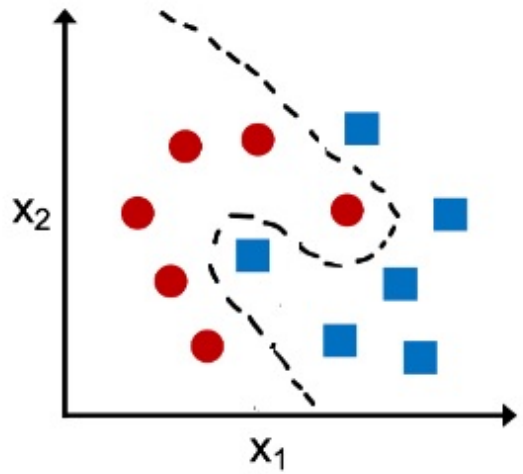
$$E[Y \mid X] = \frac{2}{3}X.$$

A la formulación general:

$$Y = f(X, \boldsymbol{\Omega}) + \varepsilon$$

Problema de clasificación

Una colección de N objetos O_i con $i \in \{1, \ldots, N\}$, cada uno representado por un vector real de dimensión d, $\boldsymbol{x}_i \in \mathbb{R}^d$; un conjunto $\{1, \ldots, M\}$ de clases o categorías; una variable Y etiquetadora que toma valores en dicho conjunto; y una función clasificadora f que mapea vectores del espacio $\mathbb{R}^d$ al conjunto de etiquetas de clase $\{1, \ldots, M\}$.



Métodos de estimación de parámetros

Mínimos cuadrados

Figura: Carl F. Gauss.

Máxima verosimilitud

Figura: Ronald A. Fisher

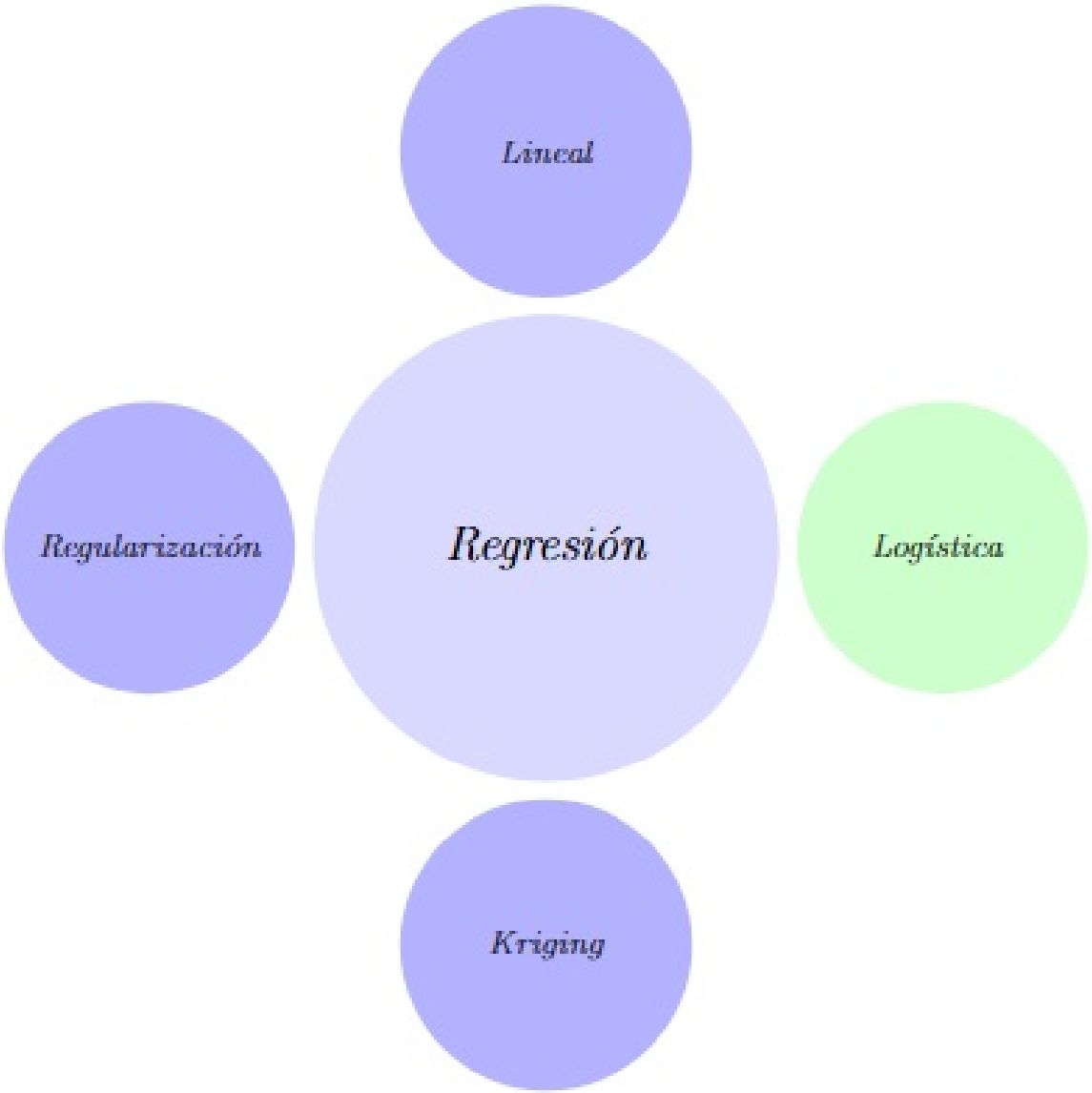
Para $f(x_k, \boldsymbol{\Omega}) = \sum_{j=1}^{m} \omega_j \, \phi_j(x_k) \; \forall \, k = 1, \ldots, n$ donde $\boldsymbol{\Omega}^T = (\omega_1, \ldots, \omega_m)$, tenemos las *ecuaciones normales de Gauss* $\forall \, l = 1, \ldots, m$:

$$\sum_{k=1}^{n} \left(\sum_{j=1}^{m} \omega_j \, \phi_j(x_k) \right) \phi_l(x_k) = \sum_{k=1}^{n} y_k \, \phi_l(x_k)$$

De la realización de una muestra aleatoria $X_1, \ldots, X_n$, proveniente de una población X con distribución de probabilidad $f_{\boldsymbol{\omega}}$ que depende de un parámetro desconocido $\boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^m$, la función protagonista es la *verosimilitud* $L(\boldsymbol{\omega}) = f_{\boldsymbol{\omega}}(x_1, \ldots, x_n)$.

Índice

Técnicas de regresión



Regresión lineal

Modelo de regresión lineal

Siendo ε la discrepancia, la variable respuesta es:

$$Y = \boldsymbol{\Phi}^T(\boldsymbol{X}) \cdot \boldsymbol{\Omega} + \varepsilon$$

donde $\boldsymbol{X}^T = (X_1, \ldots, X_p), \boldsymbol{\Phi}^T = (\phi_0, \ldots, \phi_n)$ con $\phi_0(\boldsymbol{X}) = 1$ y $\boldsymbol{\Omega}^T = (\omega_0, \ldots, \omega_n)$.

Estimación de parámetros

Considerando una muestra de entrenamiento $D = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^N \subseteq \mathbb{R}^p \times \mathbb{R}$ y siendo $\boldsymbol{y}^T = (y_1, \ldots, y_N)$, tenemos lo siguiente:

$$\hat{\boldsymbol{\Omega}} = \underbrace{\left(\boldsymbol{\Psi}^T \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Psi}^T}_{\boldsymbol{\Psi}^\dagger} \boldsymbol{y}, \quad \textit{donde} \quad \boldsymbol{\Psi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \cdots & \phi_n(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \cdots & \phi_n(\boldsymbol{x}_N) \end{pmatrix}$$

Regularización

Elastic net

Para $0 \leq \alpha \leq 1$ y $\boldsymbol{\Omega}^T = (\omega_0, \ldots, \omega_m)$, formalmente el problema es:

$$\hat{\boldsymbol{\Omega}}^{e.n.} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{n-1} \left(y_k - \omega_0 - \sum_{l=1}^{m} x_{kl}\, \omega_l \right)^2 + \lambda \sum_{l=1}^{m} \left(\alpha |\omega_l|^2 + (1-\alpha)|\omega_l| \right) \right\}$$

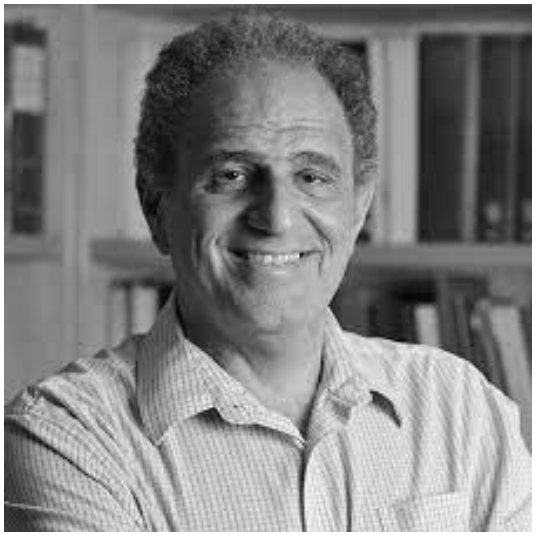


Figura: Robert Tibshirani.

Casos especiales

- $\alpha = 0 \longrightarrow$ *LASSO*.
- $\alpha = 1 \longrightarrow$ *ridge*.

Generalización

Para $p \in \mathbb{N} \cup \{\infty\}$, la técnica se puede generalizar considerando:
$\|\cdot\|_p^p : \mathbb{R}^m \to \mathbb{R}, \boldsymbol{z} \mapsto \|\boldsymbol{z}\|_p^p = \sum_{l=1}^{m} |z_l|^p$.

Kriging

Funciones kernel

Para una aplicación $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{V}, \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})$, con $\mathcal{X}$ el espacio de entrada y $\mathcal{V}$ un espacio con producto interior, una función kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ se define como:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle_{\mathcal{V}} = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}') = \sum_{i=1}^{m} \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{x}')$$

donde $\boldsymbol{\phi}^T = (\phi_1, \ldots, \phi_m)$.

Estimación en regresión de procesos gaussianos

Dado el cojunto de entrenamiento $D = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N} \subseteq \mathbb{R}^p \times \mathbb{R}$, el kernel gaussiano $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\theta \|\boldsymbol{x} - \boldsymbol{x}'\|)$ y una nueva obervación $\boldsymbol{x}_{N+1}$, los parámetros de y_{N+1} son:

$$m(\boldsymbol{x}_{N+1}) = \boldsymbol{k}^T C_N^{-1} \boldsymbol{y}; \quad \sigma^2(\boldsymbol{x}_{N+1}) = c - \boldsymbol{k}^T C_N^{-1} \boldsymbol{k}$$

Regresión logística

Estimación de parámetros en la versión binaria

Teniendo presente que $\boldsymbol{\Omega}^T = (\omega_0, \boldsymbol{\omega}^T)$ y $p(\boldsymbol{x};\boldsymbol{\Omega}) = \dfrac{1}{\exp\{-(\omega_0 + \boldsymbol{\omega}^T\boldsymbol{x})\}+1}$,
si partimos de un conjunto de entrenamiento
$D = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^N \subseteq \mathbb{R}^p \times \mathbb{R}$, tenemos que:

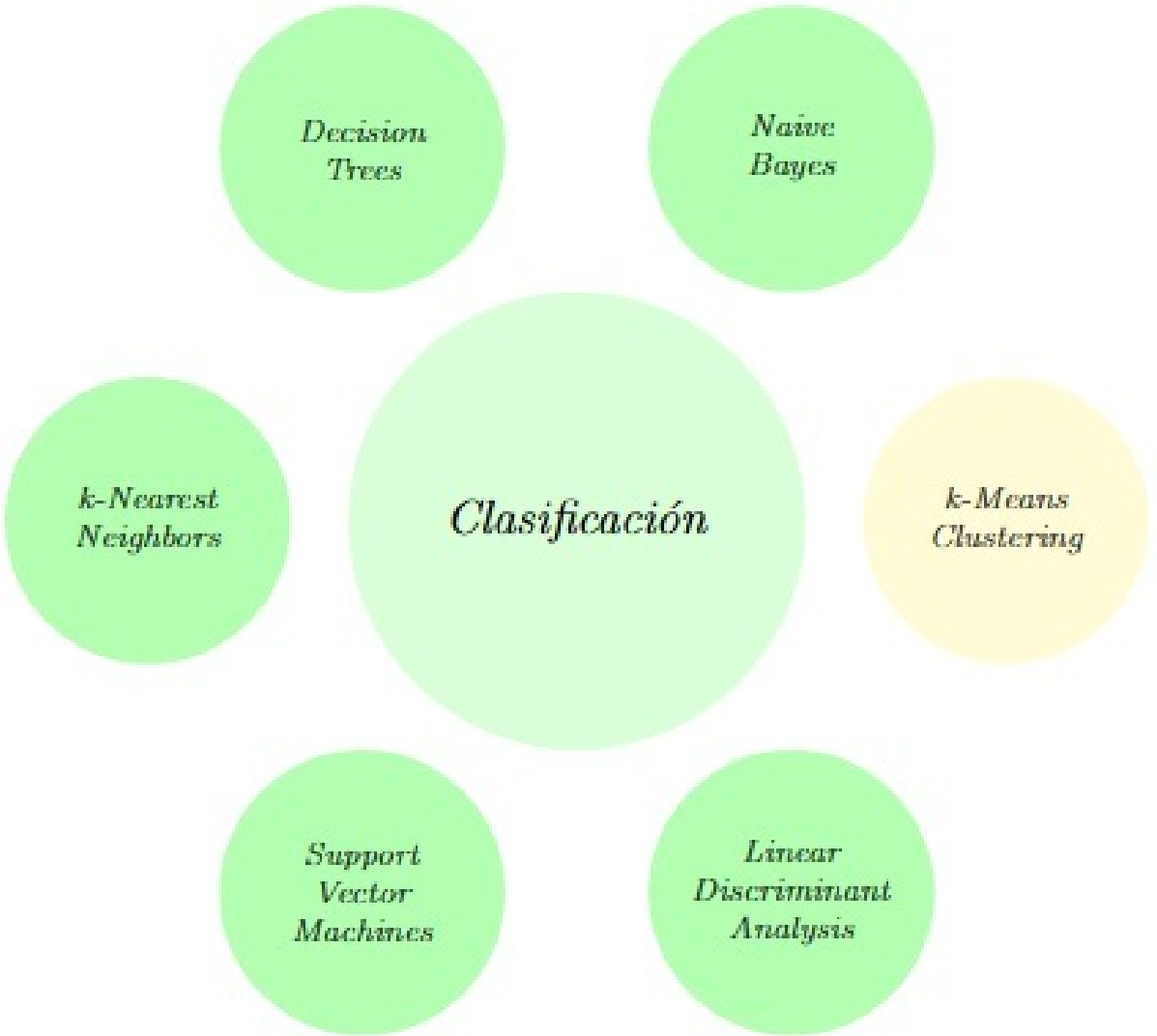
$$\begin{aligned} \boldsymbol{\Omega}^{new} &= (\mathbf{X}^T\boldsymbol{W}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{W}\boldsymbol{z} \\ donde \quad & \boldsymbol{z} = \mathbf{X}\boldsymbol{\Omega}^{old} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p}), \end{aligned}$$

$$\mathbf{W} = \begin{pmatrix} p(\boldsymbol{x}_1;\boldsymbol{\Omega}^{old})(1-p(\boldsymbol{x}_1;\boldsymbol{\Omega}^{old})) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p(\boldsymbol{x}_N;\boldsymbol{\Omega}^{old})(1-p(\boldsymbol{x}_N;\boldsymbol{\Omega}^{old})) \end{pmatrix},$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{p} = \begin{pmatrix} p(\boldsymbol{x}_1;\boldsymbol{\Omega}^{old}) \\ \vdots \\ p(\boldsymbol{x}_N;\boldsymbol{\Omega}^{old}) \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} — & \boldsymbol{x}_1^T & — \\ & \vdots & \\ — & \boldsymbol{x}_N^T & — \end{pmatrix}$$

Índice

Técnicas de clasificación



Naive Bayes

Formulación general

La hipótesis de independencia estocástica, nos permite escribir la probabilidad a posteriori como:

$$P(Y \mid \boldsymbol{X}) = \frac{P(Y)}{Z} \prod_{l=1}^{n} P(X_l \mid Y)$$

$$\textit{donde} \qquad Z = P(\boldsymbol{X}) = \sum_{k=1}^{m} P(Y = k) P(\boldsymbol{X} \mid Y = k),\ k \in \{1, \ldots, m\}.$$

El argumento de la técnica desemboca en maximizar la probabilidad a posteriori, sabiendo que $P(Y = k) = \frac{N_k}{N}$:

$$\hat{y} = \underset{y \in \{1,\ldots,m\}}{\operatorname{argmax}} P(Y = y) \prod_{l=1}^{n} P(X_l \mid Y = y)$$

k-Nearest neighbors

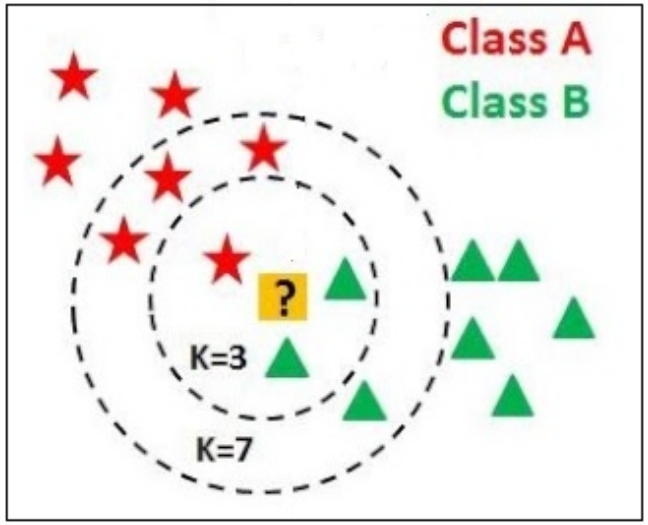
Planteaiento general

Partimos de un conjunto de datos de entrenamiento, $\mathbb{R}^n \times \mathcal{Y} \supseteq D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ con $\mathcal{Y}$ es un conjunto finito. Dada una nueva observación $\boldsymbol{x} \in \mathbb{R}^n$, la predicción de clase será:

$$\hat{y} = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \sum_{(\boldsymbol{x}_i, y_i) \in D_k(\boldsymbol{x})} \chi_c(y_i), \quad \textit{donde } c \equiv \{c\}, \quad D_k(\boldsymbol{x}) \subseteq D$$

Observación

- $\chi_A(z) = \begin{cases} 1, & z \in A \\ 0, & z \notin A \end{cases}$
- El concepto cercanía está asociado a una distancia, $d_p(\boldsymbol{r}, \boldsymbol{q}) = \left(\sum_{i=1}^{n} |q_i - r_i|^p\right)^{\frac{1}{p}}$.



Decision trees

CART

Partimos de un conjunto de datos de entrenamiento, $D = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N} \subseteq \mathbb{R}^n \times \mathcal{Y}$, donde $\mathcal{Y}$ es un conjunto finito. Dividimos el conjunto de datos en dos subconjuntos:

$$\begin{cases} D_1 &= \{(\boldsymbol{x}_k, y_i) \in D \mid x_k^j \leq s\} \\ D_2 &= \{(\boldsymbol{x}_k, y_i) \in D \mid x_k^j > s\} \end{cases}, \quad x_k^j, s \in,\ j \in \{1, \ldots, n\}.$$

Criterio

Reducción de la impureza, medida con el índice de Gini $G(t) = 1 - \sum_{k=1}^{m} p_k^2$ donde $p_k = \frac{1}{|t|} \sum_{k=1}^{|t|} \chi_c(y_k)$:

$$\Delta G = G(t) - \left(\frac{|D_1|}{|t|} G(D_1) + \frac{|D_2|}{|t|} G(D_2) \right)$$

Figura: Leo Breiman.

Support vector machines

Caso binario linealmente separable

Sea el conjunto de entrenamiento $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, con $\boldsymbol{x}_k \in \mathbb{R}^n$ e $y_k \in \{-1, +1\}$. Maximizar el margen $\frac{2}{\|\boldsymbol{\Omega}\|}$, equivale a minimizar $\frac{1}{2}\|\boldsymbol{\Omega}\|^2$.

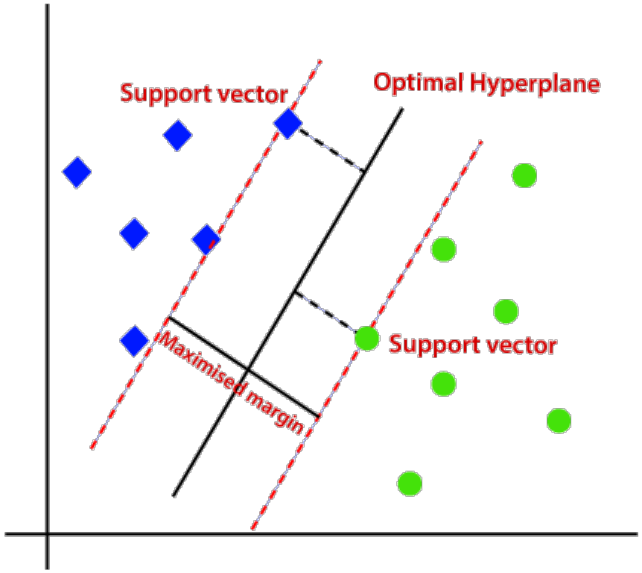
$$\begin{aligned} \text{mín}: \quad & \frac{1}{2}\|\boldsymbol{\Omega}\|^2 \\ \text{sujeto a:} \quad & y_k(\boldsymbol{\Omega}^T \boldsymbol{x}_k + b) \geq 1, \quad \forall k = 1, \ldots, N. \end{aligned}$$

Generalización

$$\text{mín}: \frac{1}{2}\|\boldsymbol{\Omega}\|^2 + C\sum_{i=1}^{n} \xi_k$$

$$\text{s.a: } y_k(\boldsymbol{\Omega}^T \boldsymbol{x}_k + b) \geq 1 - \xi_k$$

$$\xi_k \geq 0, \quad \forall k = 1, \ldots, N$$



Linear discriminant analysis

Planteamiento general

Partiendo de un conjunto de clases $\{1,\ldots,m\}$ y siendo $f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$, se quiere determinar la probabilidad a posteriori:

$$P(Y=k \mid \boldsymbol{X}=\boldsymbol{x}) \quad = \quad \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{m} f_l(\boldsymbol{x})\pi_l}, \quad \boldsymbol{x}^T = (x_1,\ldots,x_n).$$

Estimación de parámetros

Utilizando los datos de entrenamiento $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^n \times \mathbb{R}$, las estimaciones son:

- $\hat{\pi}_k = \frac{N_k}{N}$, $N_k :=$ *número de observaciones de clase* k.
- $\hat{\boldsymbol{\mu}}_k = \sum_{g_i=k} \frac{\boldsymbol{x}_i}{N_k}$, $g_i \in \{1,\ldots,m\}$.
- $\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{m} \sum_{g_i=k} \frac{(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)^T}{N-m}$.

k-Means clustering

Planteamiento general

Dado el conjunto de datos $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ con $\boldsymbol{x}_m \in \mathbb{R}^n$, el objetivo del algoritmo *k-means* es dividir estos datos en k grupos disjuntos $C_1, \ldots, C_k$ de tal forma que se minimice:

$$G_{k-means} = \sum_{m=1}^{k} \sum_{\boldsymbol{x} \in C_m} \|\boldsymbol{x} - \boldsymbol{\mu}_m\|^2, \text{ donde } \boldsymbol{\mu}_m = \frac{1}{|C_m|} \sum_{\boldsymbol{x} \in C_m} \boldsymbol{x}$$

El algoritmo puede inicializarse, elegiendo aleatoriamente k puntos como centroides, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$. Cada punto $\boldsymbol{x}_m$ se asigna al cluster cuyo centroide esté más cerca, lo que es equivalente a ir construyendo los clústeres a partir de los datos de D:

$$C_l = \left\{ x_m \in D \mid l = \underset{j \in \{1,\ldots,k\}}{\operatorname{argmin}} \|\boldsymbol{x}_m - \boldsymbol{\mu}_j\|, \ \forall\, m = 1, \ldots, N \right\}, \quad \forall\, l \in \{1, \ldots, k\}$$

Índice

Regresión

El dataset fue publicado originalmente por la Comisión de Taxis y Limusinas de Nueva York (TLC), disponibilizado en Kaggle (New York City Taxi Trip Duration) para una competencia cuyo objetivo era construir un modelo que permitiera predecir la duración total de los viajes en taxi en la ciudad de Nueva York.

Técnica	RMSLE (%)
linear regression	61,4
ridge regression	61,4
LASSO regression	62,9
elastic net regression	62,4

Cuadro: Comparativa de ténicas de regresión.

RMSLE

Para p_k la predicción de la duración del viaje y a_k la duración real del viaje, tenemos:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\log\{p_k+1\} - \log\{a_k+1\}\right)^2}$$

Clasificación

El origen de datos en este caso es el dataset MNIST, ampliamente utilizado por investigadores y estudiantes para reconocimiento de patrones y aprendizaje automático. La base de datos MNIST se derivó de datos originales del *National Institute of Standards and Technology* (NIST) y fue disponibilizada en Kaggle (MNIST Dataset) gracias a Yann A. LeCun Chief AI Scientist en Meta.

Técnica	k	Precisión (%)
k-nearest neighbours	1	95
	3	94,2
	5	93,5
	7	93
	9	92,4
SVM	–	95,1
k-means clustering	10	57,8

Cuadro: Comparativa de ténicas de clasificación y clustering.

Precisión

$$\text{Precisión} = \frac{\sum_m TC_m}{\sum_m (TC_m + FC_m)}$$

Índice

Novedades y desafíos

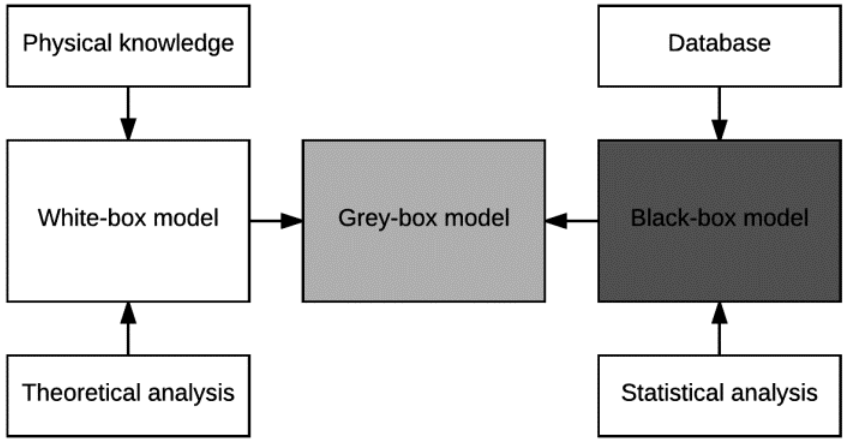


Figura: Enrico Camporeale.

El valor de las predicciones de los algoritmos de machine learning se hace evidente en el space weather, en situaciones de regresión como: valores de índices geomagnéticos, número de manchas solares, ocurrencia de erupciones solares, tiempo de propagación de un eyección de masa coronal, velocidad del viento solar y el flujo de partículas solares. Encontramos también circunstancias que dan pie a la clasificación, como: los distintos grupos de manchas solares y los tipos de viento solar de acuerdo con su origen.

Índice

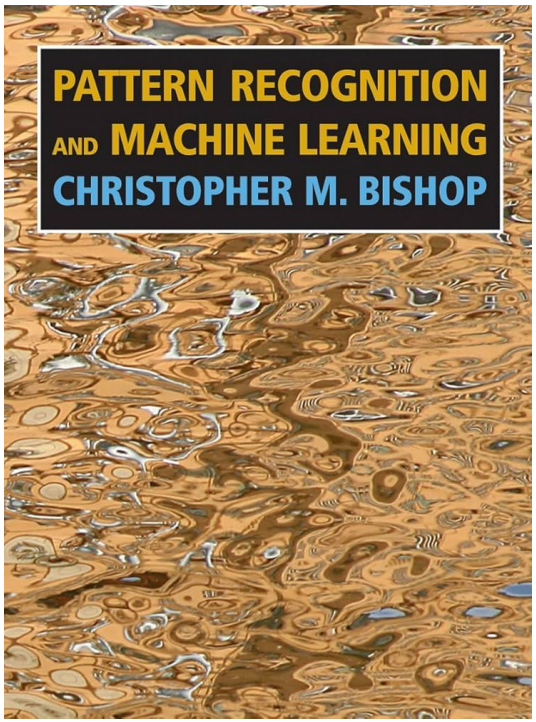
Conclusión

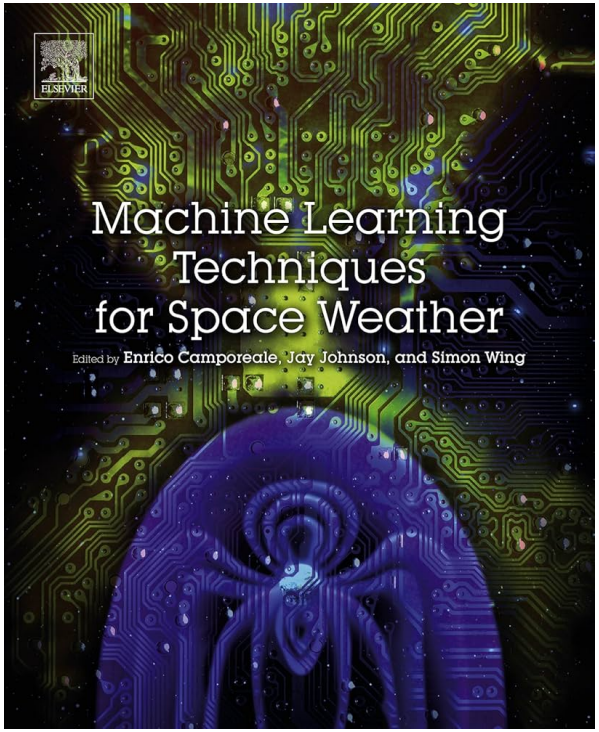
Identificamos dos puntos centrales a destacar que recogen la escencia de este proyecto y de los conocimientos adquiridos en su ejecución, estos son:

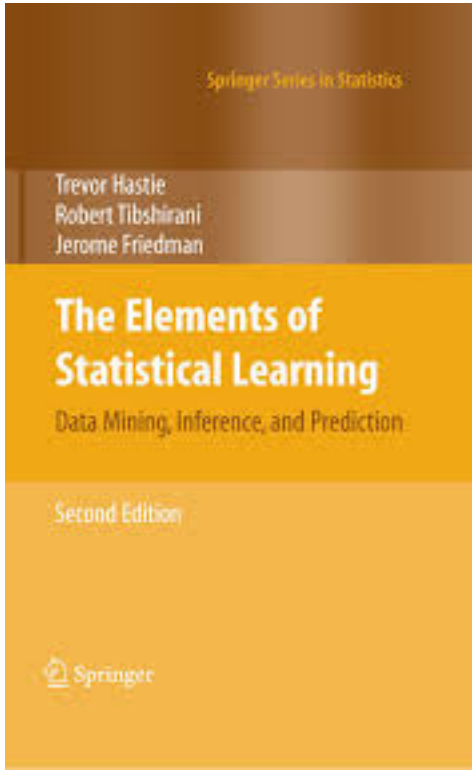
1. La profundización en algunas de las técnicas más distinguidas del aprendizaje supervisado y en una de las más conocidas de aprendizaje no supervisado.
2. La inmersión en el universo de las aplicaciones del aprendizaje automático, me permitió conocer el auténtico potencial transformador de esta herramienta, trascendiendo a las visiones parcializadas e incompletas de los medios de comunicación. A partir de esta probada capacidad revolucionaria, intuyo el enorme impacto que puede tener en disciplinas como la física, a tal punto que su irrupción puede suponer un punto de inflexión en la noble misión que históricamente se ha atribuido esta ciencia de descifrar los misterios del universo.

Índice

Referencias







¡Gracias!