https://datascience-project.eu/

Ficha de formación

Título	Introducción a Machine Learning
Palabras clave	Machine learning, aprendizaje supervisado, clasificación, regresión, IA, inteligencia artificial, naive bayes, árboles de decisión, bosque aleatorio, redes neuronales, aprendizaje profundo
Idioma	Español
Objetivos / Metas / Resultados de aprendizaje	- **Aprender sobre AI, machine learning, aprendizaje profundo, y cómo se relacionan con la ciencia de datos** - **Aprender sobre diferentes algoritmos utilizados en el aprendizaje automático, incluyendo:** - **Naive bayes** - **Árboles de decisión** - **Bosques aleatorios** - **Redes neuronales** - **Breve descripción de cómo se puede evaluar el rendimiento de los algoritmos de aprendizaje automático**

Curso de formación:	
Alfabetización en ciencia de datos	
Visualización de datos y módulo de analítica de datos	
Introducción a la ciencia de datos para las ciencias humanas y sociales	
Software	
Machine Learning	X
Ciencia de datos para el bien social	
Periodismo de datos y Storytelling	

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Descripción	Esta guía proporciona definiciones de los conceptos fundamentales del aprendizaje automático, así como descripciones de los principales métodos utilizados, incluidos algunos ejemplos y aplicaciones específicos. Puede optar por leer el guión a un nivel superficial, para adquirir una comprensión básica del campo, o leer las descripciones más profundas, en particular la sección de métodos, para obtener una comprensión de nivel intermedio del aprendizaje automático. La estadística y el aprendizaje automático proporcionan las herramientas principales para tu trabajo como científico/a de datos. Comprender los distintos métodos de aprendizaje automático -cómo funcionan, cuáles son sus principales ventajas y cómo evaluar su rendimiento en una tarea determinada- puede ayudarte a tomar mejores decisiones sobre cuándo utilizarlos y te convertirá en un experto en ciencia de datos más versátil.
Contenidos organizados en tres niveles	1. Introducción a Machine Learning La Ciencia de Datos es una disciplina empírica que combina datos con diversos métodos, extraídos principalmente de la Estadística y del Aprendizaje Automático, con el fin de resolver problemas y permitir la toma de decisiones informadas. La Estadística se ha abordado en un curso aparte, por lo que aquí nos centraremos en el campo del Machine Learning (ML). 1.1 Definiciones [BÁSICO] Hay muchas palabras de moda asociadas al ML: las dos más destacadas son Inteligencia Artificial (IA) y Aprendizaje Profundo (AD). La IA es el campo de estudio relacionado con los algoritmos que pueden realizar tareas normalmente asociadas a la "inteligencia" humana. Esto incluye cosas como algoritmos que pueden reconocer imágenes, o que parecen "entender" texto (sí, como chatGPT); que pueden moverse de forma independiente (robots, o coches autoconducidos), o tomar decisiones complejas (como a quién conceder un préstamo, o qué solicitantes de empleo contratar). Si el método para llevar a cabo estas tareas consiste en dar a la máquina instrucciones paso a paso sobre cómo hacerlo, entonces suele denominarse "IA simbólica" o "IA heurística".

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

De hecho, la IA existe desde los años 50 y, hasta que la tecnología informática se hizo más potente y los datos más abundantes (hace unos 15-20 años), la mayor parte de la IA era en realidad IA simbólica. El aumento de los datos disponibles y de la potencia de cálculo ha propiciado la popularidad y la capacidad de una segunda rama de la IA: el ML, el "aprendizaje" mediante el ejemplo. El ML es básicamente el estudio de algoritmos que pueden utilizarse para detectar patrones en los datos. En ML, se dan a la máquina las instrucciones de "cómo encontrar un patrón", así como muchos ejemplos; a partir de estos ejemplos, detecta un patrón y lo utiliza para resolver "nuevos" problemas.

El aprendizaje profundo es un subcampo del ML. Se trata de un conjunto de métodos basados en redes neuronales, que analizaremos más adelante.

1.2 Tipos de machine learning

El aprendizaje automático puede dividirse a su vez en tres clases de algoritmos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

La siguiente figura muestra los distintos tipos de aprendizaje automático y ofrece algunos ejemplos de escenarios de aplicación o casos de uso para cada tipo.

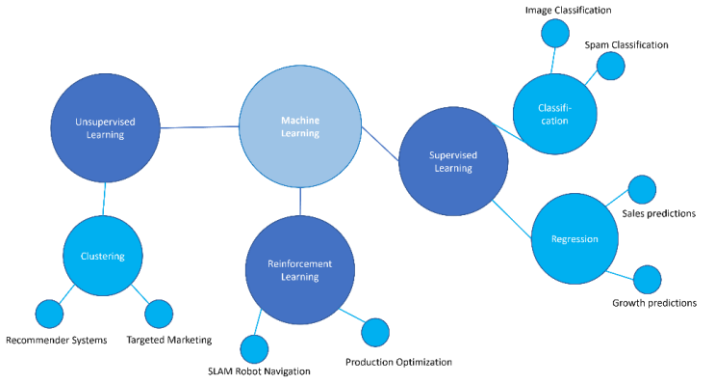


Figura 1: Typos de ML algoritmos

Aprendizaje supervisado

Todos los algoritmos de aprendizaje supervisado requieren datos etiquetados para el entrenamiento, la validación y las pruebas.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Los conjuntos de datos etiquetados son conjuntos de datos que contienen variables de características (también conocidas como variables independientes o atributos) y una variable objetivo (también llamada variable dependiente). Por ejemplo, en un algoritmo de detección de riesgo crediticio, un conjunto de datos etiquetados podría incluir elementos como la edad, el sexo, el saldo de la cuenta, la calificación crediticia y el importe del préstamo solicitado como atributos; y una variable objetivo, como si esta persona ha devuelto su préstamo o no. Otros ejemplos podrían ser un conjunto de datos de imágenes de animales domésticos, con etiquetas que indiquen qué animal se representa; o un conjunto de datos con características como el valor diario de las acciones de una empresa en los últimos 6 meses, una media anual de los últimos 5 años y el número de empleados, y la variable objetivo sería el valor de las acciones de la empresa al día siguiente.

Según el tipo de variable objetivo, el algoritmo de aprendizaje supervisado puede denominarse de clasificación o de regresión. En general, cuando la variable objetivo consta de un número finito de categorías, el algoritmo se denomina algoritmo de clasificación. Si, por el contrario, la variable objetivo es una variable cuantitativa (o numérica), el algoritmo pertenece a la clase de los algoritmos de regresión.

Aprendizaje sin supervisión

El aprendizaje no supervisado se utiliza para detectar patrones en datos no etiquetados. Algunos de los tipos más populares de aprendizaje no supervisado son:

- Agrupación: identificar grupos similares en los datos, sin saber a priori qué grupos buscar.
- Detección de anomalías: determinar qué instancias son "muy diferentes" del resto de ejemplos del conjunto de datos;
- Reducción de dimensiones: reducir la dimensión del espacio de características, lo que incluye métodos como el análisis de componentes principales o LDA.

Aprendizaje por refuerzo (RL)

El aprendizaje por refuerzo se utiliza para derivar una estrategia óptima en situaciones en las que el agente

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

algorítmico debe interactuar con un entorno determinado y tomar una secuencia de decisiones antes de conocer el resultado final (es decir, la retroalimentación no es inmediata: éxito frente a fracaso, ganar frente a perder). Los métodos de RL se utilizan sobre todo en los juegos, la conducción autónoma y la movilidad robótica.

A veces se considera una cuarta clase de algoritmos: el aprendizaje semisupervisado. Se trata de una mezcla entre el aprendizaje supervisado y el no supervisado, y ha ganado popularidad debido a lo costoso que resulta obtener datos etiquetados.

A menudo, la naturaleza del problema en cuestión y el tipo de datos disponibles te ayudarán a decidir qué clase de algoritmo de aprendizaje automático utilizar. ¿Sólo intentas identificar conjuntos de puntos de datos con algún tipo de similitud, sin tener una idea clara de cómo deberían ser estos conjuntos? Entonces lo que necesitas es aprendizaje no supervisado. ¿Tu problema consiste en desarrollar una estrategia óptima en una situación en la que la respuesta (éxito/fracaso) no es inmediata? Entonces buscas una solución de aprendizaje por refuerzo. ¿O tienes un conjunto fijo de categorías y quieres asignar automáticamente nuevos puntos de datos a esas clases predeterminadas? Entonces se trata de aprendizaje supervisado.

Sin embargo, determinar exactamente qué método de aprendizaje supervisado/no supervisado/de refuerzo elegir es un asunto mucho más complicado. El aprendizaje automático es una ciencia empírica y, por lo general, es necesario probar varios algoritmos diferentes y comparar su rendimiento para identificar "el mejor".

Por este motivo, en la siguiente sección describiremos diversas técnicas de ML y sus puntos fuertes y débiles; y en la sección final, estudiaremos cómo evaluar su rendimiento.

2. Visión general de los algoritmos de ML

Esta sección ofrece una visión general de varios algoritmos que se utilizan en el ML. Los algoritmos varían en complejidad, desde algoritmos sencillos, como los árboles de

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

decisión, hasta los más complejos, como los bosques aleatorios.

Esta sección no es en absoluto exhaustiva, pero debería darle una idea de la profundidad y variedad de técnicas disponibles en el aprendizaje automático.

2.1 Estadísticas básicas [BÁSICO]

La regresión lineal es un algoritmo utilizado para problemas de regresión de aprendizaje supervisado. La regresión logística se basa en los conceptos de regresión lineal, pero a pesar de la palabra "regresión" en el nombre, en realidad se utiliza para problemas de clasificación.

De hecho, si analizamos detenidamente muchos conceptos y algoritmos de ML, veremos que a menudo se reducen a variantes de la regresión lineal o logística. Por ejemplo, una neurona de una red neuronal suele ser una simple regresión logística (¡o algo aún más sencillo, como una línea a trozos!).

Aunque también forman parte del conjunto de herramientas de ML, la regresión lineal y logística se han estudiado ampliamente en Statistics, y no se describirán más aquí. Véase el script de STATS.

2.2 Clasificación Naive Bayes [BÁSICO]

Naive Bayes es un algoritmo de clasificación sencillo que suele utilizarse como línea de base (para comparar con otros algoritmos más complejos) en problemas de procesamiento del lenguaje natural, por ejemplo.

Naive Bayes utiliza el Teorema de Bayes para transformar el problema de determinar la probabilidad de que una instancia pertenezca a la clase Y, dados sus atributos X = [x1 , ..., xN], en el problema más sencillo de evaluar la frecuencia del atributo xi , dado que la instancia pertenece a la clase Y.

El teorema de Bayes es una sencilla fórmula matemática utilizada para calcular probabilidades condicionales. El teorema establece que:

$P(Y|X) = \frac{P(X \cap Y)}{P(X)}$, donde

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

P (Y) es la probabilidad de que se produzca el suceso Y ,

P $(X \cap Y)$ es la probabilidad de que ocurran varios sucesos,

P (Y|X) es la probabilidad de que ocurra Y dado que ocurre X (la probabilidad condicional de Y dado X).

Otra forma de escribir el Teorema de Bayes es

P $(X \cap Y)$ = P (X|Y)x P (Y) = P (Y|X)x P (X), y así es como el problema de determinar P (Y|X) puede transformarse en el problema de determinar P (X|Y).

¿Por qué es útil? Porque las frecuencias relativas de X dada Y en los datos de entrenamiento pueden realizarse para determinar P (X|Y).

Pueden dar buenos resultados cuando

- todos los atributos tienen más o menos la misma importancia para determinar la clase de objetivo.
- para una clase objetivo fija, los atributos son independientes entre sí (¿se te ocurre por qué es importante este supuesto?)

Naive Bayes se presenta en diferentes variantes:

- Gaussian NB: se utiliza cuando las variables de atributo son numéricas y se puede suponer que siguen una distribución gaussiana
- Simple NB: se utiliza cuando las variables de atributo son categóricas
- Multinomial NB: más utilizado en contextos de procesamiento del lenguaje natural, donde los atributos son palabras de un documento.

2.3 Árboles de decisión [INTERMEDIO]

Un árbol de decisión es un algoritmo de aprendizaje supervisado que puede utilizarse para modelar la clasificación y la regresión. Los árboles de decisión son tanto una forma de representar la información como un algoritmo para detectar patrones en los datos. De hecho, un algoritmo de árbol de

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

decisión muestra la información que ha "aprendido" de los datos de entrenamiento en forma de árbol de decisión.

¿Qué aspecto tiene un árbol de decisión?

- Los árboles de decisión constan de nodos y ramas, con un nodo en la parte superior
- Cada nodo "formula una pregunta" relacionada con los atributos de los datos, y tiene ramas en función de las posibles respuestas. Por ejemplo, si un atributo es "año en la universidad" y los posibles valores del atributo son (Freshman, Sophomore, Junior, Senior), entonces el nodo correspondiente a "¿qué año en la universidad?" podría tener 4 ramas. Alternativamente, en un árbol de decisión binario, un nodo siempre tendría exactamente dos ramas: por ejemplo, el nodo "¿año en la universidad = Junior?" podría ramificarse primero en "Sí" y "No", y la rama "No" podría tener otro nodo "¿año en la universidad = Freshman?" que se ramificaría en "Sí" y "no", etc.
- Los árboles de decisión se recorren desde el nodo superior hacia abajo: en cada nodo hay que decidir qué rama seguir a continuación, en función del valor o los valores de uno o varios atributos concretos.
- Así hasta llegar a los nodos terminales (u "hoja"). Estos nodos no tienen más ramas y representan la conclusión o predicción.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

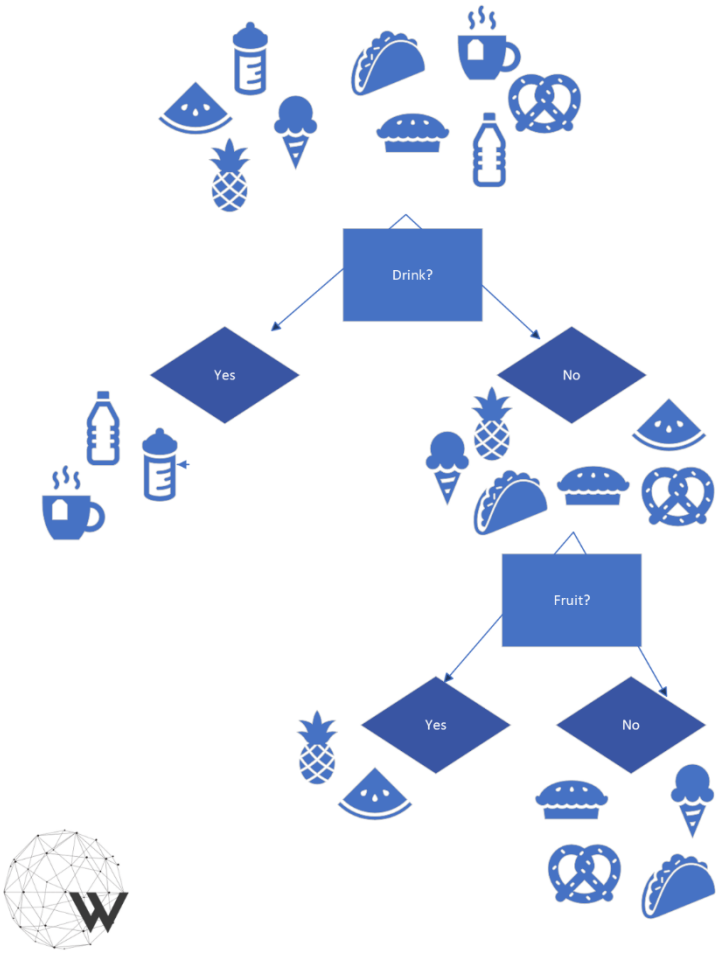


Figure 2: Árbol de clasificación

Un árbol cuyas hojas son clases, o categorías, se denomina Árbol de Clasificación. Cuando las hojas son funciones (la mayoría de las veces constantes numéricas, o bien líneas), se trataría de un Árbol de Regresión.

Los algoritmos de árboles de decisión se construyen utilizando métodos de la teoría de la información, e intentan construir un árbol según el principio de "la mayor información obtenida" en cada paso. Normalmente, el científico de datos debe elegir el número de ramas y la profundidad del árbol, por lo que suele ser necesario experimentar con distintos valores.

También es bueno tener en cuenta que disponer de árboles con un mayor número de ramas y de mayor profundidad

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

proporciona más flexibilidad, pero esto debe sopesarse cuidadosamente frente a las mayores posibilidades de sobreajuste, y al hecho de que los árboles con menos ramas y de menor profundidad son eminentemente más comprensibles.

2.4 Bosques aleatorios (decisión) [INTERMEDIO]

Un bosque aleatorio es una colección de muchos árboles de decisión que funcionan como un conjunto. Los bosques aleatorios son un tipo especial de "aprendizaje por conjuntos", una clase de métodos que combinan modelos (normalmente simples) para mejorar la precisión predictiva a través de la diversidad.

Los bosques aleatorios constan de varios árboles de decisión elegidos al azar y combinan sus predicciones. Varían en el número de árboles que contienen y en la profundidad de cada árbol.

Los bosques aleatorios suelen considerarse una combinación de la capacidad de explicación de los árboles de decisión y la potencia y mayor precisión de métodos más complejos. Los bosques aleatorios y otros métodos de conjunto basados en árboles, como el gradient boosting, siguen siendo bastante populares y pueden lograr resultados de última generación (sí, no siempre tiene que ser una red neuronal).

2.5 Agrupación jerárquica [BÁSICO]

La agrupación es un amplio conjunto de técnicas de aprendizaje no supervisado. El objetivo es detectar estructuras y similitudes en los datos: encontrar una agrupación de los ejemplos del conjunto de datos de forma que los ejemplos de un grupo sean similares entre sí y diferentes de los ejemplos de otros grupos. Una aplicación popular sería la elaboración de perfiles de consumidores: identificar "tipos" de consumidores para poder dirigir mejor los anuncios.

El clustering jerárquico y el clustering K-means son dos de las técnicas de clustering más destacadas. La agrupación

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

jerárquica produce una estructura en forma de árbol (en este caso, suele denominarse dendrograma), que comienza en un nodo superior que contiene todo el conjunto de datos, y recursivamente, en cada nodo, se ramifica en dendrogramas más pequeños, en los que los elementos "similares" van a la misma rama. Este tipo de agrupación ofrece distintos niveles de granularidad: si miramos hacia la parte superior del dendrograma, tenemos un concepto más amplio de "similar", y a medida que avanzamos hacia la parte inferior, las diferencias entre las ramas son más sutiles.

2.6 Agrupación K-Means [BÁSICO]

Mientras que la agrupación jerárquica no requiere ninguna información sobre el número de grupos, o clusters, en los que dividir los datos, el clúster de K-means sí la necesita. De hecho, en el clúster de K-means, el conjunto de datos se divide en grupos distintos K.

A menudo no está claro a priori en cuántos grupos hay que dividir un conjunto de datos. Por este motivo, parte de tu trabajo como científica de datos consistirá en experimentar con distintos valores de K para encontrar el "mejor".

El algoritmo K-means asume que cada instancia del conjunto de datos es un punto en un espacio vectorial con una función de distancia determinada (normalmente euclídea). Comienza asignando aleatoriamente cada instancia del conjunto de datos exactamente a uno de los K clústeres y, a continuación, calcula un centroide, o media, para cada clúster. A continuación, reasigna cada punto al conglomerado cuyo centroide esté más próximo; se vuelven a calcular las medias de los conglomerados y se reasignan de nuevo los puntos. Este proceso continúa hasta que el proceso de reasignación no cambia la pertenencia a un conglomerado de ninguno de los puntos del conjunto de datos.

Una advertencia: los clusters no son robustos y, en particular, las asignaciones aleatorias iniciales de puntos a los clusters influyen mucho en los resultados. Deberías ejecutar el

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

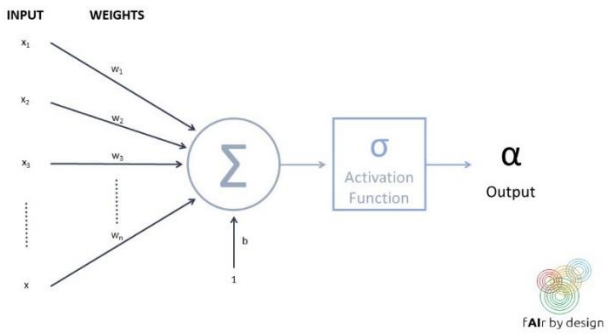
algoritmo K-means varias veces y luego elegir la mejor agrupación.

¿Y cómo es posible determinar cuál es el mejor? Si ya tenemos una noción de distancia, podemos calcular la variación entre los puntos de cada conglomerado. Tomemos la suma de todos los K grupos: Si los grupos tienen sentido y cada uno de ellos contiene puntos similares entre sí, es de esperar que la suma sea pequeña.

2.7 Redes neuronales

Una red neuronal está formada por una serie de unidades interconectadas (las llamadas "neuronas"), como la representada en la siguiente figura.

Cada neurona recibe varias entradas, asigna un peso a cada una y las combina y ejecuta una función de activación para producir una salida. A menudo se utiliza la función sigmoide como función de activación, lo que significa que la neurona actúa como una regresión logística. Pero la función de activación más utilizada actualmente es aún más sencilla: se llama unidad lineal rectificada (ReLU) y toma el valor f(x) = x cuando la entrada x es positiva, y f(x) = 0 cuando x es negativa.



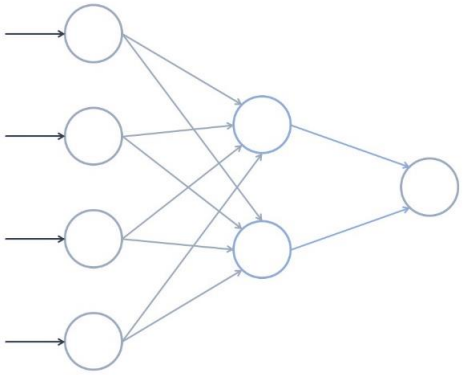
Una red neuronal se forma organizando estas llamadas neuronas en capas.

Entrenar una red neuronal significa intentar establecer los valores de los pesos de la red que minimicen el error de predicción en los datos de entrenamiento (medido por una

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

función de pérdida determinada).



Como puedes ver, los componentes básicos de una red neuronal son bastante sencillos. Lo que las hace tan complejas es el gran número de "neuronas" que tienen, el número de capas y las distintas formas en que las neuronas pueden conectarse entre sí.

3. Evaluación del rendimiento

3.1 Precisión y Co.

Hay muchas métricas que pueden utilizarse para medir el rendimiento de un modelo entrenado. Cuál utilizar depende del tipo de modelo (aprendizaje supervisado, no supervisado o de refuerzo; clasificación frente a regresión) y del contexto de uso. Nos centraremos en el aprendizaje supervisado.

En el aprendizaje supervisado, los conjuntos de datos deben dividirse en conjuntos de entrenamiento, validación y prueba. Los conjuntos de prueba no deben verse nunca en el entrenamiento ni en la validación: deben "guardarse bajo llave" y sacarse sólo al final, para comprobar cómo funciona el modelo con datos completamente nuevos. Sólo si se hace así, y sólo si los datos de prueba son representativos del contexto de uso previsto del modelo, puede considerarse que el rendimiento del modelo en los datos de prueba es una indicación de cómo funcionará "en vivo". Esto también significa que los diferentes contextos de uso requieren diferentes conjuntos de pruebas.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Los datos de validación se utilizan para ayudar a elegir el "mejor" modelo. Por ejemplo, supongamos que tenemos un clasificador de árbol de decisión, en el que intentamos decidir cuál es la mejor "profundidad", y también queremos compararlo con un clasificador Naive Bayes: utilicemos el rendimiento en el conjunto de datos de validación para hacer la comparación. Conviene repetir una cuestión importante: si un conjunto de datos se ha utilizado para la validación, no puede utilizarse como conjunto de prueba. Sin embargo, teniendo en cuenta este principio, puede utilizar los datos de validación para más de una validación o comparación de modelos.

Por último, el conjunto de datos de entrenamiento es el que se utiliza para entrenar el modelo. Lo ideal sería que los datos de validación estuvieran completamente separados de los datos de entrenamiento. Sin embargo, en circunstancias en las que los datos son escasos, es posible utilizar el bootstrapping o la validación cruzada (véase más adelante) para utilizar el conjunto de datos de entrenamiento tanto para el entrenamiento como para la validación del modelo.

Una vez establecido un conjunto de prueba o validación, también necesitamos saber cómo medir el rendimiento del modelo. Recordemos que, para un algoritmo supervisado, todos los ejemplos del conjunto de datos tienen el valor objetivo "correcto", que puede compararse con el valor predicho por el modelo.

- La métrica de rendimiento más utilizada para los modelos de regresión es el MSE. Se calcula el error cuadrático medio entre el valor objetivo real y la predicción del modelo. Esto debería haber sido cubierto en su curso de estadística, y no se discutirá aquí.
- La métrica de rendimiento más utilizada para la clasificación es la precisión: se trata simplemente del número total de clasificaciones correctas sobre el número total de instancias del conjunto de datos.

Co-funded by the Erasmus+ Programme of the European Union

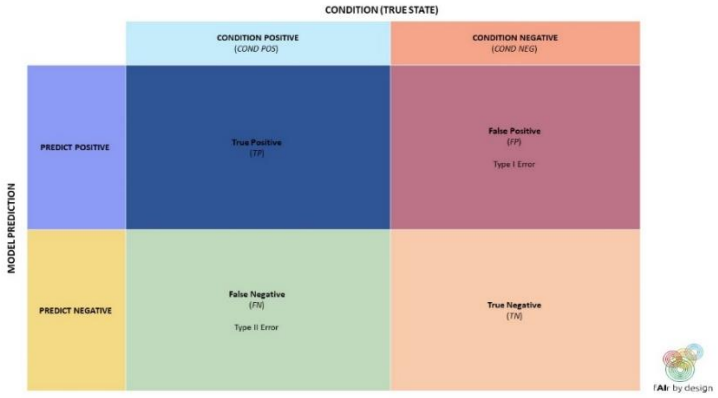
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Sin embargo, no siempre son las "mejores" métricas, como mostrarán los ejemplos siguientes.

Los clasificadores binarios son sistemas de clasificación en los que sólo hay dos clases posibles: llamémoslas POSITIVA y NEGATIVA.

Examinaremos distintas métricas de rendimiento para ellos y por qué, en determinadas circunstancias, son preferibles a la precisión.

Empecemos con una herramienta de uso común para ayudar a comprender el rendimiento de un clasificador binario: la matriz de confusión.



Utilizando la terminología de la matriz de confusión, podemos escribir una fórmula para la precisión:

Precisión = (TP + TN)/(TP + TN + FP + FN)

¿Cuándo querría utilizar una métrica distinta de la precisión?

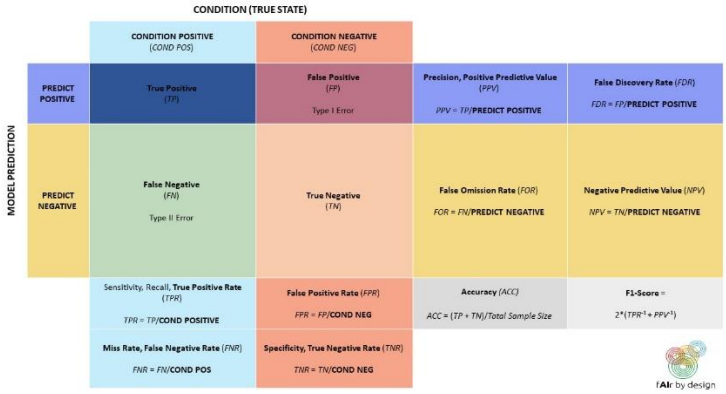
- Cuando las clases objetivo de su conjunto de pruebas están muy desequilibradas: por ejemplo, si el 95% son POSITIVAS y sólo el 5% son NEGATIVAS, entonces un clasificador que simplemente clasificara todo como POSITIVO tendría una asombrosa precisión del 95%. Pero, ¿sería útil?
- ¿Es más importante identificar correctamente todos los elementos POSITIVOS (por ejemplo, en un diagnóstico médico, se quiere estar seguro de detectar la presencia de una enfermedad, para poder iniciar el

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

tratamiento)? ¿O es más importante evitar falsos POSITIVOS?

Una versión más ampliada de la matriz de confusión, que se muestra a continuación, puede ayudar en la elección de la métrica:



Así, si necesitas identificar todos los elementos POSITIVOS, entonces tu modelo debe tener una *alta sensibilidad, o Tasa de Verdaderos Positivos (TPR).* Si, por el contrario, quieres evitar falsos POSITIVOS, entonces tu modelo debe minimizar *la Tasa de Falsos Positivos (FPR)* - lo que, examinando la matriz de confusión, equivale a maximizar la *especificidad*, o la *Tasa de Verdaderos Negativos*.

Incluso cuando esté claro que se tiene la métrica (o métricas) correcta (se puede intentar optimizar más de una o encontrar un equilibrio entre varias), ¿cuál es el punto en el que se dice "esto es suficientemente bueno" y se decide utilizar el modelo? No hay un manual que responda a esta pregunta: depende del contexto.

A modo de ejemplo, consideremos una aplicación de la "vida real": la detección automática de discursos de odio en las redes sociales.

Según los datos obtenidos en el proyecto "Barometro dell'Odio" de Amnistía Internacional Italia (véanse las diapositivas de data4good), la incitación al odio representa alrededor del 1% de los contenidos políticos en línea. Dado

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

que la clase objetivo está tan desequilibrada, la precisión no es la mejor métrica. Supongamos que se desarrolla un modelo de incitación al odio optimizado para un alto TPR y un bajo FPR: alcanza un 99% de TPR y un 1% de FPR.

Entonces, por cada 100 comentarios que el modelo clasifica como incitación al odio, ¿cuántos se puede esperar que sean en realidad comentarios neutrales? Intenta averiguarlo tú misma antes de leer la hoja de Excel.

Real World Data

Total number of comments per day: 1000000

Prevalence: 1 %

(in the context of deployment of the automated hate speech detector, what is the percent of hate speech comments out of all comments published per day)

Test Set Data

True positive rate: 99 %

(percent of correctly identified hate speech comments out of all hate speech comments in test set)

False positive rate: 1 %

(percent of comments incorrectly identified as hate speech out of all non-hate speech comments in test set)

	Predicted non-hate speech	Predicted hate speech	Row totals
True not hate speech	980100	9900	990000
True hate speech	100	9900	10000
Column Totals	980200	19800	1000000

What is the chance that a comment that is predicted to be hate speech, actually _is_ hate speech???
50 %

What is the chance that a comment that is predicted to be non-hate speech, actually _is_ hate speech???
0.01 %

Puedes utilizar la hoja de cálculo anterior para jugar con diferentes TPR, FPR y prevalencias. Esto debería darte una idea de la importancia, no sólo de las métricas basadas en tu conjunto de pruebas, sino también para tratar de entender el impacto del modelo en su contexto de uso.

Por ejemplo: sabiendo lo que sabes sobre el porcentaje de comentarios neutrales que marcaría como incitación al odio, ¿recomendarías que el modelo se utilizara para clasificar y censurar automáticamente los comentarios de incitación al odio?

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.2 Bootstrapping

El Bootstrapping se basa en hacer un Muestreo Aleatorio con Reemplazo (es decir, se coge el jarrón con bolas de colores, tan común en los textos de probabilidad, que se saca al azar una bola, se anota el color y se vuelve a tirar la bola al jarrón) sobre los datos de entrenamiento. Esto significa que una misma observación puede extraerse varias veces, mientras que otras observaciones pueden no extraerse en absoluto.

Este hecho estadístico se aprovecha: se extraen muestras de los datos de entrenamiento tantas veces como sea necesario hasta obtener un nuevo conjunto de datos de entrenamiento del mismo tamaño. Las observaciones que nunca se han extraído en este procedimiento se incluyen en el conjunto de datos de validación. Los resultados de la validación se utilizan para comparar los distintos algoritmos.

3.3 Validación cruzada

Hay diferentes formas de realizar la validación cruzada, pero nosotros nos centramos en la validación cruzada n-fold, y set n = 5 para simplificar.

El conjunto de datos de entrenamiento se divide, por muestreo aleatorio, en 5 subgrupos de tamaño aproximadamente igual.

- En la primera pasada, se toma el grupo de datos 1 como datos de validación, y se entrega con el resto de datos (grupos 2,3,4,5).
- En la segunda pasada, el segundo grupo de datos se reserva para la validación, y el algoritmo se entrena en los otros grupos de datos (1,3,4,5).
- Continúa así hasta que los 5 grupos de datos hayan servido como datos de validación exactamente una vez.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

- Así se obtienen 5 resultados de validación (por ejemplo, la tasa de error para la clasificación o el MSE para la regresión)
- Una vez completada la validación y seleccionado el "mejor" modelo, se puede volver a entrenar con el conjunto de datos completo.

3.4 Otras consideraciones

Hay situaciones en las que estas medidas de rendimiento no son suficientes. Consideremos el siguiente ejemplo, en el que un clasificador de imágenes ha detectado un patrón y puede clasificar las imágenes como "perro" frente a "lobo".

"perro" *"lobo"*

"dog"

¿Cómo crees que clasificará estas dos próximas imágenes?

La imagen de la izquierda se clasificó como "perro". La de la derecha como "lobo".

¿Por qué? Porque en realidad el modelo no detectaba perro contra lobo, sino nieve contra ausencia de nieve.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Este ejemplo se inspira en el artículo "¿Por qué debería confiar en ti?". [1]. Mientras el modelo sea demasiado complejo para que podamos entender qué patrones ha aprendido y por qué se ha hecho una predicción concreta, nos resultará difícil detectar errores. Hay situaciones en las que puede ser mucho más importante poder entender qué patrones ha aprendido el modelo, que ganar unos puntos porcentuales extra en precisión. Más allá de la explicabilidad, otros posibles requisitos del modelo podrían ser la seguridad (contra piratas informáticos o envenenadores de datos, por ejemplo), la privacidad (si el algoritmo tiene que procesar datos sensibles) o la no discriminación (véanse las diapositivas sobre datos 4 buenos). Hay muchos criterios que se combinan para crear el "mejor" modelo: la precisión puede ser sólo uno de ellos. 3.5 Lecturas adicionales Este script sólo te ha servido para empezar tu viaje por el ML. Si tienes curiosidad por aprender más y probar algunos problemas, te recomendamos encarecidamente el libro de texto "An Introduction to Statistical Learning" [2].
Test de autoevaluación (preguntas y respuestas de elección múltiple)	1. ¿Cuál de las siguientes opciones describe mejor el enfoque de aprendizaje automático para realizar una tarea? A) Seguir instrucciones paso a paso B) Detectar un patrón a partir de datos históricos o ensayos anteriores y aplicarlo. C) Seguir probando al azar hasta tener éxito 2. ¿Qué algoritmo asume que las características de entrada son mutuamente independientes? A) Redes neuronales B) Árbolo de decisiones C) Naive Bayes 3. ¿En cuántas participaciones se divide el conjunto de datos de entrenamiento de un CV en 5-fold? ¿Y cuántas veces se entrena el algoritmo en total? A) 5 participaciones; 5 entrenamientos

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	B) 4 participaciones; 5 entrenamientos C) 5 participaciones; 6 entrenamientos
Recursos (videos, enlaces de referencia)	▪ https://medium.com/@lyon-nlp/labeling-tools-for-nlp-36a8179f15d8 ▪ https://towardsdatascience.com/introduction-to-machine-learning-with-graphs-f3e73c38d4f8 ▪ James, G. et al, *An Introduction to Statistical Learning,* 2nd ed., 2021. Available at https://www.statlearning.com/
Material relacionado	• Photo by Colin Davis on Unsplash: red setter • Photo by Ashlee Marie on Unsplash: dog in meadow • Photo by Oscar Sutton on Unsplash: labrador • Photo by ractapopoulos on Pixabay: wolf howling • Photo by StormmillaGirl on Pixabay: wolf in snowy Woods • Photo by Leila LaRochelle on pexels: wolf in snow from above • Photo by Steve on pexels: brown wolf, no snow • Photo by Maurizio Izzo on Pixabay: German shepherd in snow
PPT relacionado	1. Estadísticas 2. Datos para el bien social
Bibliografía	[1] Ribeiro, M. et al, "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Available on arxiv: https://arxiv.org/abs/1602.04938 [2] James, G. et al, *An Introduction to Statistical Learning,* 2nd ed., 2021. Available at https://www.statlearning.com/
Proporcionado por	[Women in AI Austria]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.