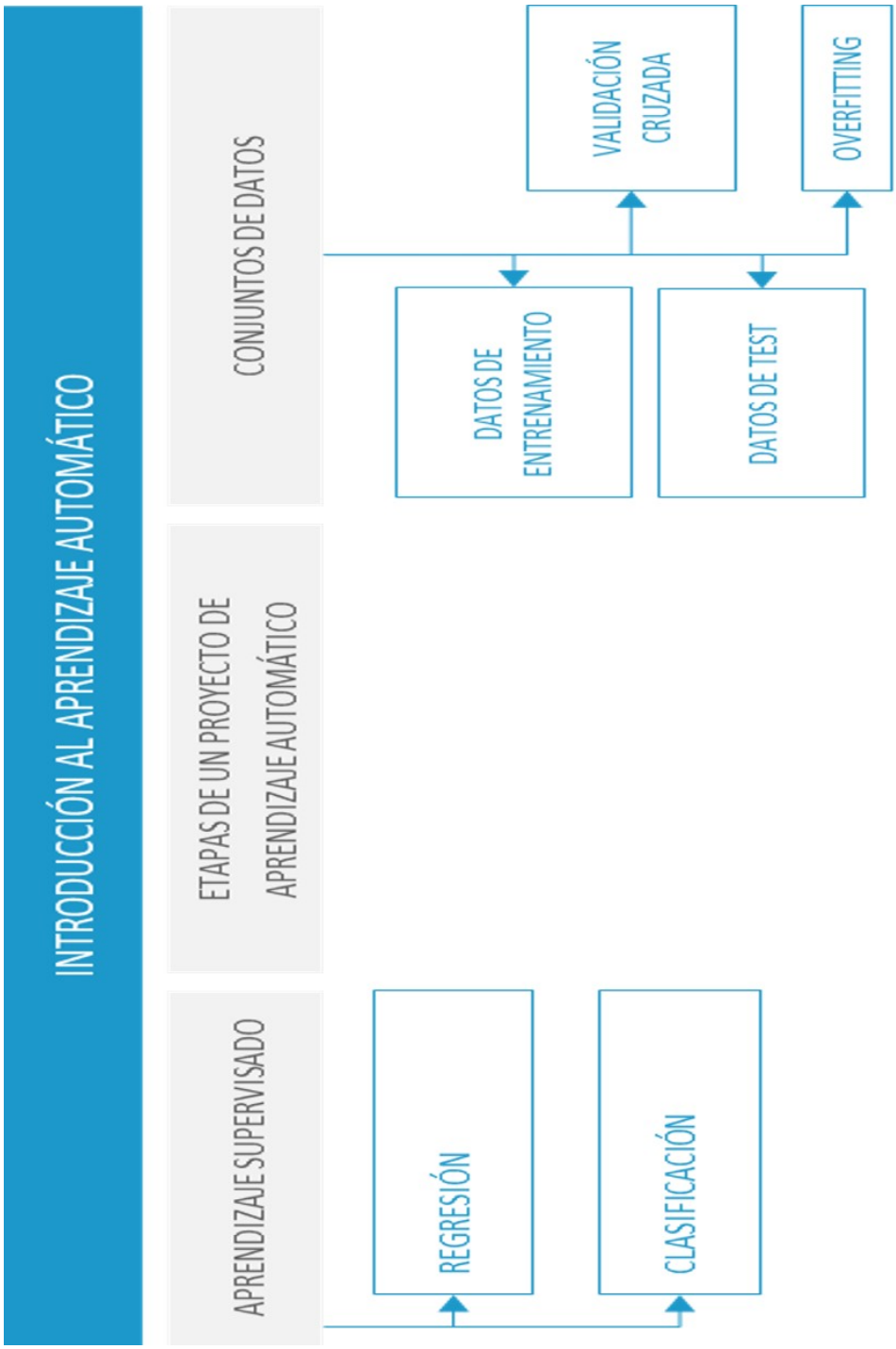
Técnicas de Aprendizaje Automático

Tema 1. Introducción al aprendizaje automático

Índice

Esquema



1.1. Introducción y objetivos

En este tema introduciremos los conceptos básicos del **aprendizaje automático**. Se entiende por aprendizaje automático aquellos algoritmos que se ejecutan en los ordenadores para aprender automáticamente en base a los datos proporcionados. Se trata de crear programas capaces de generalizar comportamientos a partir de los datos suministrados en forma de ejemplos. Es, por tanto, un proceso de inducción del conocimiento. El aprendizaje automático es fundamental para aplicaciones como la clasificación de imágenes, el procesamiento de lenguaje natural y la toma de decisiones automatizadas.

El aprendizaje automático, también conocido como *machine learning*, se divide tradicionalmente en tres amplías categorías, que se corresponden con paradigmas de aprendizaje según la naturaleza de los datos y de la disponibilidad de los mismos. Estas tres categorías son: **aprendizaje supervisado, aprendizaje no supervisado**, y aprendizaje por refuerzo. El aprendizaje supervisado utiliza ejemplos conocidos para obtener las inferencias mientras que el aprendizaje no supervisado no dispone de ejemplos con un objetivo o etiqueta conocidos. Finalmente, en el aprendizaje por refuerzo un agente *software* aprende a medida que interactúa con el entorno. Los datos los obtiene a partir de la respuesta del entorno a las acciones que el agente realiza.

A su vez, los **problemas** se pueden dividir en los siguientes cuatro subtipos:

- Aprendizaje supervisado.
  - Problemas de regresión: tienen como objetivo predecir un valor numérico continuo.
  - Problemas de clasificación: tienen como objetivo predecir el valor de una etiqueta/categoría.

- Aprendizaje no supervisado.
  - Problemas de agrupamiento: buscan encontrar patrones en los datos mediante la relación de grupos de datos con similitud entre sí.
  - Problemas de detección de anomalías: buscan encontrar datos que no siguen patrones establecidos en el resto del conjunto de datos.

En esta asignatura se abordarán las técnicas del aprendizaje supervisado, en las cuales un modelo se entrena utilizando un conjunto de datos etiquetado. Para ello, se proporciona al algoritmo de aprendizaje automático un conjunto de ejemplos de entrada (características) junto con las salidas deseadas (etiquetas o resultados) correspondientes. El objetivo del aprendizaje automático supervisado es que el modelo aprenda a mapear las entradas a las salidas de manera que pueda hacer predicciones precisas sobre nuevos datos nunca vistos. Además, el aprendizaje se llama supervisado porque se puede supervisar, es decir, evaluar, el comportamiento de este. En este tema se describen las **principales características y diferencias** de estos cuatro grupos de problemas que forman parte del aprendizaje automático supervisado. Además, se describen las diferencias entre los conjuntos de datos de entrenamiento, test y validación cruzada.

1.2. Conceptos clave del aprendizaje supervisado

Antes de seguir avanzando en nuestro conocimiento del aprendizaje automático supervisado, es importante dejar claros algunos conceptos:

- **Conjunto de datos etiquetados:** es el conjunto de datos de ejemplo junto con las salidas deseadas, que se utilizarán como datos de entrada. Se llaman datos etiquetados porque cada registro de datos de entrada está asociado a una salida.
- **Entrenamiento:** es la etapa en la que el modelo de aprendizaje automático utiliza los datos de entrada para aprender la relación existente entre las entradas y las salidas, y así conseguir el aprendizaje.
- **Predicción:** es la etapa siguiente al entrenamiento. Después de que un modelo ha sido entrenado, y por lo tanto ha aprendido a establecer patrones entre datos etiquetados y salidas, puede utilizarse para hacer predicciones sobre nuevos datos de entrada. El modelo aplicará la relación aprendida en el entrenamiento para generar predicciones basadas en los datos de entrada proporcionados.
- **Evaluación del rendimiento:** en todo modelo de aprendizaje supervisado es fundamental medir cómo de bueno ha sido el aprendizaje. Para realizar esta evaluación, se usarán métricas como la precisión, la exactitud, o el rendimiento del modelo en datos que no fueron utilizados durante el entrenamiento. Esto se hace para medir si el modelo ha sido capaz de generalizar el aprendizaje o, por el contrario, simplemente está memorizando el conjunto de datos de entrenamiento.

El aprendizaje supervisado es uno de los paradigmas más comunes dentro del campo del *machine learning*, y se utiliza en un gran abanico de soluciones del mundo real, como, por ejemplo, en los filtros de correo electrónico, en diagnóstico médico, en clasificación de imágenes, sistemas de recomendación, detección de fraude, entre otras.

1.3. Aprendizaje supervisado: problemas de regresión

El aprendizaje automático **surge a mediados de los años 80** con la aplicación de las redes de neuronas y los árboles de decisión. El aprendizaje automático se empezó a utilizar en problemas de predicción complejos donde los modelos estadísticos clásicos no eran muy buenos como, por ejemplo, el reconocimiento de voz e imágenes, la predicción de series temporales no lineales, la predicción de los mercados financieros, el reconocimiento de texto escrito, etc.

Recordemos que el aprendizaje supervisado es un tipo de aprendizaje automático donde se realizan inferencias por medio de una función que establece una correspondencia entre las entradas y la salida del sistema. En el aprendizaje supervisado tenemos datos que son generados por una «caja negra» donde un vector de variables de entrada x (llamadas variables independientes) entran por un lado y, por otro lado, las variables respuesta y son obtenidas.

Figura 1. El aprendizaje automático busca el proceso que relaciona las variables de entrada x con las variables respuesta y. Fuente: elaboración propia.

En el aprendizaje supervisado para cada observación de las variables predictoras x existe una medida de la variable respuesta y. En el caso concreto de los problemas de regresión, la variable respuesta y del sistema que se desea inferir o generalizar es una variable cuantitativa, es decir una variable numérica continua.

Ideas clave

El objetivo del aprendizaje supervisado es **predecir las respuestas que habrá en el futuro** con nuevas variables de entrada. Para ello se utilizan algoritmos y se trata al mecanismo de generación de los datos como algo desconocido. Es decir, se considera el interior de la caja como algo complejo y desconocido. Por tanto, el enfoque es buscar una función $f(x)$ que opere con los datos x para producir las respuestas y.

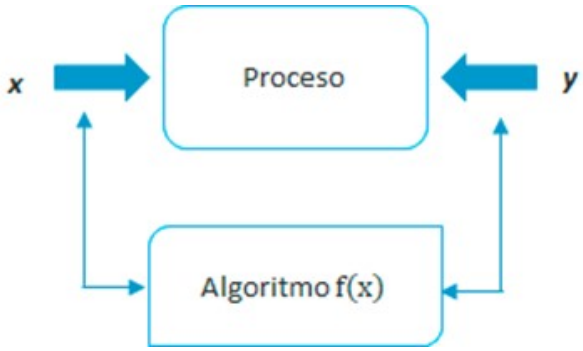


Figura 2. En el aprendizaje supervisado por medio de algoritmos se busca la función $f(x)=y$ que relaciona la entrada con la salida. Fuente: elaboración propia.

La evaluación de este modelo se lleva a cabo por medio de la capacidad predictiva del modelo.

El aprendizaje supervisado tiene como objetivo generalizar las respuestas sobre datos no observados, utilizando para ello ejemplos observados previamente. En el caso de los problemas de regresión, la variable respuesta y es una variable numérica continua.

Algunos ejemplos clásicos de regresión dentro del aprendizaje supervisado son:

- Predecir el precio de una vivienda. Dadas unas características como el número de habitaciones, metros cuadrados, el barrio, etc., predecir el valor de la vivienda.

- Previsión del precio de acciones. Utilizando datos históricos del mercado, predecir el precio de una acción.
- Predicciones de temperaturas. Predecir la temperatura de una región en base a datos históricos observados.
- Predicción de ventas. Prever las ventas futuras de un producto en función de variables como el precio, la temporada y las estrategias de marketing.
- Predicción de Consumo de Energía. Estimar el consumo de energía eléctrica de un hogar o una empresa según el historial de uso y factores ambientales.
- Predicción de Puntuaciones de Crédito. Predecir la puntuación crediticia de un individuo basándose en su historial financiero, ingresos y otros factores relevantes.
- Predicción de Tiempo de Entrega. Prever el tiempo de entrega de un paquete en función de la distancia, el tráfico y otras variables logísticas.

Estos son algunos ejemplos de problemas de regresión comunes en el aprendizaje automático supervisado. Se puede observar que, en todos ellos, el objetivo es predecir un valor numérico continuo a partir de un conjunto de características relevantes. La elección de las características y el algoritmo de regresión concreto dependerá de la naturaleza específica del problema y de los datos disponibles.

1.4. Aprendizaje supervisado: problemas de clasificación

El otro gran grupo de algoritmos de aprendizaje supervisado son los que están enfocados a problemas de clasificación. En los problemas de clasificación, el aprendizaje supervisado utiliza ejemplos conocidos para inferir la etiqueta (clasificar) de los vectores de entrada x eligiendo una de entre varias categorías o clases. En un problema de clasificación, las categorías o clases son las etiquetas que se les intenta asignar los datos. Estos algoritmos, al igual que en los problemas de regresión, utilizan ejemplos etiquetados previamente para aprender los patrones para llevar a cabo una clasificación.

En este tipo de problemas la variable respuesta y es una variable con dos o más categorías. Un ejemplo sería asignar a un correo electrónico la categoría de spam o no spam en función de los correos recibidos previamente. Otro ejemplo es realizar un diagnóstico a un paciente en función de sus características (sexo, presión sanguínea, colesterol, etc.). La etiqueta de clase es, por tanto, una variable discreta. La variable de clase representa las diferentes categorías o clases a las que se asigna cada instancia de datos. Estas categorías son finitas y distintas, lo que significa que son valores discretos. Por ejemplo, en un problema de clasificación binaria, la variable de clase podría tener dos valores discretos, como 0 y 1, donde cada valor representa una clase diferente. En un problema de clasificación multiclase, la variable de clase podría tener varios valores discretos, cada uno correspondiente a una clase específica.

En los **problemas de clasificación**, utilizando aprendizaje supervisado, el objetivo es identificar a qué categoría o clase pertenece una nueva observación, utilizando para ello una serie de observaciones y categorías conocidas previamente.

Ideas clave

Los **problemas de clasificación** se dividen en dos grandes grupos: clasificación binaria y clasificación multiclase. Los problemas de clasificación binaria buscan diferenciar las nuevas observaciones entre una de las dos clases posibles (ejemplo spam y no spam). Los problemas de clasificación multiclase conllevan asignar una nueva observación a una de entre varias clases posibles.

El aprendizaje supervisado tiene como objetivo generalizar utilizando para ello ejemplos conocidos. En el caso de los problemas de clasificación la variable respuesta y es una variable con dos o más categorías o clases.

Algunos ejemplos de problemas de clasificación dentro del aprendizaje supervisado son:

- Detección de spam en correos electrónicos: clasificar correos electrónicos como «spam» o «no spam» para filtrar mensajes no deseados.
- Diagnóstico médico: clasificar imágenes médicas (por ejemplo, radiografías o resonancias magnéticas) para detectar enfermedades o condiciones específicas.
- Clasificación de texto: determinar la categoría o tema de un artículo, comentario o revisión de un producto.
- Reconocimiento facial: identificar y clasificar caras en imágenes o vídeos en función de las personas que aparecen en ellas.
- Clasificación de documentos jurídicos: categorizar documentos legales según su contenido o tema, como contratos, testamentos, demandas, etc.
- Detección de enfermedades en plantas: clasificar imágenes de hojas de plantas para identificar enfermedades o plagas.

Todo esto son problemas típicos de clasificación, donde el objetivo está en asignar una etiqueta de clase a unos datos de entrada, donde las etiquetas de clase están predefinidas y son conocidas.

1.5. Etapas en un proyecto de aprendizaje automático

A lo largo de los siguientes temas el estudiante aprenderá cuáles son los principales algoritmos de aprendizaje supervisado, la implementación y cómo aplicarlos a casos reales. Estos algoritmos van a permitir resolver una amplia variedad de problemas de *machine learning*. Sin embargo, antes de lanzarse a resolver estos problemas ejecutando algunos de los algoritmos que se explicarán, es importante tener en cuenta que esta no es la manera de abordar un problema de aprendizaje automático. El algoritmo solo es una pequeña pieza más de un proceso mucho más amplio de análisis de datos y toma de decisiones. Para poder resolver de forma satisfactoria un problema de aprendizaje automático, es necesario dar un paso atrás y considerar el problema general.

Lo primero será pensar a qué pregunta hay que responder. ¿Se quiere hacer un análisis exploratorio de los datos y simplemente ver si se encuentra algo interesante en ellos? ¿O existe un objetivo particular que se quiera resolver? En el segundo caso, imaginemos que se quieren detectar fraudes en transacciones bancarias, hacer recomendaciones a usuarios, o encontrar planetas desconocidos. Si es ese uno de los objetivos buscados, primero hay que pensar en cómo definir y medir el éxito, y cuál sería el impacto de una solución exitosa. Con todo esto se plantean las siguientes cuestiones:

- ¿Tengo los datos correctos y necesarios para llevar a cabo esta tarea?
- ¿Cómo mido si mi predictor/clasificador está realmente funcionando?
- Si la tarea resulta exitosa, ¿cuál será el impacto?

Ideas clave

Una vez planteadas las cuestiones, si ya está definido el problema que se quiere resolver, se sabe que una solución podría tener un impacto significativo para el proyecto y se dispone de la información para evaluar el éxito. Entonces, los siguientes pasos suelen ser adquirir los datos y construir un prototipo funcional. A lo largo de los siguientes temas se abordarán los conceptos necesarios sobre cómo crear un modelo, cómo emplearlo y cómo evaluarlo.

De una manera más esquemática, podemos dividir un proyecto de aprendizaje automático en las siguientes etapas:

- Definición del problema: identificar el problema que se desea resolver y el objetivo del aprendizaje automático. Además, será necesario definir las métricas para evaluar el éxito del modelo.

- Recopilación de datos: identificar y recopilar los datos necesarios para llevar a cabo la tarea. En ocasiones, esta etapa puede necesitar realizar una búsqueda de fuentes de datos, adquisición de los mismos y realizar un almacenamiento adecuado.

- Tratamiento de los datos: en esta etapa hay que realizar todas estas transformaciones necesarias para tener los datos correctos. Habrá que eliminar datos erróneos, identificar los datos faltantes, normalizar, codificar variables categóricas, etc.

- Exploración de características: analizar y visualizar los datos para comprender mejor sus características y relaciones. En esta etapa se podrán realizar hipótesis sobre los datos, identificar patrones, tendencias y posibles correlaciones que después podrán ser útiles al modelo. Una vez realizado todo esto, se podrá hacer la selección de las características más relevantes para el aprendizaje.

- Elección del modelo: seleccionar el algoritmo de aprendizaje más apropiado en base a los datos disponibles y la tarea que se quiere resolver. Además, habrá que configurar inicialmente los hiperparámetros y la arquitectura del modelo.

Ideas clave

- Entrenamiento del modelo: esta etapa es el núcleo del proceso de aprendizaje automático. En esta etapa, se entrena un modelo para que tome entradas y prediga una salida con el menor error posible. Con modelos más grandes, y especialmente con conjuntos de datos grandes, este paso puede volverse difícil de realizar. Dado que la memoria es generalmente un recurso finito para los cálculos, la distribución eficiente del entrenamiento del modelo es crucial.

- Evaluación del modelo: en esta etapa se utilizarán las métricas apropiadas para determinar el rendimiento del modelo en base al error obtenido o a la precisión.

- Ajuste de hiperparámetros y optimización: en base a los resultados obtenidos en la etapa anterior, habrá que realizar un ajuste de los hiperpárametros del modelo para elegir aquel que obtenga mejores resultados. Esta etapa conducirá a iterar en un proceso de entrenamiento y evaluación hasta conseguir el ajuste óptimo para el modelo.

- Realizar predicciones o desplegar el modelo: una vez realizado el entrenamiento, ajuste y análisis del modelo, está listo para realizar las predicciones.

Como se indicó más arriba estas pueden ser las etapas de un proyecto de aprendizaje automático, pero es importante señalar que todos los componentes del ciclo deben estar orquestados y ejecutados en el orden correcto. Además, no se puede olvidar que el ciclo de modelos de aprendizaje automático es iterativo y puede requerir volver a etapas anteriores para realizar mejoras o cambios a medida que se va adquiriendo mayor conocimiento sobre los datos y el problema.

1.6. Datos de entrenamiento y datos de test: overfitting y evaluación cruzada

La clave de cualquier modelo de aprendizaje automático es su capacidad de generalizar situaciones del futuro en función de los datos históricos observados. Antes de introducir los tipos de problemas que se pueden resolver con aprendizaje supervisado, es necesario establecer algunos conceptos relativos a los datos, muy importantes en los problemas de *machine learning*.

El conjunto de datos, o los datos históricos, es una colección de varios tipos de datos almacenados en un formato digital. Los datos son el componente clave de cualquier proyecto de *machine learning* y consisten principalmente en imágenes, textos, audio, vídeos, puntos de datos numéricos, etc.

Según el informe The State of Data Science (2021), la preparación y la comprensión de los datos es una de las tareas más importantes y que consume más tiempo del ciclo de vida del proyecto de *machine learning*. Dicho informe muestra que la mayoría de los científicos de datos y los desarrolladores de IA pasan casi el 70 % de su tiempo analizando conjuntos de datos. El tiempo restante se dedica a otros procesos, como la selección de modelos, el entrenamiento, las pruebas y el despliegue.

Encontrar un conjunto de datos de calidad es un requisito fundamental para construir la base de cualquier aplicación de IA. Sin embargo, los conjuntos de datos del mundo real son complejos, desordenados y desestructurados. El rendimiento de cualquier modelo de *machine learning* depende de la cantidad, la calidad y la relevancia del conjunto de datos.

Ideas clave

Gracias a las iniciativas Open Data actualmente se dispone de conjuntos de datos de código abierto que han motivado a la comunidad de IA y a los investigadores a realizar investigaciones y trabajos de vanguardia en el campo de la IA. Sin embargo, a pesar de la disposición de los datos, estos no garantizan que sean de la calidad adecuada para algunos problemas. Enumeramos a continuación, los desafíos existentes que limitan a los científicos de datos a la hora de crear aplicaciones basadas en *machine learning*:

- **Datos insuficientes:** falta de disponibilidad de muestras grandes de puntos de datos requeridos por los algoritmos de *machine learning*.
- **Sesgo y error humano:** la mayoría de las herramientas utilizadas para la recopilación de datos conducen a errores humanos o sesgos hacia un aspecto.
- **Calidad:** los conjuntos de datos del mundo real están desorganizados y son complejos. Son de baja calidad casi por defecto.
- **Privacidad y cumplimiento:** la mayoría de las fuentes no comparten sus datos debido a algunas normas de privacidad y cumplimiento. Por ejemplo, médicos, seguridad nacional, etc.
- **Proceso de anotaciones de datos:** por lo general, las intervenciones humanas se utilizan para etiquetar manualmente los conjuntos de datos por su calidad, lo que genera un error. Lleva mucho tiempo y es caro.

A continuación, puedes visualizar la sesión sobre los conceptos fundamentales de *machine learning*.

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=18cb6be2-0522-4f22-b3c1-b17e00e9daac

Datos de entrenamiento y datos test

Tal y como se indicaba en secciones previas, dentro de las etapas de un proyecto de aprendizaje automático existen dos etapas diferenciadas, la etapa del entrenamiento del modelo, y la etapa de predicciones que llamaremos etapa de test. En ambas etapas son necesarios los datos, en la etapa del entrenamiento los datos son los que permiten el aprendizaje de la función f(x) que empareja entradas y salidas, en la etapa de test, los datos son necesarios para comprobar si ese aprendizaje es general o no.

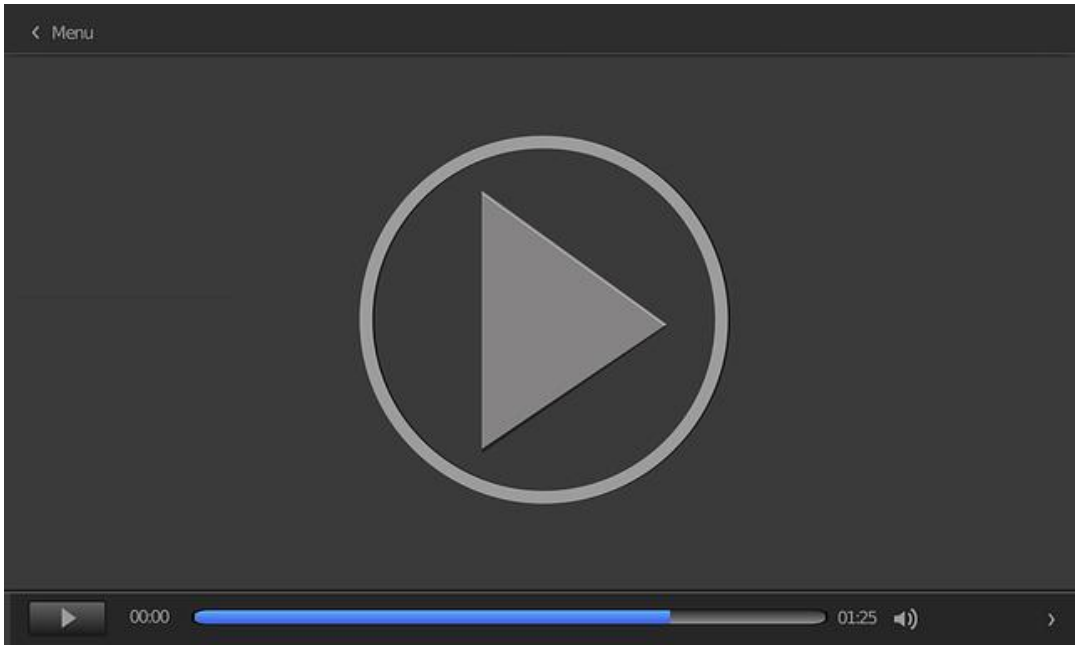
Por lo tanto, en el aprendizaje supervisado, queremos construir un modelo a partir de los datos de entrenamiento y luego poder hacer predicciones precisas sobre datos nuevos que tengan las mismas características que el conjunto de entrenamiento que se ha utilizado. Si un modelo es capaz de hacer predicciones precisas sobre datos nuevos, decimos que es capaz de generalizar el aprendizaje desde el conjunto de entrenamiento al conjunto de prueba. El objetivo del aprendizaje supervisado es construir un modelo que sea capaz de generalizar con la mayor precisión posible. Por ello, para estas etapas se utilizan diferentes conjuntos de datos.

Ideas clave

- **Datos históricos observados:** son todos los registros de información que se tienen almacenados para aprender de ellos.
- **Conjunto de datos de entrenamiento:** es la porción de datos del conjunto de datos históricos utilizada para realizar el aprendizaje.
- **Conjunto de datos de test:** es la porción de datos del conjunto de datos históricos utilizada para evaluar la calidad del aprendizaje.

Comúnmente, construimos un modelo de tal manera que pueda realizar predicciones precisas en el conjunto de datos de entrenamiento. Si los conjuntos de entrenamiento y test tienen suficientes características en común (igual distribución de datos), esperamos que el modelo también sea preciso en el conjunto de test. Sin embargo, hay algunos casos en los que esto puede salir mal. Por ejemplo, si permitimos construir modelos muy complejos, siempre podrán ser tan precisos como queramos en el conjunto de entrenamiento.

En el siguiente vídeo vamos a tratar el término de overfitting.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=b772bdbe-ef12-4a3a-b4e3-b17e00e9da84

En la Figura 3 se describe el proceso de **entrenamiento** y predicción a alto nivel. En la fase de entrenamiento (a) se extraen las variables relevantes de los datos de entrada para construir un modelo por medio de algoritmos de aprendizaje automático. Posteriormente, en la fase de **predicción** (b) se realiza una extracción de variables similar sobre las que se aplica el modelo previamente entrenado para obtener el resultado estimado.

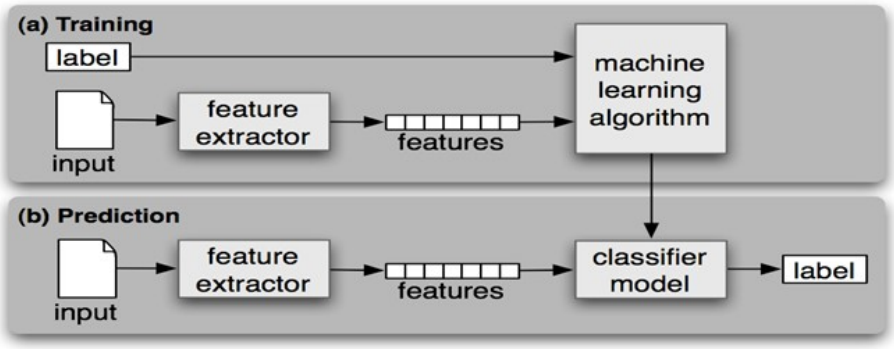


Figura 3. Ejemplo de las fases de entrenamiento (a) y predicción (b) en los procesos de aprendizaje automático. Fuente: NLTK 3.0., 2019.

Durante la fase de construcción de los modelos de aprendizaje automático las diferentes implementaciones *software* proporcionan métricas de error. Estas métricas suelen obtenerse con el conjunto de datos utilizado para realizar el entrenamiento. Este **conjunto de datos** se conoce con el nombre de *training set* o **conjunto de entrenamiento**. Las métricas obtenidas con los datos de entrenamiento deben utilizarse solo como referencia, pues no son buenos indicadores del comportamiento futuro.

Un modelo puede ser capaz de tener un error mínimo con el conjunto de entrenamiento y no ser capaz de predecir bien los valores futuros. Por esta razón, el

error en el conjunto de entrenamiento suele ser un valor muy optimista y debe de ser interpretado con cautela.

Para solucionar el problema anterior, se utiliza un **conjunto de datos de test**. Este conjunto de datos de test puede estar formado con un subconjunto de los datos de entrenamiento.

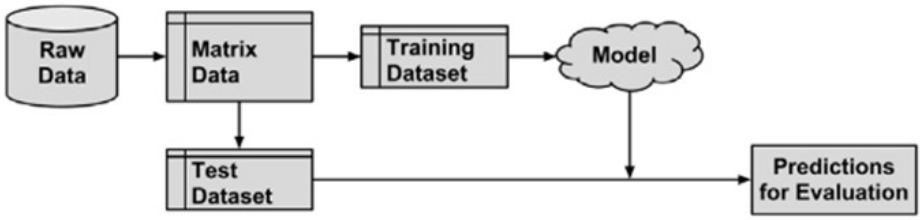


Figura 4. Ejemplo de utilización de un conjunto de test para evaluar el rendimiento de un modelo de aprendizaje supervisado.

El procedimiento de dividir los datos de los que se dispone entre conjuntos de entrenamiento y conjuntos de test se conoce con el nombre de ***hold-out***. En este procedimiento, el conjunto de entrenamiento se utiliza para crear el modelo que es evaluado utilizando el conjunto de test. Normalmente, se utilizan particiones del conjunto de datos históricos de entre un 70 y un 90 % de los datos para crear el conjunto de entrenamiento y el % restante de los datos para generar conjunto de test. Para asegurarse de que no existen diferencias sistemáticas entre los dos grupos, es necesario obtener las instancias de forma aleatoria.

Validación cruzada

La repetición de la técnica de *hold-out* es la base para la técnica de **validación cruzada**. La técnica de validación cruzada es el estándar de la industria para estimar el rendimiento de los modelos. Esta técnica también es conocida como ***k-fold***, donde k es un valor que indica que se han dividido los datos históricos en k particiones separadas llamadas *folds* o conjuntos.

Aunque `k` puede ser cualquier número, lo habitual es utilizar `k=5` o `k=10`. Esto se debe a que la evidencia empírica indica que hay poco beneficio en utilizar más de 10 *folds*. En este caso, para cada uno de los 10 *folds* (que comprenden un 10 % del total de los datos) se crea un modelo en los 9 *folds* restantes (90 % de los datos) y se evalúa en ese 10 %. Este proceso se realza 10 veces, y la media y la desviación típica de las ejecuciones es reportada. En el Tema 11 de la asignatura se entrará en el detalle de esta técnica.

Figura 5. Ejemplo de división de un esquema de validación cruzada con 5 folds. Fuente: Anant, 2020.

El conjunto de entrenamiento en aprendizaje supervisado proporciona un error que debe de ser evaluado con cautela. Es más riguroso evaluar a los modelos utilizando un conjunto separado e independiente llamado **conjunto de test**. Esta separación entre conjunto de test y entrenamiento se puede repetir a lo largo de los datos disponibles utilizando la técnica de **validación cruzada**.

Overfitting

El término *overfitting* o **sobreajuste** hace referencia al fenómeno que se produce en un modelo de aprendizaje supervisado cuando el aprendizaje se ajusta demasiado a las particularidades del conjunto de datos de entrenamiento y obtiene un modelo que funciona bien en el conjunto de entrenamiento, pero no es capaz de generalizarse a nuevos datos. Por otro lado, si su modelo es demasiado simple, y aprende una $f(x)$ muy general, es posible que no pueda capturar todos los aspectos y la variabilidad en los datos, y el modelo funcionará mal incluso en el conjunto de entrenamiento. Elegir un modelo demasiado simple se llama **desajuste** o *underfitting*.

Cuanto más complejo se permita que sea el modelo, mejor podrá predecir sobre los datos de entrenamiento. Sin embargo, si el modelo se vuelve demasiado complejo, entonces comienza a centrarse demasiado en cada punto de datos individual del conjunto de datos de entrenamiento y el modelo no generalizará bien a datos nuevos. Hay un punto óptimo intermedio que producirá el **mejor rendimiento de generalización**. Y ese es el modelo que se quiere encontrar, aquel que permite generalizar el aprendizaje, pero con un alto rendimiento en ambos conjuntos de datos.

1.7. Referencias bibliográficas

Anomnachi, O. (2020). *Spark and Cassandra For Machine Learning: Cross-Validation*. Anant. https://anant.us/blog/modern-business/spark-and-cassandra-for-machine-learning-cross-validation/

Bird, S., Klein, E. y Loper, E. (2019). *Natural Language Processing with Python*. NLTK 3.0. https://www.nltk.org/book/ch06.html

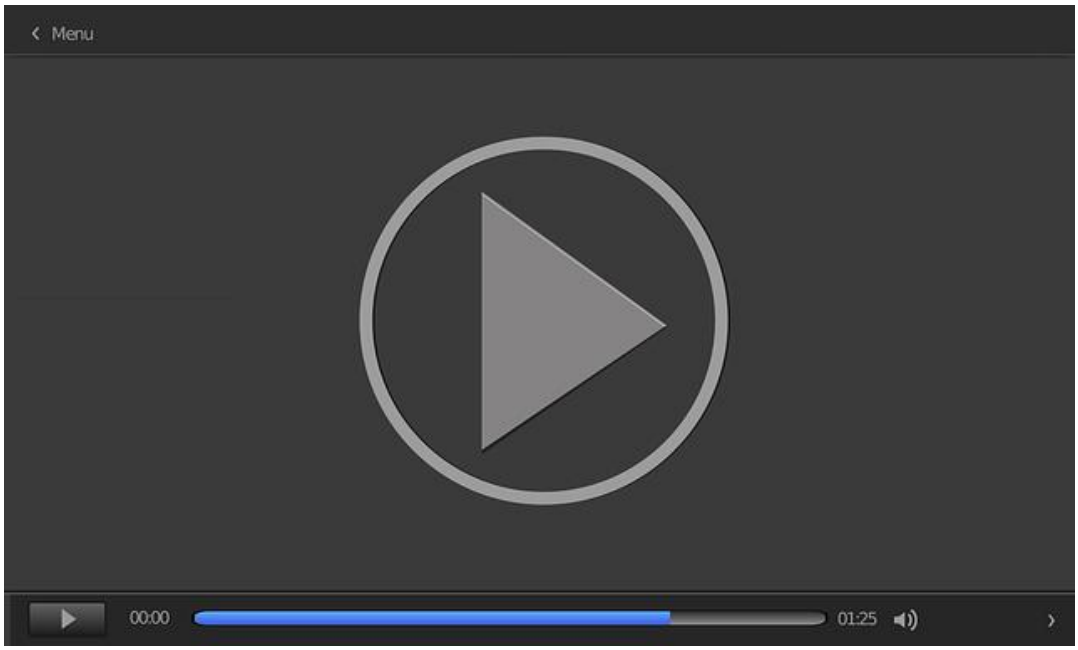
A Few Useful Things to Know about Machine Learning

Domingos, P. (2012). *A Few Useful Things to Know About Machine Learning*. University of Washington. https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Artículo de Pedro Domingos en el que explica conceptos importantes sobre *machine learning*.

Train/Test en Scikit-learn

Udacity. (2015, febrero 23). *Train/Test Split in sklearn - Intro to Machine Learning* [Vídeo]. YouTube. https://www.youtube.com/watch?v=lSwvUmZCvco



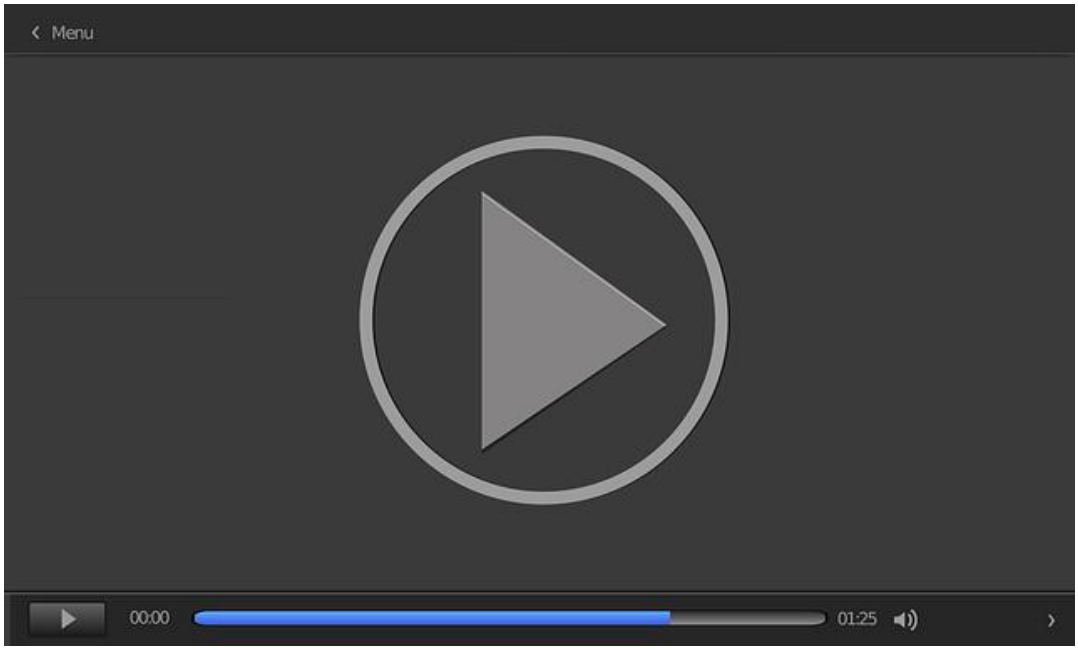
Accede al vídeo:

https://www.youtube.com/embed/lSwvUmZCvco

Vídeo que describe el proceso de generar conjuntos de train y test con Python.

Validación cruzada

Stanford Online. (2020, abril 17). *Lecture 8 - Data Splits, Models & Cross-Validation | Stanford CS229: Machine Learning (Autumn 2018)* [Vídeo]. YouTube. https://www.youtube.com/watch?v=rjbkWSTjHzM



Accede al vídeo:

https://www.youtube.com/embed/rjbkWSTjHzM

Sesión del curso de *machine learning* de Stanford sobre validación cruzada.

Introducción a Machine Learning en Python

Klein, B. (2024). *Machine Learning with Python*. python-course.eu. https://python-course.eu/machine-learning/machine-learning-with-python.php

En esta web se describen los principales conceptos de *machine learning* vistos en este tema y se presentan diferentes ejemplos de aplicación.

Test

1. ¿Cuáles de las siguientes afirmaciones son correctas?

   A. El aprendizaje automático utiliza siempre ejemplos con clases conocidas previamente.

   B. El aprendizaje automático sirve únicamente para resolver problemas de predicción numérica.

   C. El aprendizaje supervisado busca automáticamente los mecanismos que relacionan una entrada con una salida.

   D. B y C son correctas.

2. En el caso de los problemas de regresión:

   A. La variable respuesta que se desea predecir es de tipo cualitativa.

   B. La variable respuesta que se desea predecir es de tipo cuantitativa.

   C. No siempre existe una variable respuesta.

   D. Ninguna de las anteriores es correcta.

3. En los problemas de clasificación:

   A. La variable respuesta contiene siempre más de dos categorías.

   B. La variable respuesta contiene siempre dos o más categorías.

   C. La variable respuesta es de tipo numérico.

   D. Ninguna de las anteriores es correcta.

4. En la fase de entrenamiento de los modelos:

   A. Se realiza la extracción de características y se utiliza para generar posteriormente una predicción.

   B. Se elige qué modelo es el mejor.

   C. Se aprende un modelo que podrá ser utilizado posteriormente.

   D. Ninguna de las anteriores es correcta.

5. En el aprendizaje automático:

A. El conjunto de entrenamiento se utiliza para construir un modelo.

B. El conjunto de test se utiliza para evaluar un modelo.

C. Si un modelo tiene un error mínimo en el conjunto de entrenamiento también lo tendrá en el conjunto de test.

D. Todas las anteriores son correctas.

6. ¿Cuál es la función principal del conjunto de entrenamiento y del conjunto de test en *machine learning*?

A. El conjunto de entrenamiento se utiliza para evaluar el rendimiento del modelo, mientras que el conjunto de test se utiliza para construir y ajustar el modelo.

B. Ambos conjuntos se utilizan para construir y ajustar el modelo durante la fase de entrenamiento.

C. El conjunto de entrenamiento se utiliza para construir y ajustar el modelo, mientras que el conjunto de test se utiliza para evaluar el rendimiento del modelo en datos no vistos.

D. El conjunto de test se utiliza exclusivamente para construir y ajustar el modelo durante la fase de entrenamiento.

7. ¿Cuál es la principal característica del método de *hold-out* en el contexto de validación de modelos en *machine learning*?

A Consiste en dividir los datos en k particiones para realizar múltiples iteraciones de entrenamiento y evaluación.

B. Consiste en separar los datos disponibles en dos conjuntos: uno para entrenamiento y otro para test.

C. Se utiliza para realizar ajustes continuos de los hiperparámetros del modelo durante la fase de entrenamiento.

D. Es exclusivamente aplicable a problemas de clasificación y no a problemas de regresión.

8. El método k-cross validation:

A. Consiste en dividir los datos disponibles k grupos de tamaño variable cada uno de ellos.

B. Consiste en dividir los datos disponibles en k grupos del mismo tamaño.

C. Garantiza la representación equitativa de todas las clases en cada conjunto de validación, lo que puede llevar a una evaluación sesgada en problemas desbalanceados.

D. Ninguna de las anteriores es correcta.

9. ¿Cuál de las siguientes afirmaciones describe correctamente el overfitting en el contexto del aprendizaje automático?

A. El overfitting ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos.

B. El overfitting se refiere a la falta de ajuste del modelo a los datos de entrenamiento, lo que resulta en un rendimiento deficiente.

C. El overfitting es beneficioso ya que permite al modelo adaptarse perfectamente a los datos de entrenamiento.

D. El overfitting solo afecta a modelos lineales y no a modelos no lineales.

10. ¿Cuál de las siguientes afirmaciones describe correctamente una diferencia clave entre regresión y clasificación en el contexto del aprendizaje automático?

A. La regresión se utiliza para predecir valores numéricos, mientras que la clasificación se utiliza para asignar instancias a categorías discretas.

B. Tanto la regresión como la clasificación son métodos intercambiables y se pueden utilizar de manera indistinta para cualquier tipo de problema.

C. En la regresión, la variable de salida es siempre categórica, mientras que en la clasificación puede ser numérica.

D. La regresión se aplica únicamente a problemas de clasificación binaria, mientras que la clasificación puede manejar problemas con más de dos clases.