Módulo 1

Lección 1

Conceptos de Aprendizaje Automático (Machine Learning)

Contenido

Haz clic sobre los títulos para navegar en cada tema.

Introducción

El aprendizaje automático es una forma de estadística aplicada con el uso de computadoras para estimar funciones complejas. Los algoritmos de aprendizaje automático se pueden dividir en las categorías de aprendizaje supervisado y no supervisado.

Un algoritmo de aprendizaje automático se construye combinando:

Un algoritmo de optimización

Una función de costo

Un modelo

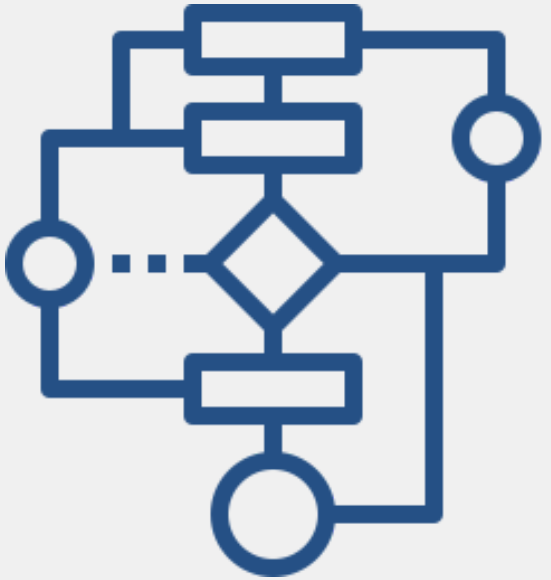
Un conjunto de datos

Volver a Contenido

Un algoritmo de aprendizaje es un algoritmo que puede aprender de los datos. Pero, ¿qué se entiende por aprender?

Mitchell (1997), proporciona la definición:

"Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tareas T y la medida de rendimiento R, si su Rendimiento en Tareas mejora con la Experiencia".



Tema 2. Tareas en aprendizaje automático

Una tarea es el proceso que se lleva a cabo sobre la entrada, para obtener una salida determinada. El aprendizaje automático se enfrenta con tareas difíciles de resolver para programas fijos escritos y diseñados por seres humanos.

Entendemos por aprendizaje el medio para obtener la capacidad de realizar la tarea.

Volver a Contenido

Conozcamos algunas tareas comunes en aprendizaje automático:

Clasificación y clasificación con entradas faltantes

Regresión

Máquina traductora

Salida estructurada

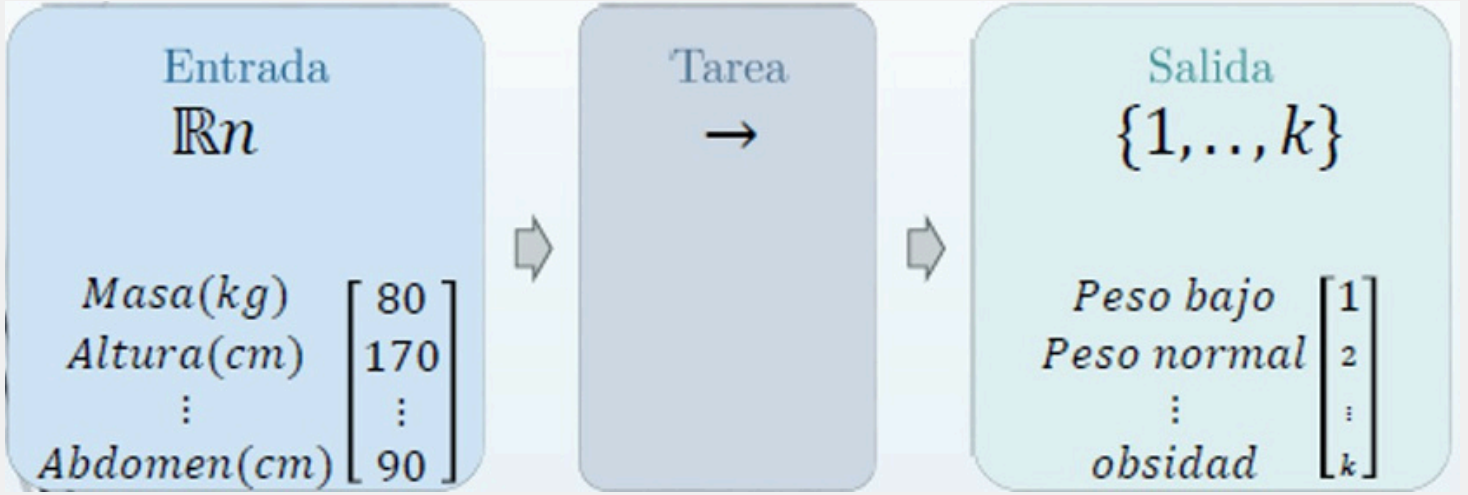
Limpieza de ruido

Estimación de función de densidad o masa de probabilidad

Veamos más a continuación.

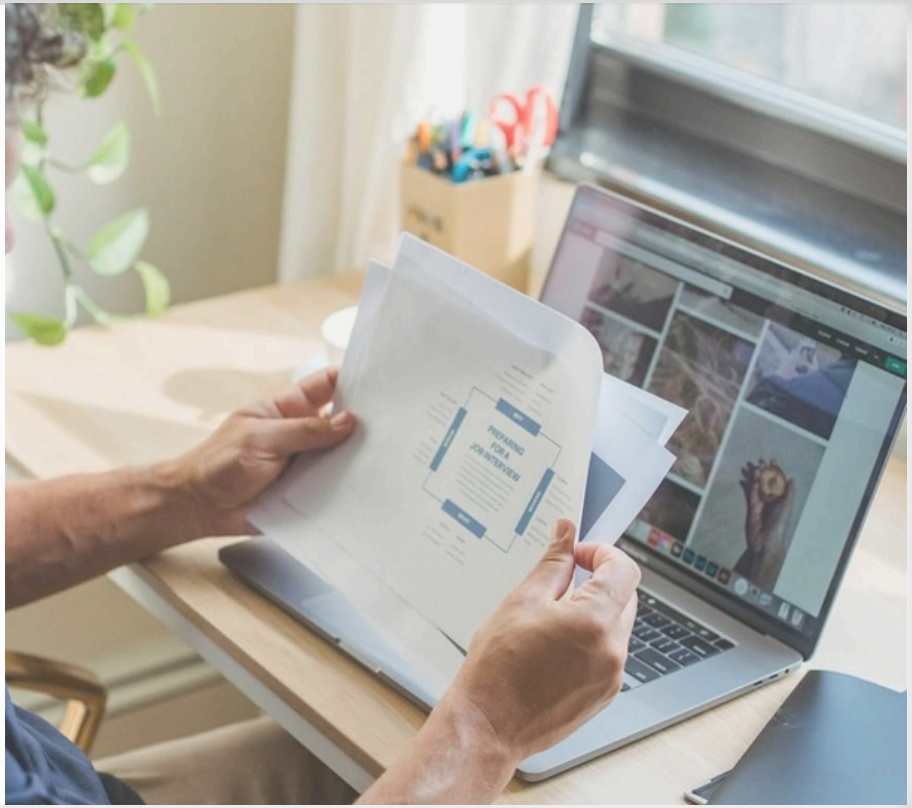
Tareas de clasificación

Consisten en determinar una categoría k en función de las características de entrada:



DISCUSIÓN GRUPAL SOBRE TAREAS DE CLASIFICACIÓN

- Ejemplos prácticos
- Casos de uso
- Aplicaciones en la vida real

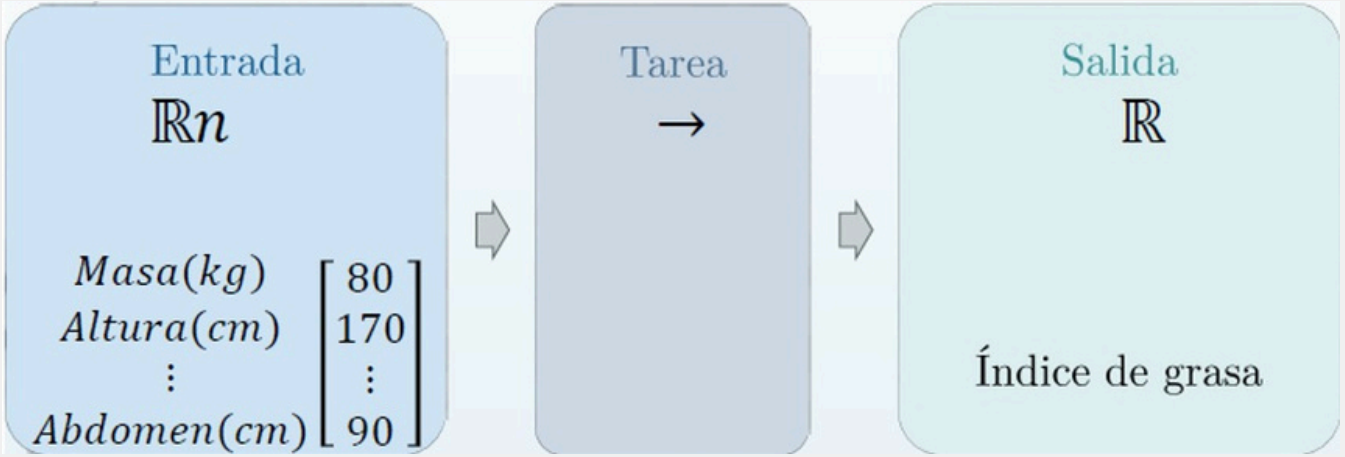


Clasificación con entradas faltantes

No garantiza todo el vector de entrada. También, el algoritmo aprende un conjunto de funciones.

Regresión

Obtiene algún valor numérico a partir de alguna entrada. Es similar a la clasificación pero con formato de salida diferente.

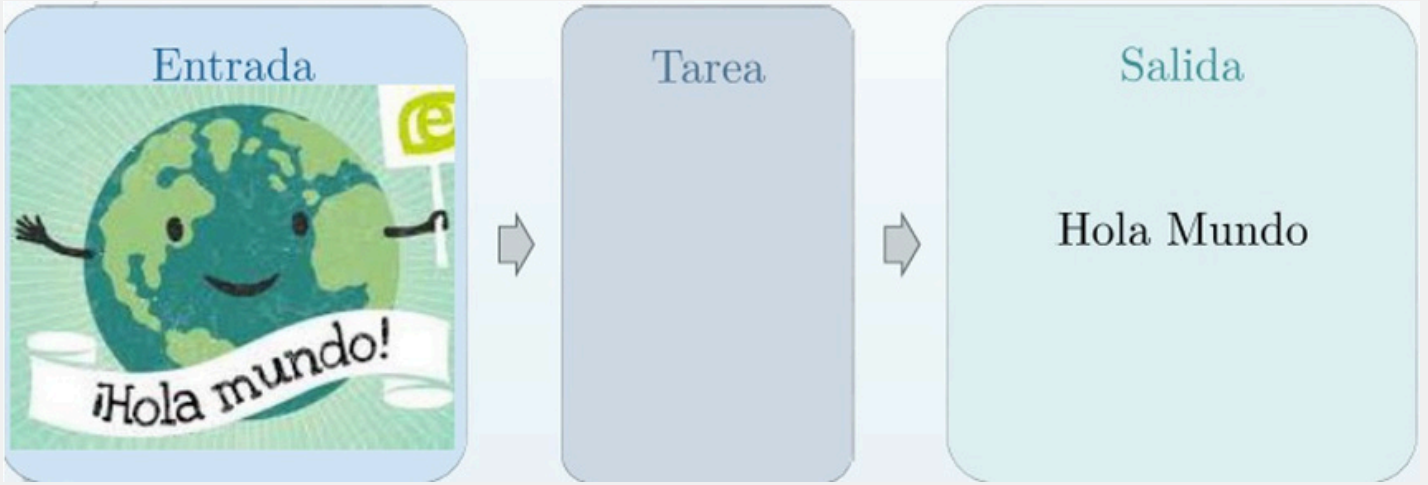


DISCUSIÓN GRUPAL SOBRE TAREAS DE REGRESIÓN

- Ejemplos concretos
- Relación con problemas del mundo real

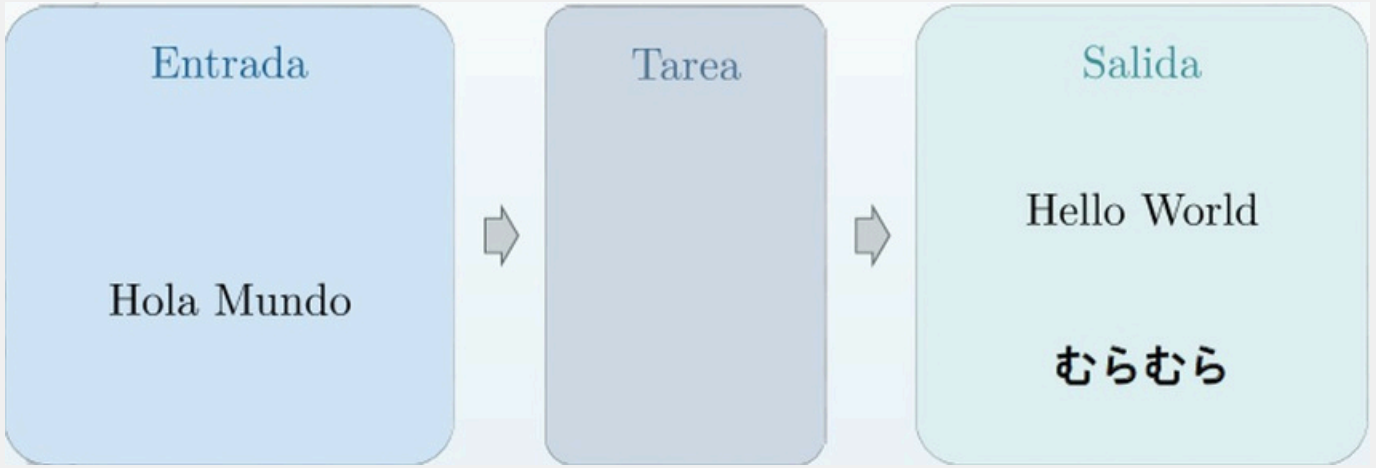
Transcripción

Parte de una representación desestructurada y la transcribe en una forma discreta y textual.



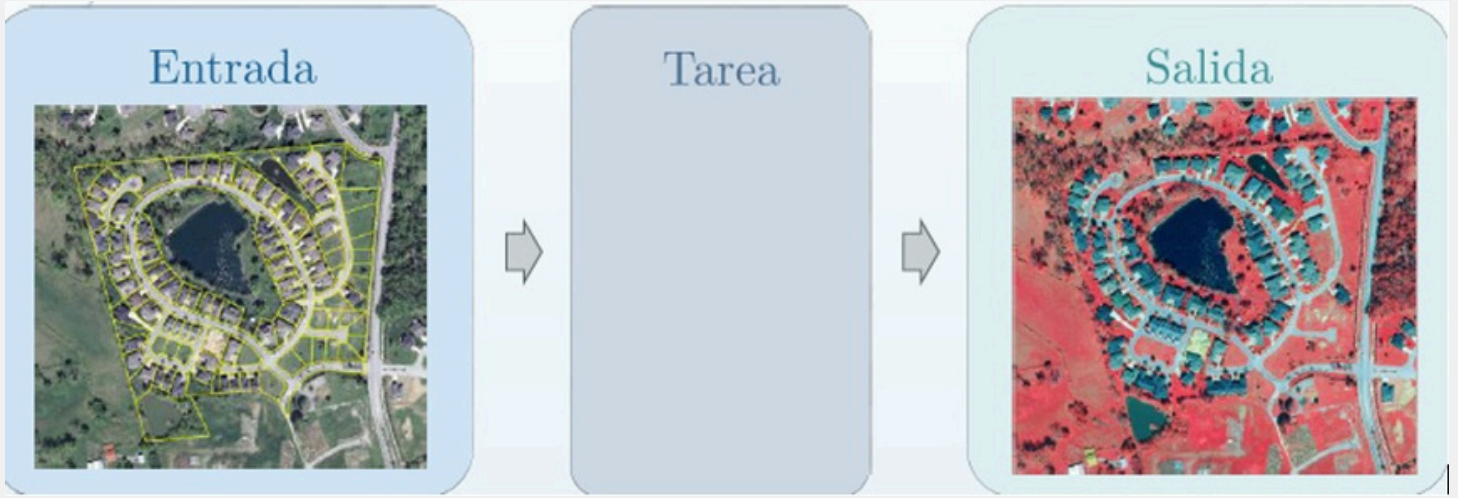
Tareas de traducción automática

Parte de una secuencia de símbolos en algún idioma, y el agente debe convertir esto en una secuencia de símbolos en otro idioma.



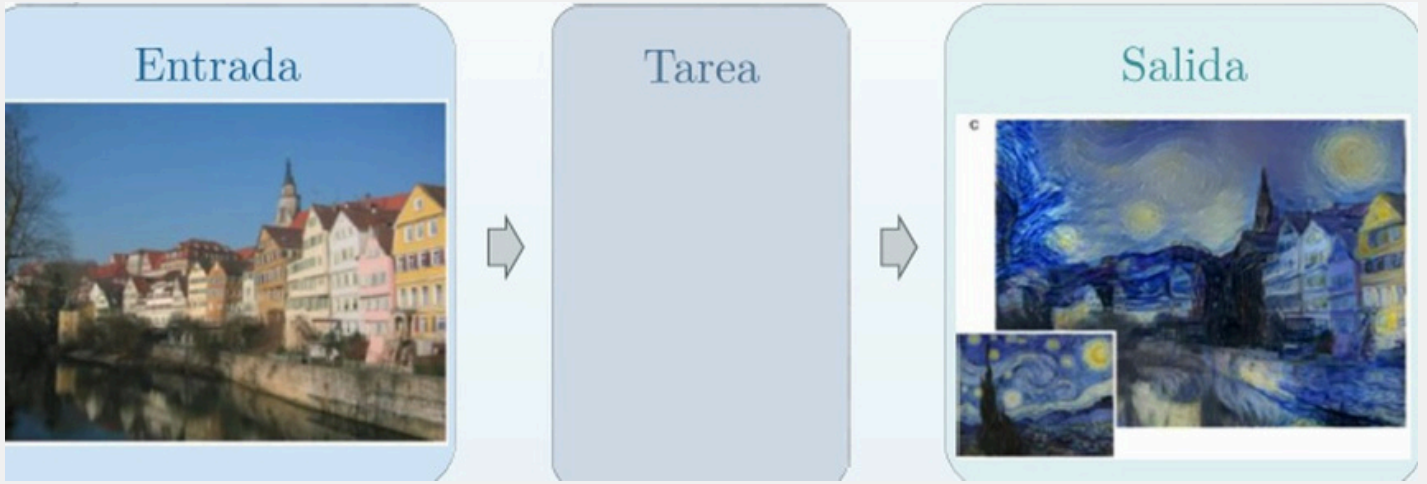
Tareas de salida estructurada

La salida es un arreglo u otra estructura de datos con múltiples valores que relaciona los elementos (transcripción y traducción).



Tareas de síntesis y toma de muestras

El agente genera nuevos ejemplos que sean similares a los de los datos de entrenamiento. Pueden haber múltiples salidas correctas.



Tareas de filtración de ruido

Parte de un ejemplo corrupto desconocido a partir de un ejemplo limpio X a partir de su versión corrompida.

$\tilde{x} \in \mathbb{R}n$ obtenido por un proceso de corrupción

$x \in \mathbb{R}n$ para obtener el ejemplo limpio

Tema 3. Conjuntos de datos y divisiones

Conjunto de Entrenamiento

Conjunto de Prueba

Conjunto de Validación

Volver a Contenido

Conjunto de Entrenamiento

El conjunto de entrenamiento es una parte del conjunto de datos utilizado para entrenar modelos de aprendizaje supervisado. Su función principal es permitir que el modelo aprenda patrones y relaciones entre las características de entrada y las salidas esperadas.

Las características de entrada

Las salidas esperadas

La calidad del modelo depende en gran medida de la representatividad del conjunto de entrenamiento. Debe abarcar la diversidad de casos que el modelo encontrará en situaciones reales para generalizar de manera efectiva.

Conjunto de Prueba

El conjunto de prueba se utiliza para evaluar el rendimiento del modelo después de que ha sido entrenado en el conjunto de entrenamiento. También, proporciona una evaluación imparcial del modelo, permitiendo medir su capacidad para generalizar a datos no vistos.

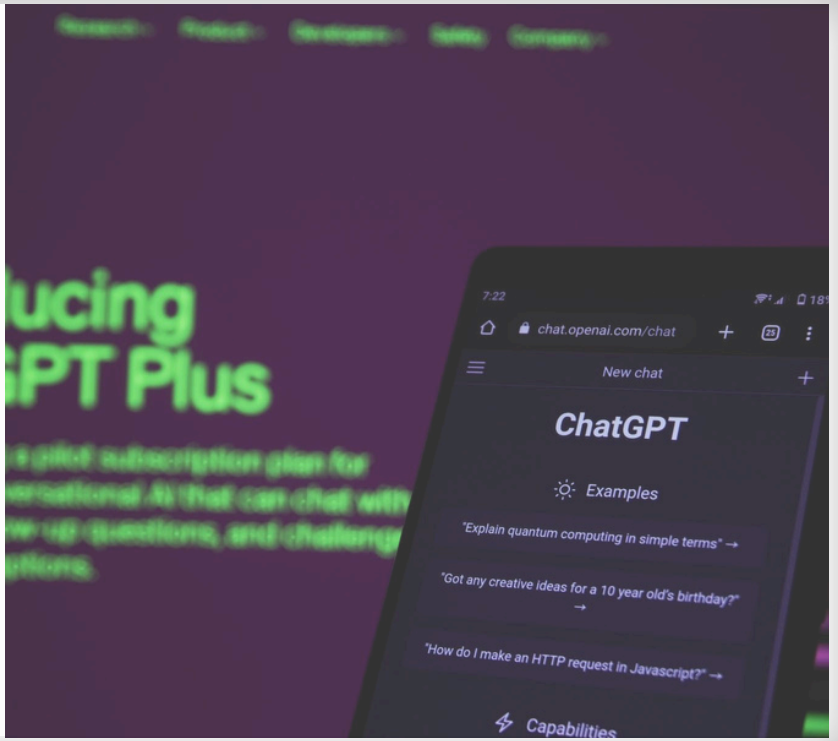
Es esencial que el conjunto de prueba sea independiente y no esté sesgado hacia ninguna característica específica. De lo contrario, las métricas de rendimiento pueden no reflejar la verdadera capacidad del modelo.

Conjunto de Validación

El conjunto de validación se utiliza durante el proceso de entrenamiento para ajustar hiperparámetros y evitar el sobreajuste. Este también permite:

Consideraciones generales

- El tamaño adecuado de los conjuntos de entrenamiento, prueba y validación depende del tamaño total del conjunto de datos y la complejidad del problema.
- Es crucial estratificar los conjuntos para mantener la proporción de clases en problemas de clasificación, evitando así desequilibrios.

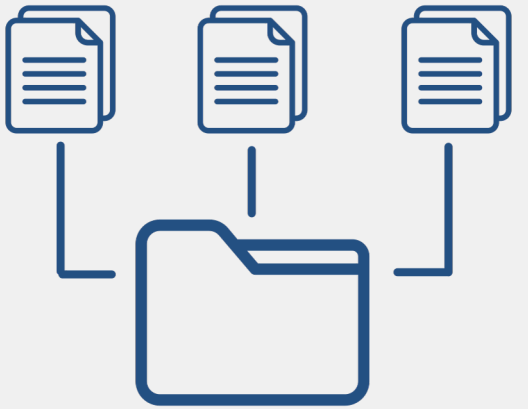


Consideraciones generales

- La correcta gestión de conjuntos de datos y divisiones es esencial para el desarrollo de modelos precisos y generalizables.
- Cada conjunto cumple un papel único en el proceso de aprendizaje supervisado, desde el entrenamiento inicial hasta la evaluación final.
- La representatividad y la imparcialidad son claves para obtener resultados confiables y aplicables en entornos del mundo real.

Tema 4. Técnicas de evaluación

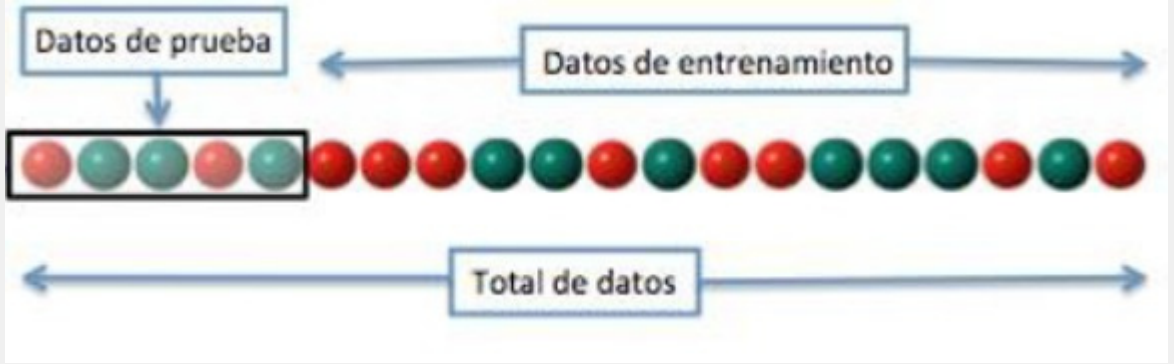
Una de las técnicas de evaluación fundamentales para valorar el desempeño de los modelos de aprendizaje automático es la validación cruzada. Se utiliza para evaluar el rendimiento de un modelo de manera robusta y evitar el sobreajuste. En lugar de depender de un solo conjunto de datos de entrenamiento y prueba, la validación cruzada divide los datos en múltiples conjuntos, permitiendo una evaluación más precisa del modelo.



Volver a Contenido

Medidas en el conjunto de validación

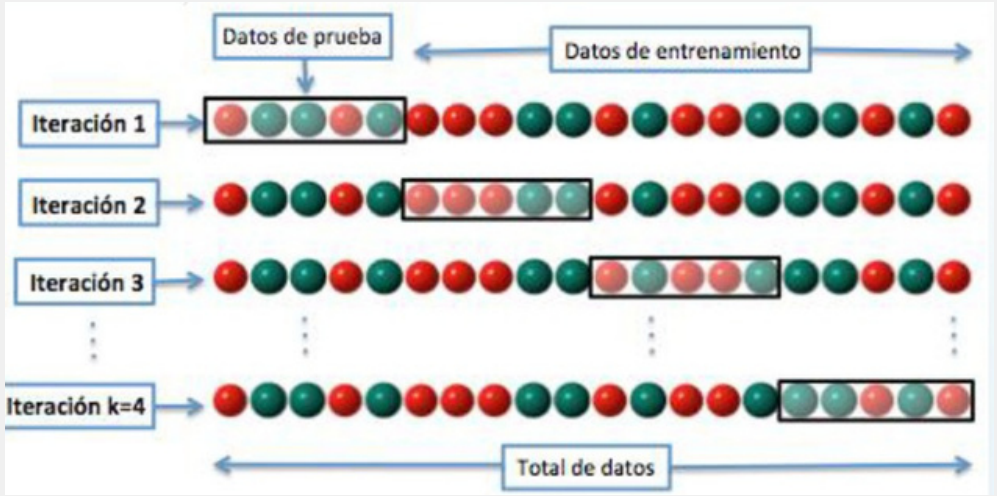
Veamos una representación del rendimiento del modelo en conjunto de prueba (error de generalización).



Se debe tener en cuenta que esta medida puede tener sesgos importantes.

Concepto fundamental

La validación cruzada implica dividir el conjunto de datos en k pliegues (o "folds"). Luego, el modelo se entrena k veces, utilizando k-1 pliegues para entrenamiento y el pliegue restante para prueba en cada iteración, esto se repite k veces, y los resultados se promedian para obtener una medida más robusta del rendimiento del modelo.



Pasos para implementar la validación cruzada

División de datos: divide el conjunto de datos en k pliegues.

Iteración del modelo: inicia un bucle que va desde 1 hasta k.
En cada iteración, selecciona k–1 pliegues para entrenar el modelo.

Evaluación del modelo: utiliza el pliegue restante para evaluar el rendimiento del modelo.

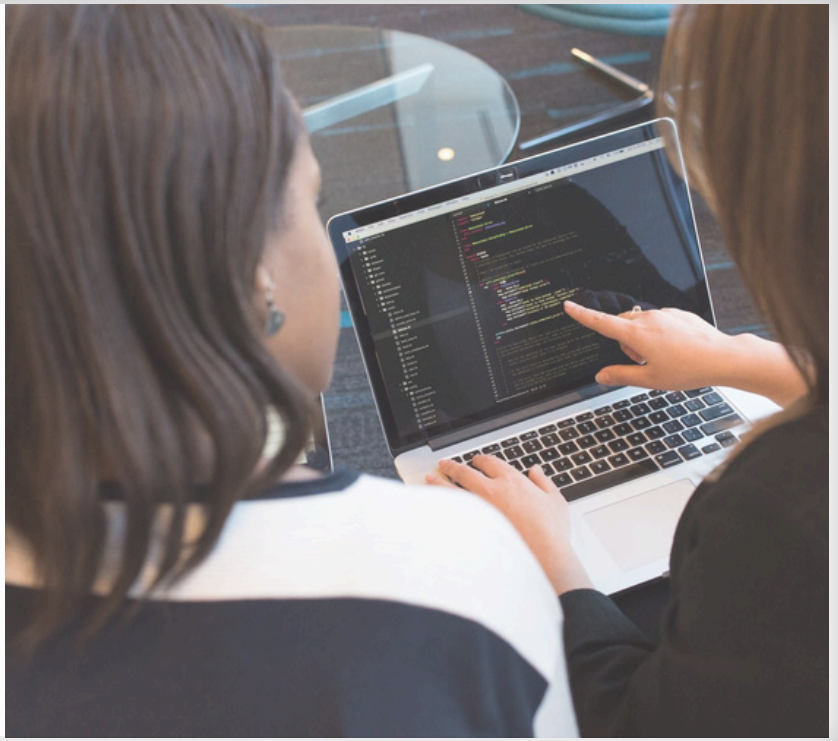
Promedio de resultados: repite el proceso k veces, registrando el rendimiento en cada iteración.

Calcula el promedio de los resultados para obtener una métrica global de rendimiento.

Ventajas de la validación cruzada

1. Mejora de la generalización: al evaluar el modelo en diferentes subconjuntos de datos, la validación cruzada proporciona una evaluación más precisa del rendimiento general del modelo.

2. Reducción del sobreajuste: al utilizar múltiples divisiones de datos, se reduce la probabilidad de que el modelo se ajuste demasiado a un conjunto de datos específico.



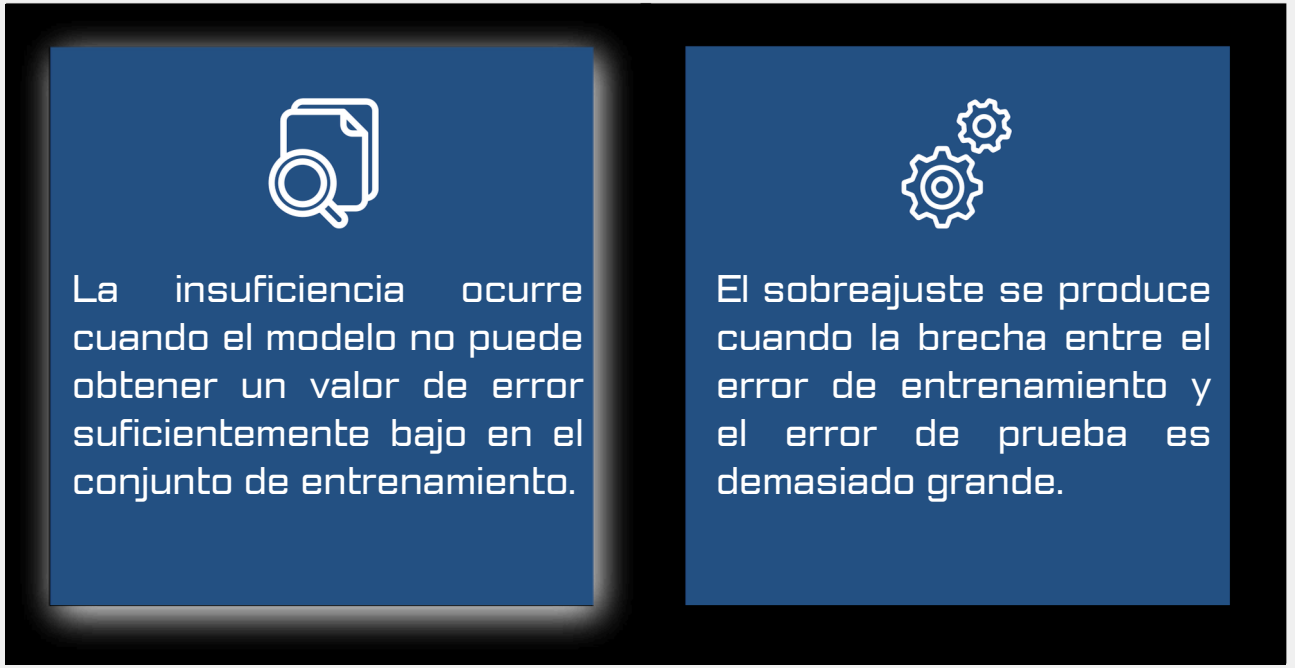
Ventajas de la validación cruzada

3. Mayor utilización de datos: aprovecha al máximo los datos disponibles para entrenamiento y prueba en múltiples combinaciones.

4. Evaluación fiable del rendimiento del modelo en comparación con una simple división de datos al en conjunto de entrenamiento y prueba:
implementar esta técnica, los practicantes del aprendizaje automático pueden tomar decisiones más informadas sobre la capacidad predictiva de sus modelos.

Capacidad del modelo

La capacidad de un modelo de adaptarse puede llegar a dos extremos:



Capacidad de generalización

Generalización es la capacidad de un algoritmo de obtener un buen desempeño para entradas previamente no observadas. El error de generalización o error de prueba, se define como el valor esperado del error en una nueva entrada y se estima al medir el rendimiento en un conjunto de pruebas.

Lo que se mide:

$$\frac{1}{m^{(\text{train})}}||\boldsymbol{X}^{(\text{train})}\boldsymbol{w} - \boldsymbol{y}^{(\text{train})}||_2^2$$

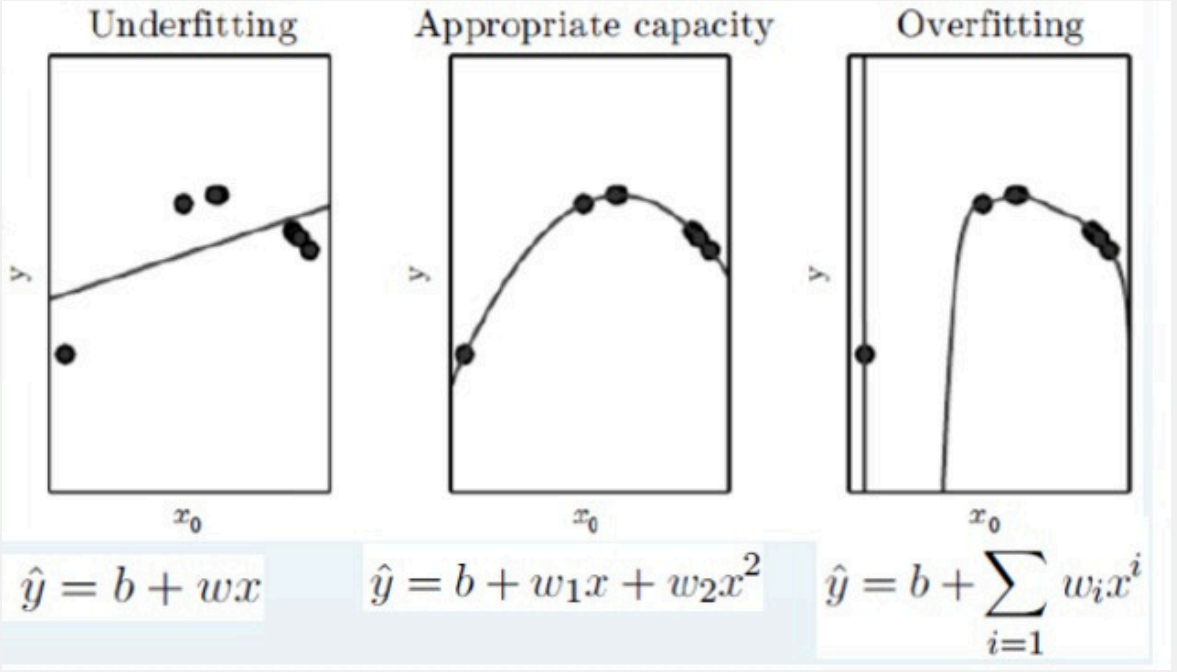
Lo que se quiere mejorar:

$$\frac{1}{m^{(\text{test})}}||\boldsymbol{X}^{(\text{test})}\boldsymbol{w} - \boldsymbol{y}^{(\text{test})}||_2^2$$

DISCUSIÓN GRUPAL ACERCA DEL SOBREAJUSTE

- Identificación del sobreajuste y sus consecuencias
- Estrategias para mitigar el sobreajuste

Ejemplo de regresión polinomial



Capacidad vs Error