Fases para abordar un problema con machine learning

Generalmente hay tres fases para abordar un problema con machine learning:

Fase 1 - Fase de entrenamiento: Esta es la fase donde los datos de entrenamiento se usan para entrenar el modelo emparejando la entrada dada con la salida esperada. El resultado de esta fase es el propio modelo de aprendizaje.

Fase 2: Fase de validación y prueba: Esta fase consiste en medir cuánto de bueno es el modelo de aprendizaje que ha sido entrenado y estimar las propiedades del modelo, tales como medidas de error, memoria, precisión y otros. Esta fase utiliza un conjunto de datos de validación, y la salida es un modelo de aprendizaje sofisticado.

Fase 3 - Fase de aplicación: En esta fase, el modelo está sujeto a los datos del mundo real para los cuales los resultados deben derivarse.

Pasos para construir un modelo de machine learning

Construir un modelo de Machine Learning , no se reduce solo a utilizar un algoritmo de aprendizaje o utilizar una librería de Machine Learning ; sino que es todo un proceso que suele involucrar los siguientes pasos:

1. **Recolectar los datos.** Podemos recolectar los datos desde muchas fuentes, podemos por ejemplo extraer los datos de un sitio web o obtener los datos utilizando una API o desde una base de datos. Podemos también utilizar otros dispositivos que recolectan los datos por nosotros; o utilizar datos que son de dominio público. El número de opciones que tenemos para recolectar datos no tiene fin!. Este paso parece obvio, pero es uno de los que más complicaciones trae y más tiempo consume.
2. **Preprocesar los datos**. Una vez que tenemos los datos, tenemos que asegurarnos que tiene el formato correcto para nutrir nuestro algoritmo de aprendizaje. Es prácticamente inevitable tener que realizar varias tareas de preprocesamiento antes de poder utilizar los datos. Igualmente este punto suele ser mucho más sencillo que el paso anterior.
3. **Explorar los datos.** Una vez que ya tenemos los datos y están con el formato correcto, podemos realizar un pre análisis para corregir los casos de valores faltantes o intentar encontrar a simple vista algún patrón en los mismos que nos facilite la construcción del modelo. En esta etapa suelen ser de mucha utilidad las medidas estadísticas y los gráficos en 2 y 3 dimensiones para tener una idea visual de cómo se comportan nuestros datos. En este punto podemos detectar valores atípicos que debamos descartar; o encontrar las características que más influencia tienen para realizar una predicción.
4. **Entrenar el algoritmo .** Aquí es donde comenzamos a utilizar las técnicas de Machine Learning realmente. En esta etapa nutrimos al o los algoritmos de aprendizaje con los datos que venimos procesando en las etapas anteriores. La idea es que los algoritmos puedan extraer información útil de los datos que le pasamos para luego poder hacer predicciones.
5. **Evaluar el algoritmo .** En esta etapa ponemos a prueba la información o conocimiento que el algoritmo obtuvo del entrenamiento del paso anterior. Evaluamos qué tan preciso es el algoritmo en sus predicciones y si no estamos muy conforme con su rendimiento, podemos volver a la etapa anterior y continuar entrenando el algoritmo cambiando algunos parámetros hasta lograr un rendimiento aceptable.
6. **Utilizar el modelo.** En esta última etapa, ya ponemos a nuestro modelo a enfrentarse al problema real. Aquí también podemos medir su rendimiento, lo que tal vez nos obligue a revisar todos los pasos anteriores. El modelo determina cómo los datos de entrada para cada solicitante se pueden utilizar para predecir mejor el resultado del préstamo. Al encontrar y usar patrones en el conjunto de entrenamiento, ML produce un modelo (que se puede pensar en esto como una caja negra) que produce una predicción del resultado para cada nuevo solicitante, basado en los datos de ese solicitante.

Conjunto de datos

Se distinguen dos tipos, el conjunto de entrenamiento y el conjunto de prueba. Para obtener estos, dividimos los datos de muestra en dos partes; una parte se utiliza como conjunto de entrenamiento para determinar los parámetros del clasificador y la otra parte, llamada conjunto de prueba (test o conjunto de generalización) se utiliza para estimar el error de generalización ya que el objetivo final es que el clasificador consiga un error de generalización pequeño evitando el sobreajuste (sobre-entrenamiento), que consiste en una sobrevaloración de la capacidad predictiva de los modelos obtenidos: en esencia, no tiene sentido evaluar la calidad del modelo sobre los datos que han servido para construirlo ya que esta práctica nos lleva a ser demasiado optimistas acerca de su calidad.

El conjunto de entrenamiento suele a su vez dividirse en conjuntos de entrenamiento (propiamente dicho) y conjunto de validación para ajustar el modelo

Se suelen utilizar el 80 % de los datos para entrenar a la máquina, el 20 % como conjunto de validación.

Evaluación de modelos

Una buena práctica consiste en dividir el conjunto de datos etiquetados en 2 subconjuntos ,uno de entrenamiento y otro de test.

Normalmente se debe decidir el tamaño de cada uno de los conjuntos. Aunque no existe una regla de oro para decidir cuántas instancias asignamos a cada uno de ellos,es frecuente ver problemas donde un 80% corresponde a datos de entrenamiento y un 20% corresponde a datos de test

- **el conjunto de entrenamiento** se emplea para aprender un modelo de clasificación o regresión
- **el conjunto de test** se emplea para predecir el valor de una determinada instancia de nuestro modelo