Inteligencia Artificial para profesionales TIC

ML: Entrenamiento y predicción

valgrai.eu

Índice

1. Entrenamiento y predicción

Bienvenido a la IA discriminativa o predictiva

Entrenamiento y predicción

En el proceso de entrenamiento, se consideran **un conjunto de características relevantes de los datos** de entrada para construir el modelo utilizando cualquiera de los diferentes algoritmos de aprendizaje automático

En el proceso de predicción, de nuevo se considera un conjunto de las mismas características de los nuevos datos y se utiliza el modelo, utilizando el modelo entrenado anterior, para **predecir** el resultado estimado

El primer proceso es el **entrenamiento** con el conjunto de datos disponible, con el fin de crear el modelo

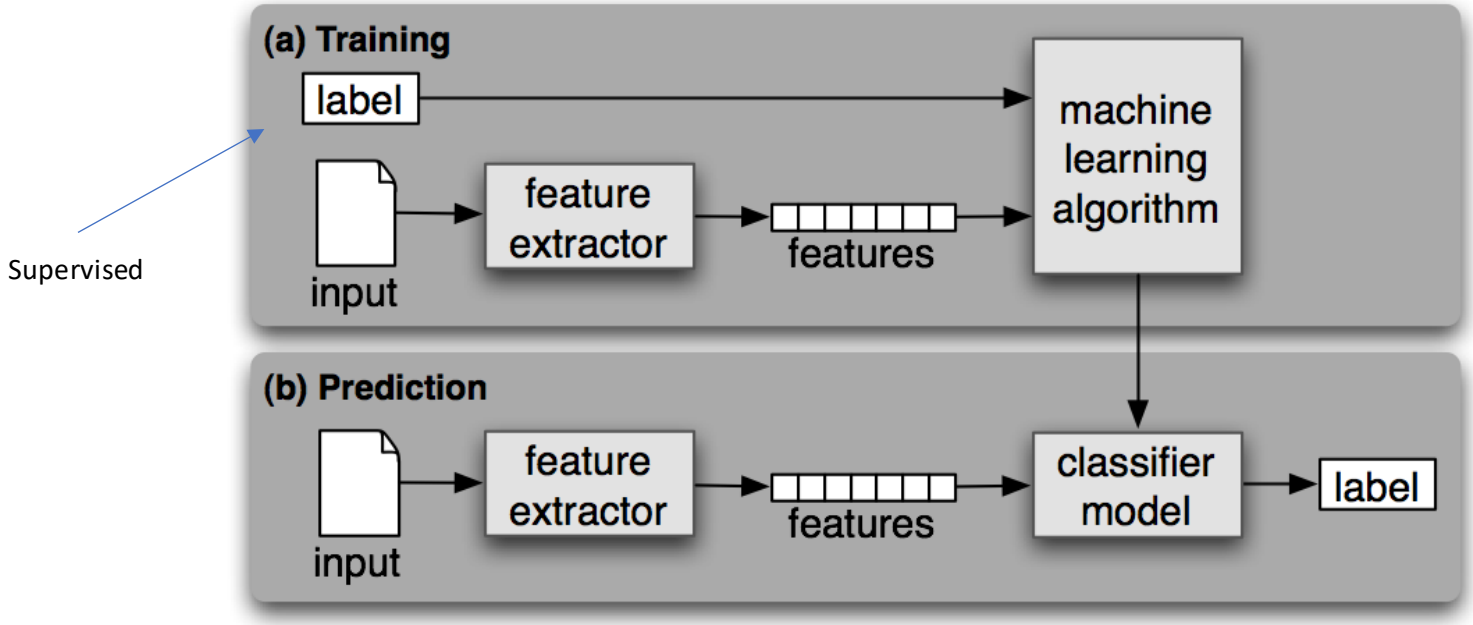
Una vez que tenemos el modelo, el objetivo es hacer **predicciones** con nuevos datos

Nuestro objetivo: nuestro modelo debe ser bueno para generalizar, es decir, ser bueno para predecir frente a lo no visto o los ejemplos utilizados en la creación del modelo

2. Proceso principal

Enfocado en el aprendizaje supervisado

Estamos tratando de predecir situaciones futuras a partir de datos observados (aprendizaje supervisado)



http://www.nltk.org/book/ch06.html

3. Métricas de error

Solo una primera introducción

Las **métricas de error** o cómo podemos medir objetivamente la calidad de nuestro modelo es fundamental, está en el núcleo de cualquier proceso de ingeniería.

Aquí solo una introducción, lo cubriremos en tema específico

¿Qué tan bueno es nuestro modelo? Esta es una cuestión fundamental en la que tenemos que centrarnos desde el principio de la creación de nuestro modelo

Estas métricas, como veremos, son diferentes en el caso del aprendizaje supervisado y no supervisado, y son esenciales antes de desplegar nuestro modelo

En el caso del **aprendizaje supervisado**, tenemos lo que llamamos el conjunto de entrenamiento: una colección de ejemplos (vectores de características) con categorías o valores numéricos bien conocidos

Podemos usar métricas para ver qué tan bien funciona nuestro modelo con los datos de entrenamiento, pero como veremos, es una muy mala idea considerarlas solo ya que nuestro modelo podría haber aprendido los datos de entrenamiento y ser muy malo en la 'generalización' o predicción de resultados para nuevas entradas (lo llamaremos sobreajuste)

Por lo tanto, una buena práctica es usar un **conjunto de prueba (test/validación).** Esto significa que, de alguna manera, no usamos toda nuestra información disponible para entrenar el modelo, sino solo una parte de ella, y luego medimos la precisión de nuestro modelo con este conjunto de datos

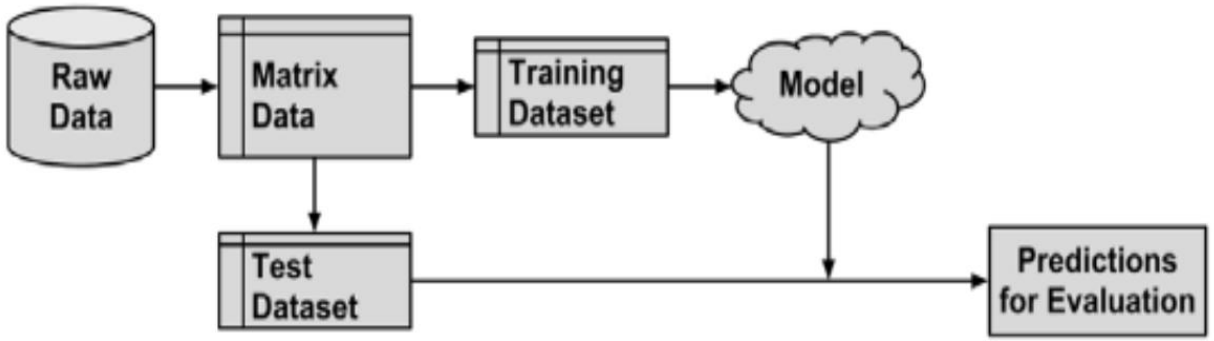
3. Métricas de error

Solo una primera introducción

Nuestro modelo tiene que aprender de los ejemplos, pero ¿es bueno memorizarlos?

Si el modelo memoriza los ejemplos de entrada, ¿será bueno para **generalizar**? ¿Dará buenas predicciones frente a nuevos ejemplos? La respuesta es no.

Generalizar significa lograr que nuestro modelo aprenda la "**esencia**" del problema a partir de los ejemplos, los patrones subyacentes, pero hasta cierto punto.



3. Métricas de error

Ajuste insuficiente, sobreajuste, nuestros enemigos

Underfitting se define cuando el modelo no es lo suficientemente potente como para aprender del conjunto de datos de entrada.

Overfitting aparece cuando el modelo es realmente bueno para los datos de entrenamiento, pero no es capaz de predecir datos futuros.

Necesitamos entonces una métrica basada en datos de prueba (no datos de entrenamiento), en un intento de evaluar cómo podría funcionar nuestro modelo a la hora de predecir nuevas entradas.

A esto se le llama generalmente **hold-out** (Por ejemplo, dividimos nuestros datos de entrada en 80-20)

Incluso con esta división, no podemos estar realmente en lo correcto estadísticamente hablando

Por lo tanto, utilizamos una técnica llamada validación cruzada o **cross-validation** (una repetición sistemática de la técnica de exclusión) que es el estándar a la hora de estimar el error de los métodos

3. Métricas de error

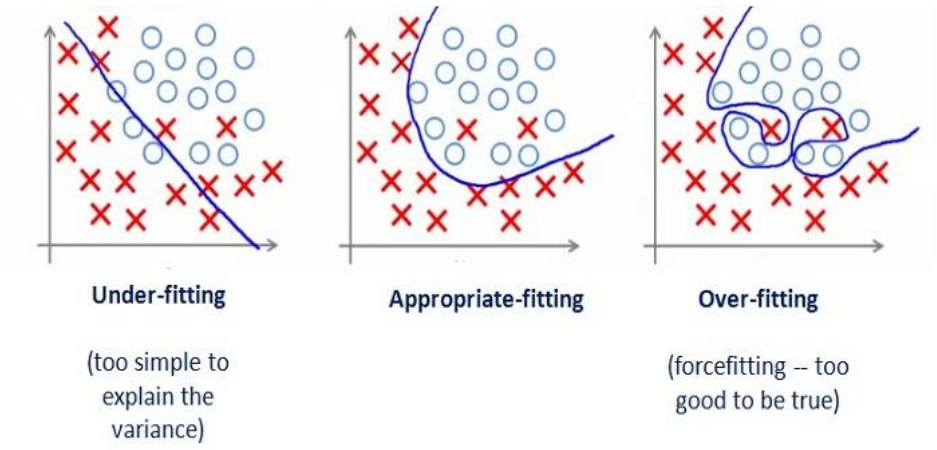
Ajuste insuficiente, sobreajuste, nuestros enemigos

Las **métricas** son extremadamente importantes y debemos medirlas adecuadamente.

Sin embargo, nuestro modelo debe ser capaz de proporcionar una "generalización", lo que significa buenos resultados frente a nuevos datos "invisibles"

Es por eso que introducimos los factores relacionados con la capacidad del modelo para ser realmente 'general'':

- **Overfitting**
- **Underfitting**



3. Métricas de error

Tenga cuidado con el modelo en producción

Un error común, realmente muy común, es hacer un buen trabajo en el entrenamiento y luego no prestar mucha atención al proceso de predicción.

En la predicción, cuando el modelo está en producción:

- Asegúrese de realizar exactamente el mismo proceso con las nuevas entradas en términos de transformación de datos
- Utilice los mismos criterios para obtener los valores de las características de los valores de entrada que al entrenar el modelo
- Asegúrese de que los nuevos datos de entrada para obtener predicciones pertenezcan a la misma distribución de la población objetivo
- Realice una evaluación de métricas durante la producción para ver si las métricas son las mismas

No considerar estos aspectos llevan a **model degradation**.

4. Enfoque centrado en los datos

Sesgo y calidad de los datos

La **IA centrada en los datos** se basa en la idea de que, actualmente, los modelos ya no son el problema. Tenemos muchas alternativas y herramientas muy interesantes para encontrar sus hiperparámetros y sacar lo mejor de ellos

El punto principal aquí en este momento son los **DATOS** y, principalmente, su **calidad**

Uno de los problemas más comunes que tienen muchos modelos de ML cuando se ponen en producción es lo que llamamos "**sesgo**"

Sesgo significa una tendencia inapropiada a proporcionar predicciones que no son realmente "objetivas".

Las técnicas de ML no suelen ser las principales responsables de este hecho; Los datos utilizados para la formación son los principales culpables

El uso de datos de calidad es fundamental para evitar resultados considerados no éticos.

La forma en que se recopilan los datos es esencial (evitar el **sesgo de muestreo**, es decir, **el conjunto de datos debe tener la misma distribución de la población objetivo**), y un paso preliminar importante antes de entrenar cualquier modelo

La cantidad de datos puede ser decisiva para algunos métodos de ML (por ejemplo, el aprendizaje profundo)

El **equilibrio de datos** también es un factor crítico: los modelos de ML tienden a predecir más valores en categorías en las que hay más datos disponibles en el proceso de entrenamiento. Equilibrar los datos va a ser esencial.

Otros elementos a tener en cuenta es el tratamiento de los valores faltantes (en características o 'columnas'), tener que escalar características, codificación de características (de categoría a numérica), etc.

Valencian Graduate School and Research Network of Artificial Intelligence

valgrai.eu

valgrai.eu

Valencian Graduate School and Research Network of Artificial Intelligence