Aprendizaje Supervisado

Modelos de clasificación

Hugo Franco

Universidad Central - Depto de Ingeniería de Sistemas

16 de mayo de 2022

UNIVERSIDAD
CENTRAL

Repaso: Aprendizaje supervisado

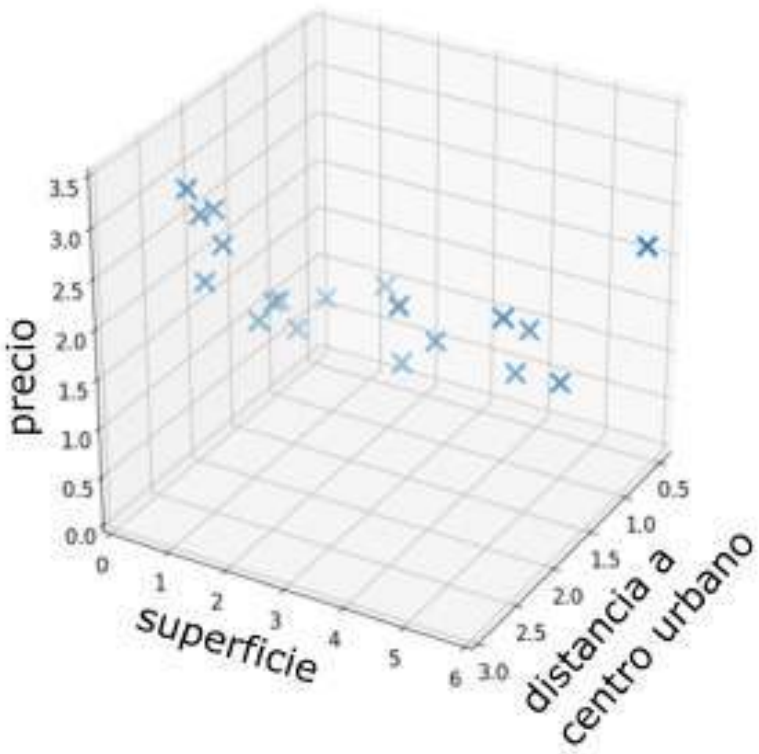
UNIVERSIDAD CENTRAL

Aprendizaje supervisado

En el aprendizaje supervisado, la entrada al algoritmo de entrenamiento corresponde a un conjunto de entradas (instancias) conocidas, determinadas por los valores de sus características, y cada una de ellas asociada a salida esperada del modelo (etiqueta). Esto es:

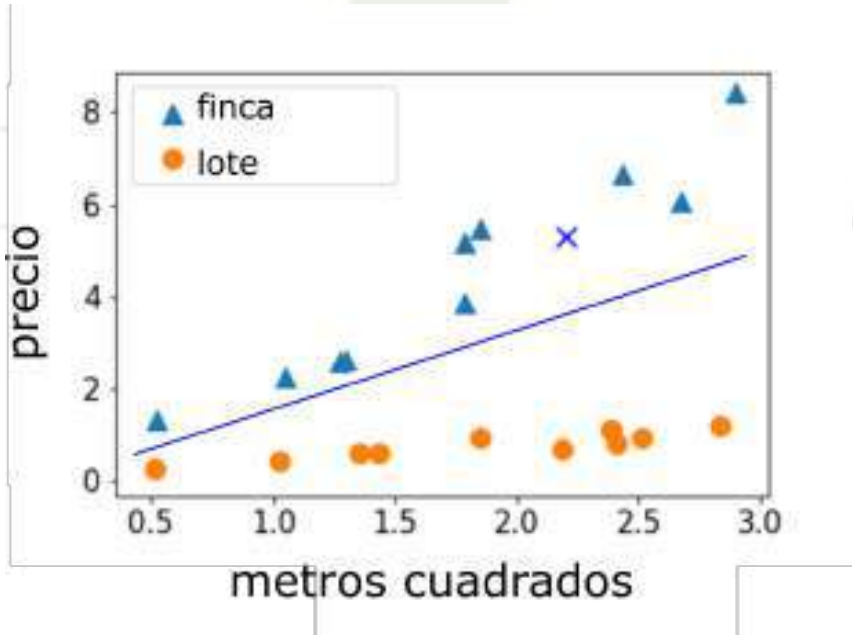
$$\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(m)}\right) \quad \rightarrow \quad y_j$$
$$\mathbf{x}_j \in \mathbb{R}^m$$

- El dataset corresponde entonces a la lista $\{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_n, y_n)\}$ para n «muestras» distintas, con $\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(m)}\right)$; las m características definen cada muestra
- y_j corresponde a la etiqueta o salida esperada para $\mathbf{x}_j$, puede ser continua (regresión) o categórica (clasificación).
- Ejemplo: en el caso de las fincas, se tienen dos características (superficie y distancia a centro urbano) asociadas a la etiqueta o salida esperada (precio de la finca)

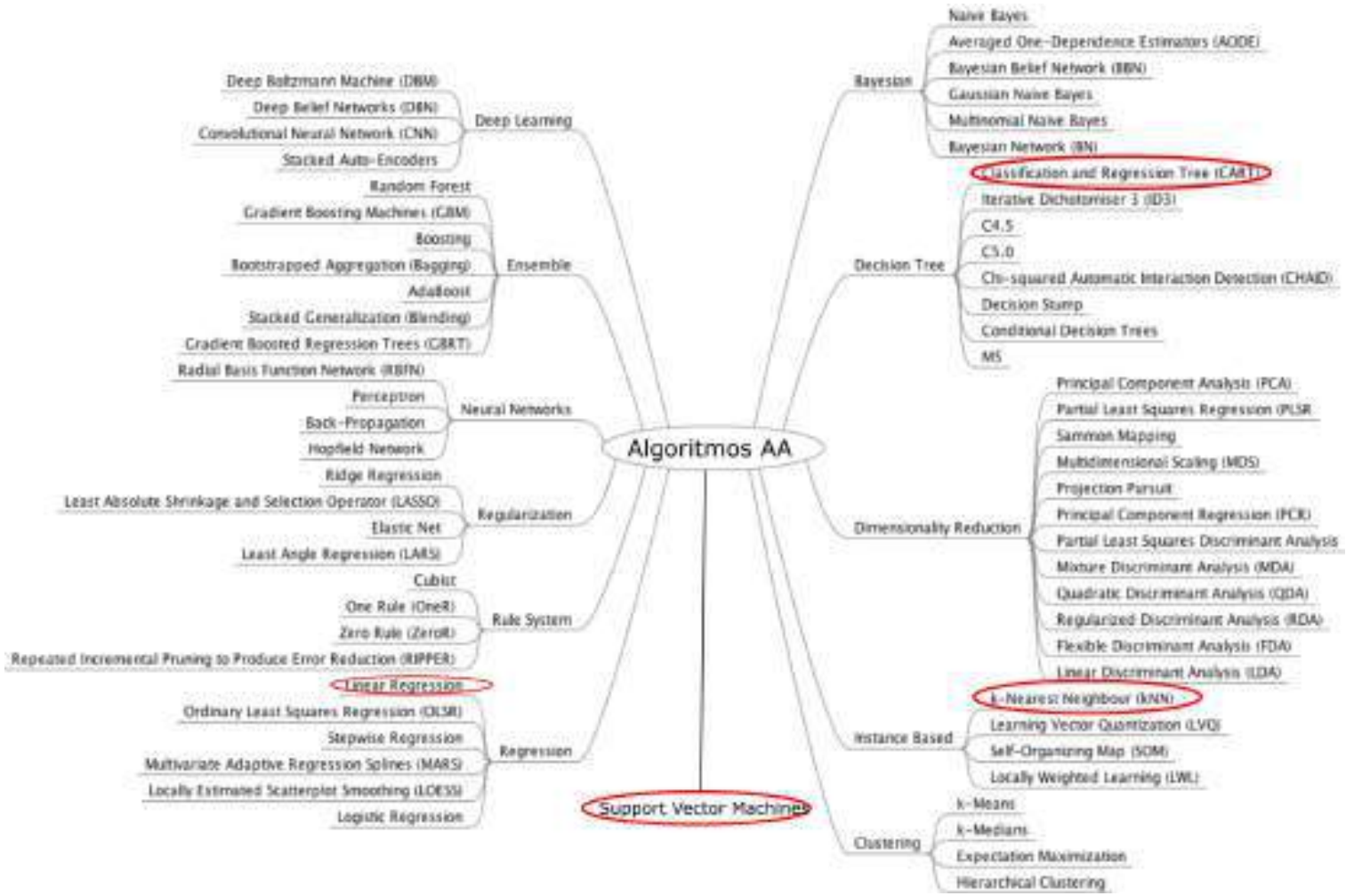


Tipos de métodos de aprendizaje supervisado

- Regresión: si $y_j \in \mathbb{R}$ es una variable continua.
  - En el ejemplo, predicción de precios dependiendo de características del predio
- Clasificación: la salida y_j del j-ésimo patrón de entrenamiento es una variable discreta (categórica)
  - En el ejemplo, identificar los tipos de predio según sus características



Métodos de aprendizaje supervisado



Ejemplo: Problemas de clasificación y regresión

- *Izquierda:* identificar clientes potenciales para venta de vivienda de acuerdo con el historial de los clientes de una entidad financiera. La variable objetivo es discreta (categórica). Clasificación
- *Derecha:* predecir la velocidad del viento de acuerdo con medidas previas de las condiciones atmosféricas registradas en un lugar. La variable objetivo es continua. Regresión.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Métodos de clasificación supervisada

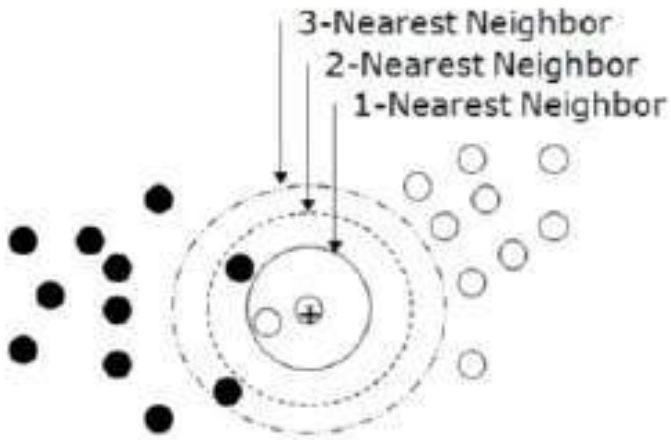
UNIVERSIDAD CENTRAL

Algoritmo KNN

UNIVERSIDAD CENTRAL

Algoritmo KNN

- **KNN:** *K vecinos más próximos*: cada nueva instancia se asigna a una clase según su proximidad a muestras previamente clasificadas dentro de una vecindad determinada por el valor de K.
- Es ineficiente ante conjuntos grandes de datos.



Algoritmo KNN

- Calcular la distancia a los demás elementos del dataset
- Determinar los K vecinos más cercanos
- Ejecutar una «votación por etiquetas», en la que la clase ganadora (mayor número de vecinos cercanos) se asigna a la nueva instancia.
- La función de distancia se define según la naturaleza de las características ($x^{(j)}$)

Árboles de decisión

UNIVERSIDAD CENTRAL

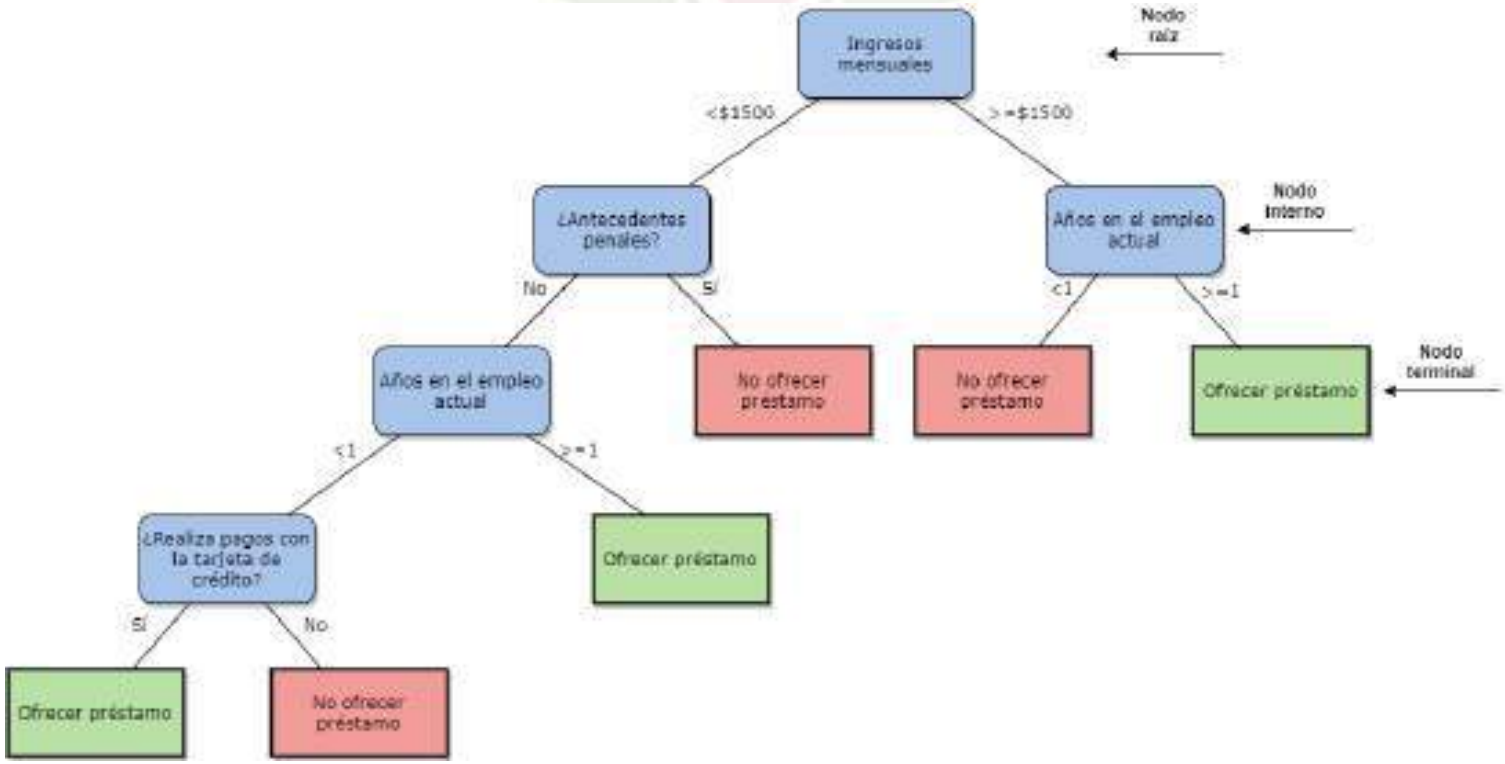
Árboles de decisión

- Un **árbol de decisión** es una estructura en árbol similar a un diagrama de flujo basado en la evaluación de condicionales

- El árbol de decisión se entrena («aprende») mediante el *ajuste de las reglas de decisión* de acuerdo con el valor recibido en cada atributo y su incidencia en la llegada al vértice terminal (hoja) correcto, en función del valor esperado de salida .
- Los árboles de decisión representan de forma más literal el flujo del pensamiento humano a la hora de ejecutar procesos de clasificación.
  - Por eso se consideran «cajas blancas» con notorias ventajas en la interpretación y, consecuentemente, se emplean en procesos de **selección de características**.
- También se pueden usar en problemas de regresión

Estructura de los árboles de decisión

- **En un árbol de decisión**
  - Los vértices internos representan las características en el dataset
  - Las aristas representan reglas de decisión
  - Cada una de las hojas (vértices terminales) representa un resultado posible dentro del conjunto de valores posibles de salida (etiquetas del dataset).



Algoritmos y medidas para el entrenamiento

- Dentro de los múltiples algoritmos para entrenar árboles de decisión, los más conocidos son los algoritmos **ID3** o **C4.5.**
  - Empieza por responder ¿cuál atributo del dataset debe ser la raíz del árbol?
  - Se evalúa cada atributo usando un test estadístico para determinar que tan correctamente clasifica los patrones de entrenamiento (el más representativo o el que mejor describe tal conjunto)
- Sea $X = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ el dataset y $c_1, \ldots, c_k$ el conjunto de clases. Los árboles de decisión utilizan medidas de información como:
  - La entropía de Shannon (medida de la información): $S(X) = -\sum_{j=1}^{k} p_j \log_2 p_j$ con p_j la probabilidad de encontrar un elemento de la clase c_i en el dataset X
  - La entropía condicional : $S(X|a) = \sum_{v \in vals(a)} \frac{|E_a(v)|}{|X|} S(E_a(v))$, con $E_a(v) = \{(\mathbf{x}, y) \in X | x_a = v\}$ y $v \in val(a)$ los valores que puede tomar x_a, generando una partición sobre X mediante un umbral definido sobre la variable x_a
  - La ganancia: $G(X) = S(X) - S(X|a)$: reducción esperada en entropía causada por conocer el valor de x_a (umbral)

Algoritmo C4.5

Entrada: (sub)conjunto de datos X

1. Evaluar
   1. si todas las muestras de la lista pertenecen a la misma clase: se crea un nodo hoja para el árbol de decisión que elige tal clase.
   2. si ninguna de las características provee ganancia de información: se crea un nodo de decisión más arriba en el árbol usando el *valor esperado* de la clase.
   3. si se encuentra una instancia de una clase nunca antes vista: se crea un nodo de decisión más arriba en el árbol usando el *valor esperado* de la nueva clase.
2. Para cada atributo x_i,
   1. calcular la ganancia de la información normalizada: dividir $G(X)$ por la entropía $S(E_a(v))$.
   2. determinar la tasa de ganancia de información normalizada al separar el conjunto X según x_i.
3. Identificar x_{best}: la característica (x_i) con la mayor ganancia de información normalizada.
4. Crear un nodo de decisión que se divida sobre x_{best}
5. Hacer recorrido recursivo con el subconjunto obtenido al dividir sobre x_{best} y agregar tales nodos como hijos del nodo actual.

Sobreajuste (Overfitting) - caso: árboles de decisión

El ***overfitting*** (sobreajuste) en un modelo de aprendizaje automático se produce cuando el entrenamiento ajusta los parámetros de manera que el modelo «memoriza» los datos de entrenamiento

- Cuando ello sucede, el modelo pierde su capacidad de **generalización** para las futuras predicciones
- *Consecuencia: no siempre un error grande en la medida de exactitud es indeseable*

- Se aborda con técnicas de poda de árboles:
  - **Técnicas de pre-poda:** limitan el crecimiento del árbol antes de que éste llegue a adaptarse perfectamente al conjunto de entrenamiento.
  - **Técnicas de post-poda:** permiten que el árbol se sobreajuste a los datos y luego se efectúa sobre él una poda. Es la más usada en casos prácticos
- La evaluación más exhaustiva de la capacidad de generalización se suele implementar mediante ***validación cruzada*** (ver más adelante)

Bosques aleatorios (Random forests)

UNIVERSIDAD CENTRAL

Métodos combinados o «*de ensemble*»

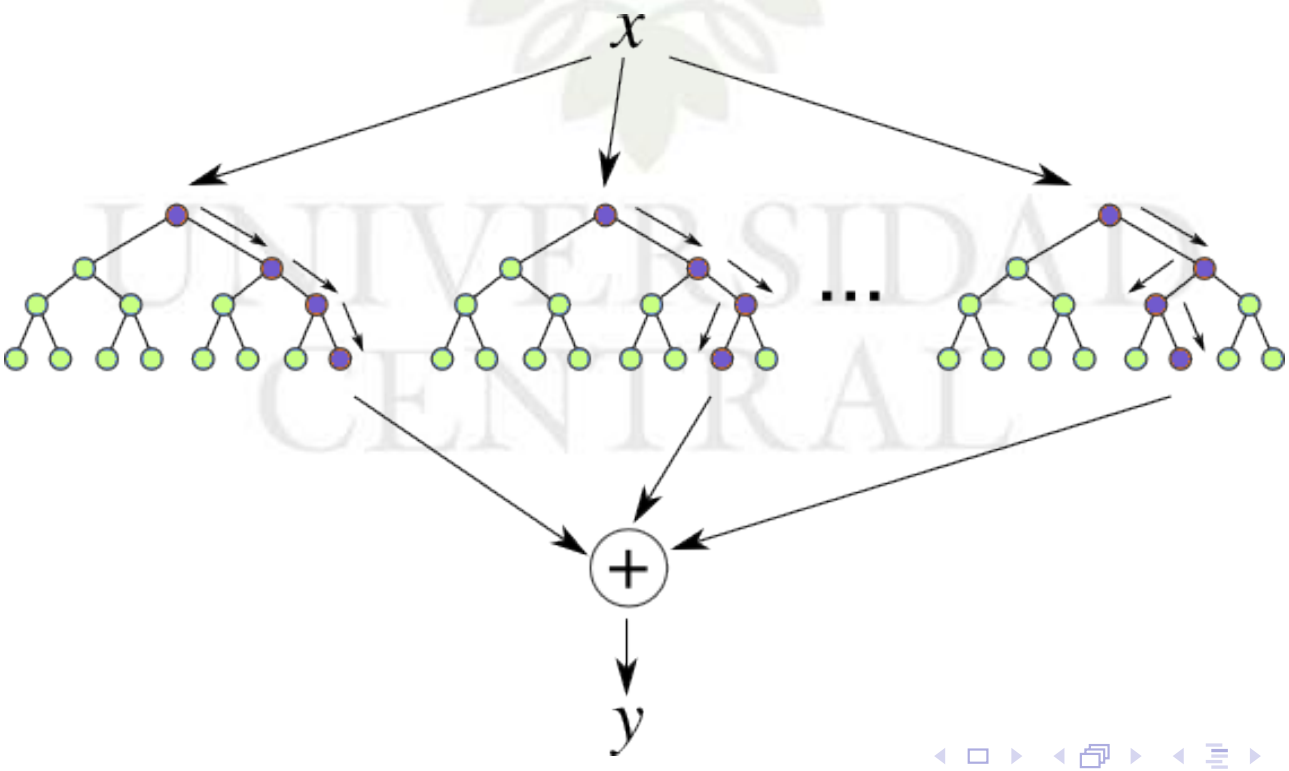
- Múltiples modelos de aprendizaje automático se pueden combinar para producir soluciones más fiables de acuerdo con el resultado obtenido por cada uno de ellos por separado.
- A estos se les conoce como «métodos combinados» o, más comúnmente, **métodos de ensemble** (*ensemble methods*).

Bosques aleatorios

Aunque casi cualquier modelo de aprendizaje automático podría ser usado bajo este enfoque, los «bosques aleatorios (*random forests*) están entre los más comunes.

Bosques aleatorios (random forests)

- Su resultado conjunto es una votación ponderada entre los resultados obtenidos por cada modelo por separado.
- Los árboles se construyen para que sean independientes entre sí.
- Cada árbol usa un subconjunto de las características ($x_k^{(j)}$) que conforman los patrones de entrada
  - Las características que usa cada árbol y los segmentos del dataset que usa para su entrenamiento se establecen de forma aleatoria (***bagging***)



Máquinas de Vector Soporte (SVM)

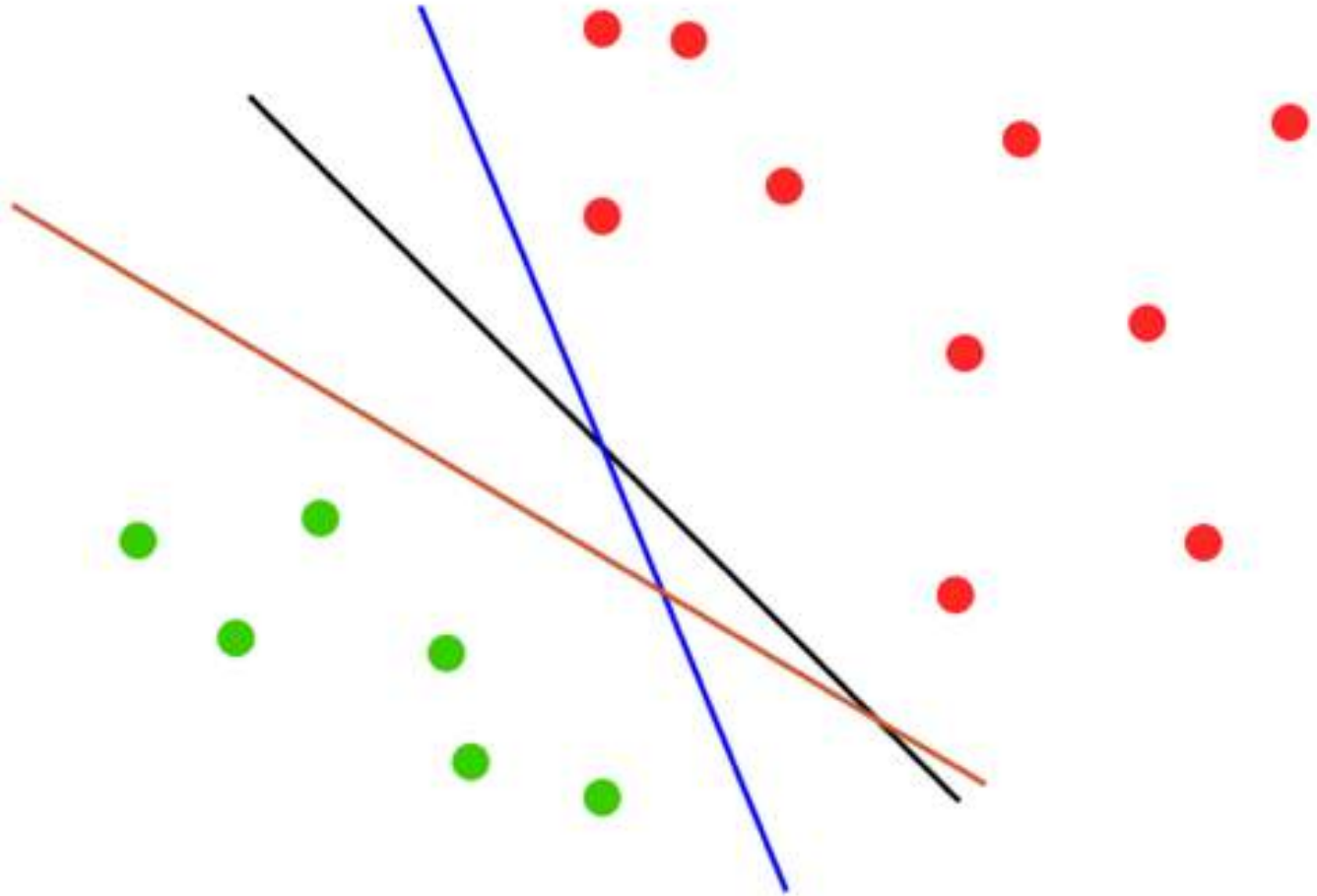
UNIVERSIDAD CENTRAL

Máquinas de vector soporte (SVM)

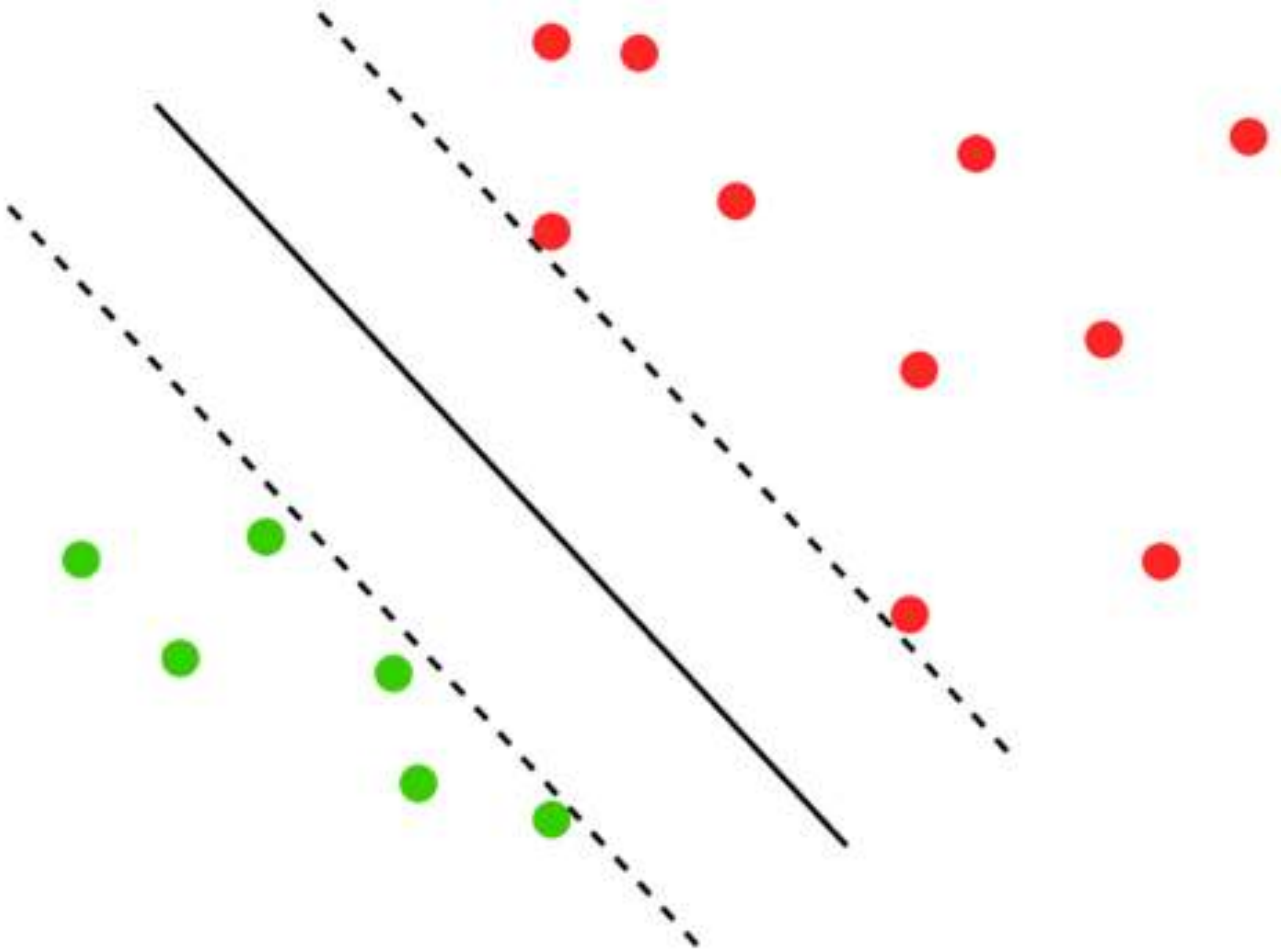
- La SVM es un modelo matemático cuyo entrenamiento consiste en determinar un conjunto de *hiperplanos* (*hipersuperficies*) en un espacio multidimensional ($\mathbb{R}^n$) que mejor separan los elementos asociados a diferentes clases.

- Los algoritmos usuales que implementan las SVM generan un *conjunto óptimo de hipersuperfices* de manera iterativa, minimizando una función de error.
- Pueden emplearse sobre representaciones de múltiples variables, continuas y categóricas.
- En general, las máquinas de vector soporte (Support Vector Machines) se consideran una técnica de clasificación, pero pueden utilizarse en regresión.

Multiplicidad de la solución



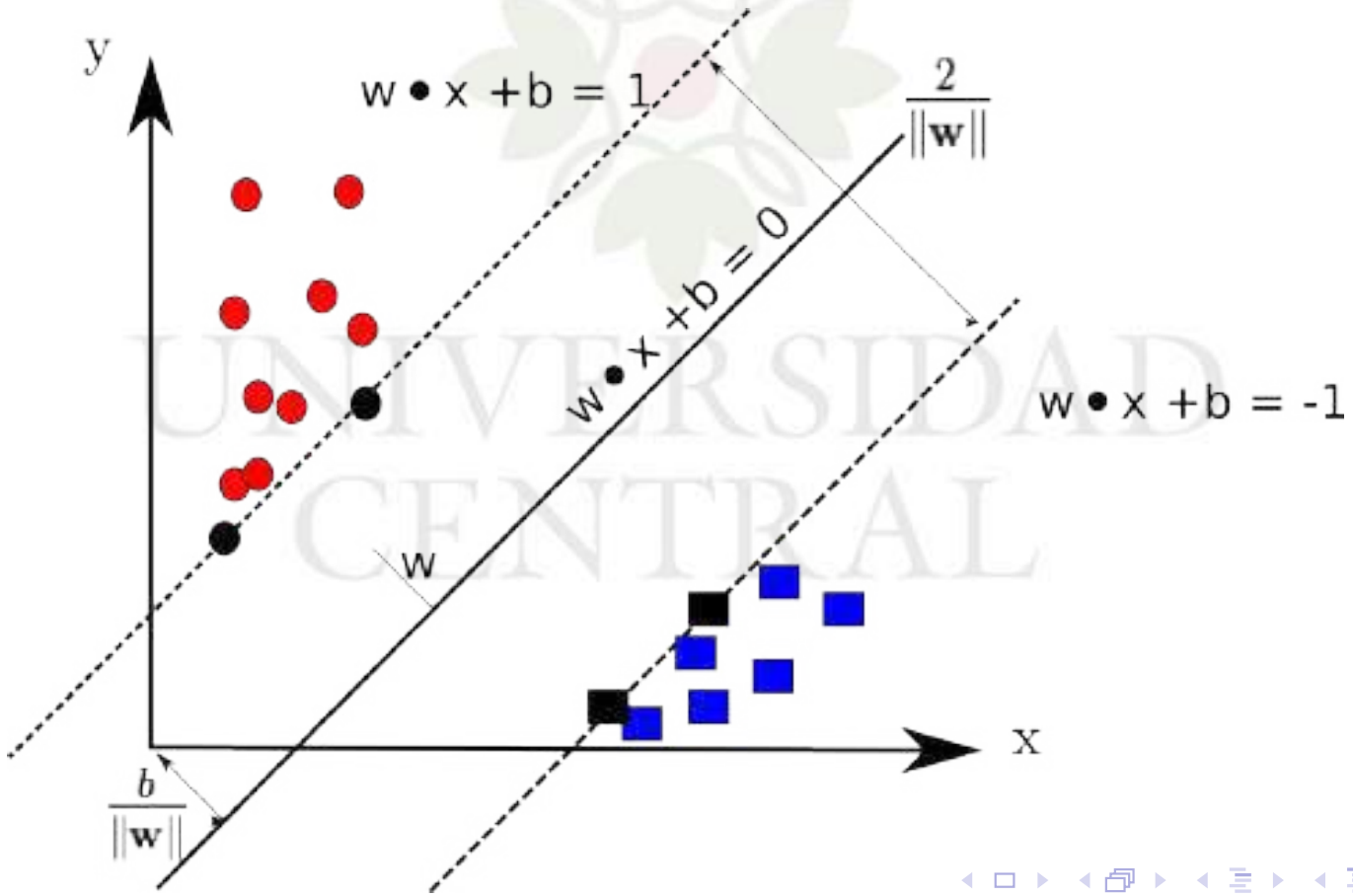
Búsqueda de una solución



Planteamiento del problema de optimización

Dado el problema de clasificación, se puede encontrar una solución única planteando encontrar la franja (banda) cuya recta central maximice la separación de conjuntos (optimización)

- $\mathbf{w} \cdot \mathbf{x}_i + b \geq 1$ si $y_i = 1$
- $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$ si $y_i = -1$



UNIVERSIDAD CENTRAL

Configuración de programación matemática

La separación de los conjuntos viene dada por $\frac{2}{\|\mathbf{w}\|}$. La tarea es *maximizar* esa distancia.

- Lo anterior es equivalente a *minimizar* $\frac{1}{2}\|\mathbf{w}\|$ o, por lo mismo, $\frac{1}{2}\|\mathbf{w}\|^2$ (para hacer convexo el problema, que siempre exista un mínimo):
  - Restricciones: Separación con margen,
    $\mathbf{w} \cdot \mathbf{x}_i + b \geq 1$ si $y_i = 1$
    $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$ si $y_i = -1$

  Esto a su vez equivale a la restricción $y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$. Así, el problema de optimización se puede formular como:

Minimizar $\frac{1}{2}\|\mathbf{w}\|^2$ sujeto a $y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 = 0$, para $1 \leq i \leq m$

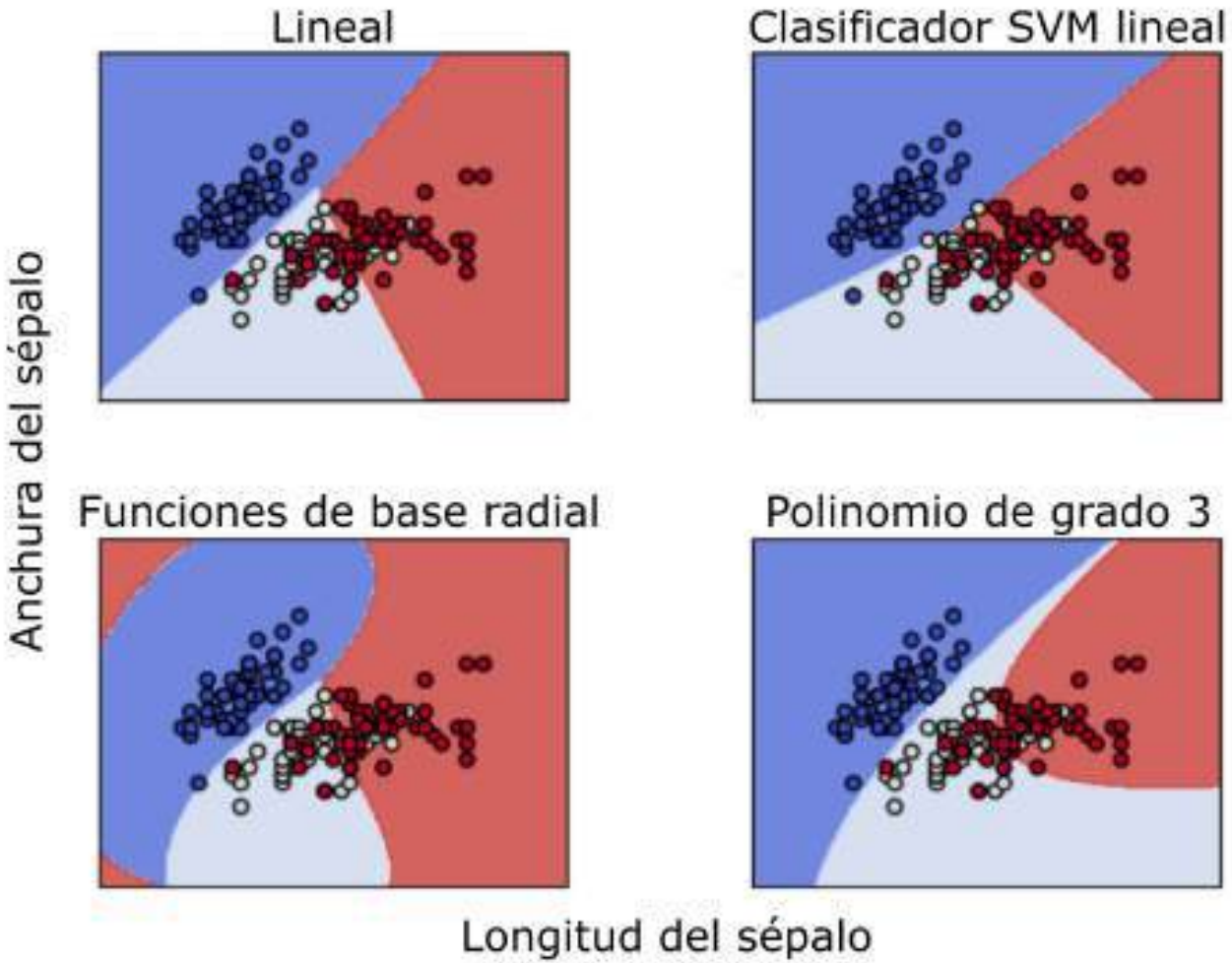
- El problema es convexo, por lo tanto, tiene un único valor mínimo. Existen algoritmos eficientes para resolverlo

Otras funciones base

El modelo presentado corresponde a una aproximación lineal (kernel). Se pueden usar otras funciones para identificar categorías dentro de datasets que no se puedan separar mediante una aproximación lineal (hiperplanos)

- Si se reemplaza el hiperplano de separación $(\mathbf{w} \cdot \mathbf{x}_i + b)$ por una superficie arbitraria determinada por la función $\mathbf{w} \cdot f(\mathbf{x}_i) + b$, se tiene que $f(x_1, \ldots x_m)$ determina la separación entre las dos regiones.
- Para ser usada como kernel en clasificación mediante SVM, una función debe cumplir la condición de Mercer: $\int\int k(x,y)g(x)g(y) \geq 0$ para toda $g(x)$ de cuadrado integrable (esto es $\int_{-\infty}^{+\infty} |g(x)|^2 dx < \infty$)
- Entre las funciones más usadas más usadas están:
  - Las funciones polinomiales $k(x,z) = (\mathbf{z}^T\mathbf{x} + c)^n$
  - Las funciones de base radial: $k(x) = \sum_{i=1}^{n} \mathbf{w}_i^T f\left(\frac{\|\mathbf{x}-\mathbf{x}_i\|}{\sigma}\right)$ con f una función exponencial para n el número de elementos en el dataset y σ un parámetro de normalización

Ejemplo: particiones obtenidas por *SVM* con diferentes fuciones de kernel



Evaluación de modelos de clasificación

UNIVERSIDAD CENTRAL

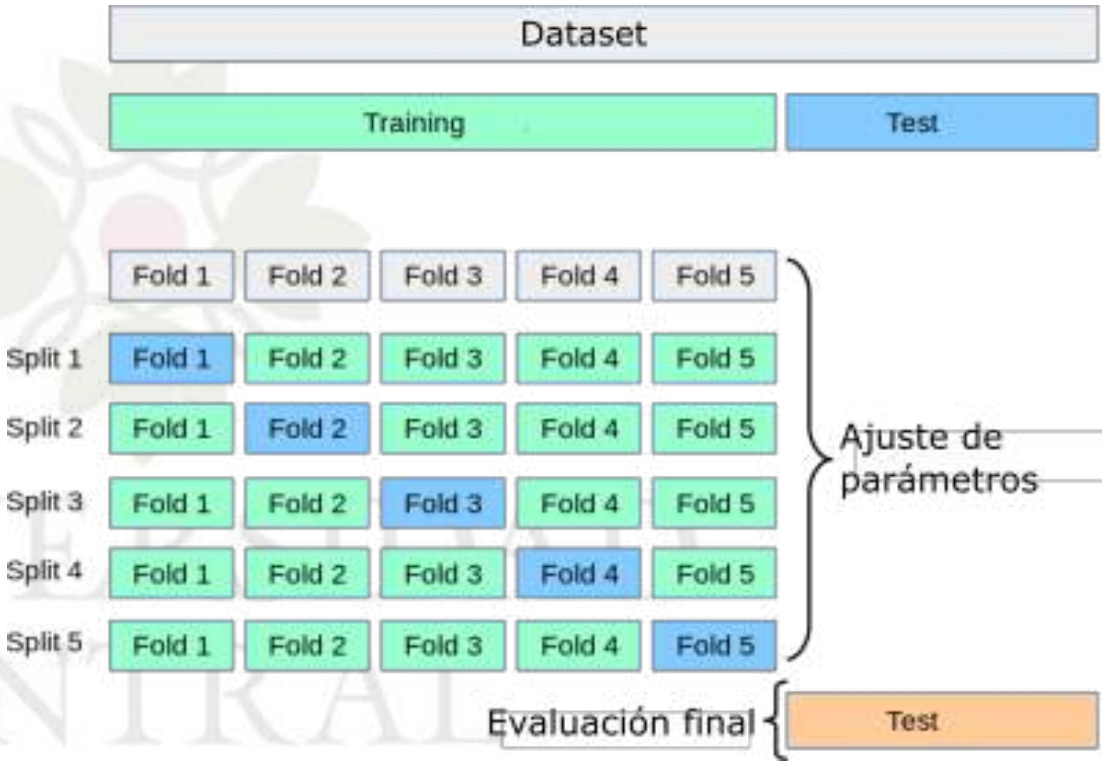
Entrenamiento, validación y prueba

Para evaluar el desempeño efectivo de un modelo de clasificación se deben usar datos de prueba –pares $(\mathbf{x}_i, y_i)$– que no hayan sido empleados en el entrenamiento (eliminar sesgos en la evaluación)

- Los modelos de clasificación, entre otros, suelen ser evaluados mediante técnicas estadísticas como la validación cruzada
- Una forma de evitar medidas sesgadas del desempeño en el entrenamiento es evaluar variables representativas del mismo para diferentes combinaciones de datos de entrenamiento y prueba

Validación cruzada *K-fold*

- A la derecha: K-fold cross validation con K=5. El desempeño general se reporta sobre un conjunto de datos jamás usado en el entrenamiento
- La precisión promedio (average accuracy) suele ser una medida de desempeño apropiada para este tipo de tareas



Medidas de desempeño de la clasificación

La matriz de confusión permite observar el desempeño de un modelo en función de sus capacidades para distinguir (clasificar) unas clases de otras sobre un datset determinado.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

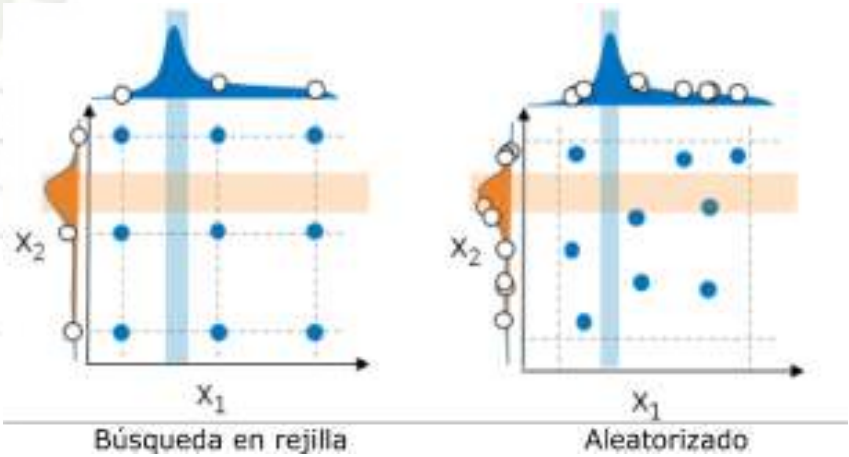
Medidas de desempeño:

- Accuracy $= \frac{\text{elementos bien clasificados}}{\text{número total de elementos del dataset}} = \frac{VP+VN}{VP+VN+FP+FN}$
- Precisión $= \frac{\text{elementos bien asignados a la clase}}{\text{total de elementos asignados a la clase}} = \frac{VP}{VP+FP}$
- Recall $= \frac{\text{elementos bien asignados a la clase}}{\text{total de elementos de la clase en el dataset}} = \frac{VP}{VP+FN}$
- Especificidad $= \frac{\text{elementos bien excluidos de la clase}}{\text{total de elementos de otras clases en el dataset}} = \frac{VN}{VN+FN}$
- F1-score $= 2\frac{\text{Precisión} \times \text{Recall}}{\text{Precisión}+\text{Recall}}$

Exploración de parámetros - Selección de modelos

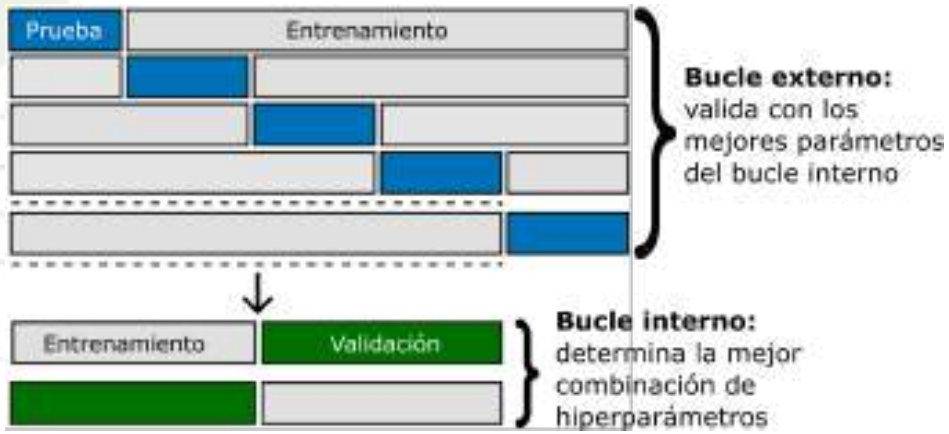
- Los modelos de clasificación (o regresión) pueden tener múltiples parámetros cuyos valores determinan la capacidad del modelo para dar una solución satisfactoria al problema de aprendizaje automático
- Los parámetros del modelo cuyo valor no cambia con el proceso de entrenamiento (aprendizaje) se conocen como «hiperparámetros»

- Pueden estar relacionados con la arquitectura del modelo (red, árbol de decisión, random forest) o con parámetros de control del proceso de aprendizaje (tasas de aprendizaje, coeficientes de las tasas de cambio)
- Se suelen usar métodos de exploración de parámetros por fuerza bruta (como Grid Search, izquierda) de búsqueda aleatoria (derecha) para encontrar la combinación de hiperparámetros que ofrece el mejor desempeño



Métodos de validación cruzada anidada

En muchos problemas, el tamaño de los datos en el *dataset*, o la variabilidad y sesgo de los mismos, genera incertidumbre en la capacidad de generalización de los modelos de aprendizaje automático.

Por ello, se usan técnicas de «validación cruzada anidada» que permitan garantizar la calidad del modelo.

- El bucle interno de estos métodos suele encargarse de la identificación de los mejores hiperparámetros para una partición conocida de entrenamiento y validación.
- El bucle externo toma este resultado y lo evalúa en el marco de un proceso de validación cruzada conocido.