

Introducción a Machine Learning

Sebastián Arpón, PhD
Físico y Data Scientist
MetricArts



METRICARTS



github.com/sarpon



www.linkedin.com/in/sarpon



Agenda

1. Aprendizaje de maquina.
2. Aprendizaje supervisado y no supervisado.
3. La dinámica del aprendizaje de maquina.
 1. Definiciones y terminología.
 2. Entendimiento del Problema.
 3. Preparación de datos.
 4. Conjunto de Entrenamiento, Conjunto de Test.
 1. K-fold Cross Validation.
 5. Selección de algoritmo de aprendizaje.
 6. Validación.
4. Como comparar modelos.
5. Como mejorar modelos.
6. Modelo Regresión Logística.
7. k-means.



Aprendizaje de máquina

¿Qué es el Aprendizaje de Maquina?

“Machine Learning as a field of study arose as a subfield of Artificial Intelligence, which was concerned with methods for improving the knowledge or performance of an intelligent agent over time, in response to the agent’s experience in the world. Such improvement often involves analyzing data from the environment and making predictions about unknown quantities, and over the years this data analysis aspect of machine learning has come to play a very large role in the field.”

Data Science for Bussiness, Foster Provost & Tom Fawcett.

Somos buenos prediciendo... pero

- El aprendizaje de maquina no se preocupa de estudiar causalidad.
- El aprendizaje de maquina asume que las tendencias se mantendrán iguales en el tiempo.



Aprendizaje supervisado y no supervisado

Aprendizaje Supervisado

- Se requieren datos con una etiqueta.
- Esta etiqueta puede ser categórica (ej Hombre o Mujer) o continua (cuanta cantidad de dinero gana).



Algunos ejemplos

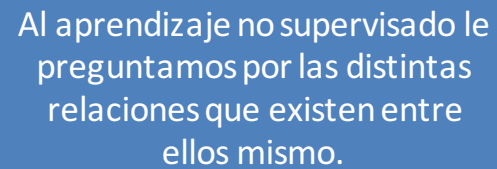
- Regresión Lineal.
- Regresión Logística.
- Árboles de Decisión.
- Support Vector machines.
- Y siguen apareciendo mas....

Aprendizaje No Supervisado

- No se requieren etiquetas en los datos.



Datos
Independientes



Al aprendizaje no supervisado le
preguntamos por las distintas
relaciones que existen entre
ellos mismo.

Algunos ejemplos

- k-meas.
- Expectation Maximization.
- Mean Shift.
- Y siguen apareciendo mas....

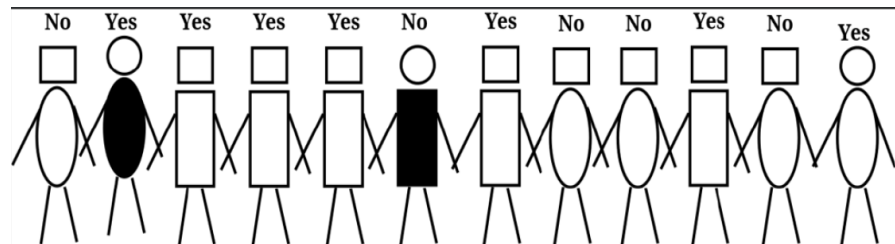


La dinámica del aprendizaje de maquina



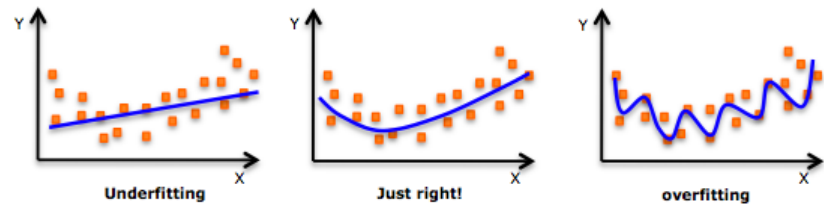
Definiciones y terminología

- Atributo (variable independiente):
 - Forma de la cabeza.
 - Forma del cuerpo.
 - Color del cuerpo.
- Objetivo o Etiqueta (variable dependiente): “Yes” o “No”
- Predicción: Son los valores que resultan de nuestro algoritmo (“Yes”S o “No”).



Definiciones y terminología

- Overfitting: Hacerlo tan bien en un conjunto de datos que perdemos generalidad.
- Underfitting: Hacerlo mal en un conjunto de datos, perdiendo generalidad e incluso tendencias en los datos.



$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Definiciones y terminología

- Entrenamiento: Ajustar los parámetros del algoritmo de forma tal de que se minimicen la cantidad de predicciones que no correspondan a la etiqueta original.
- Recall: Porcentaje de clasificados correctamente como positivos sobre todos los que realmente eran positivos.
- Precision: Porcentaje de clasificados correctamente como positivos sobre todos los clasificados como positivos.
- Accuracy: Porcentaje de clasificados correctamente.

Entendimiento del problema

- Que queremos predecir.
- Que datos necesitamos.
- Que datos podemos obtener.

Entendimiento del problema

- Que queremos predecir.
- Que datos necesitamos.
- Que datos podemos obtener.

Preparación de los datos

- Tomar los datos previamente recopilados, y les damos un formato adecuado. Esto quiere decir, por ejemplo, la variable que habla del color de pelo de una persona puede tener varios valores (rubio, castaño, pelirojo, negro... etc), en algunos casos los algoritmos necesitan que se les entreguen los datos como números, es decir por ejemplo rubio=0, castaño=1, pelirojo=2, negro=3 (existen librerías que hacen eso por nosotros).

Conjunto de Entrenamiento, Cross Validation, Conjunto de Test.

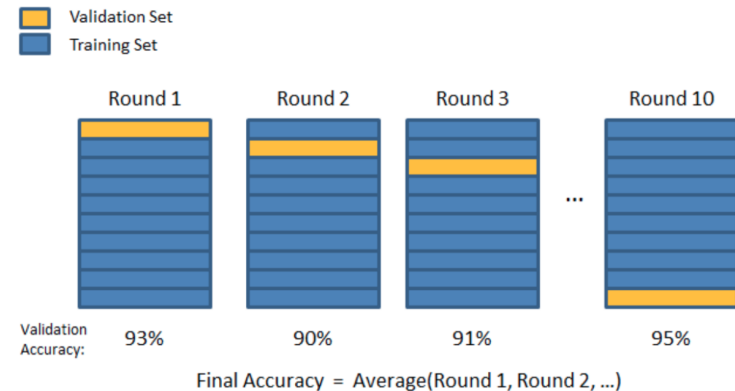
- Para poder realizar una buena comprobación del desempeño de nuestro entrenamiento es importante separar nuestros datos en conjunto de:
 - Entrenamiento.
 - Conjunto de test.

Porque usamos estos 2 conjuntos de datos?

- En la realidad nuestro modelo será aplicado a datos que es nunca vio previamente, entonces nos gustaría tener una intuición de cuan bien lo haría en estos datos.
- IDEA: tomemos una parte de los datos y no la usemos en el entrenamiento, así nuestro algoritmo nunca lo habrá visto anteriormente. Pero como sabemos el verdadero valor de la variable objetivo tenemos la posibilidad de comparar lo que predeciría nuestro algoritmo y el verdadero valor (Conjunto de Test).

K-fold Cross Validation.

- Como veremos mas adelante los algoritmos tiene distintos parámetros que debemos decidir a priori (por ejemplo en un árbol de decisión seria el numero máximo de “filtros” que hay). Para evitar que el ajuste de esos parámetros nos guie a overfitting, podemos separar el conjunto de entrenamiento de la siguiente forma.



Selección del algoritmo de aprendizaje.

- La selección del algoritmo de aprendizaje va a estar guiada por el tipo de etiqueta que se tenga (continua, discreta).
- La interpretación también es una herramienta que ayuda mucho, pues hay algoritmos que son muy fáciles de interpretar y usar (ej árbol de decisión).
- Una vez que elegimos nuestro algoritmo, entrenamos realizando el k-fold (siempre que sea posible) si su resultado es satisfactorio, entrenamos usando todo el training set como uno solo.
- Recomendación: probar con distintos algoritmos y ver los resultados.

Validación.

- Hasta este punto nunca hemos tocado el test set, esto es sumamente importante pues queremos que en este paso el algoritmo nunca haya visto esos datos.
- En la validación obtenemos todos los indicadores (Accuracy, Precision, Recall) de los datos de Test. Esto nos permitirá saber como se comportara nuestro algoritmo en datos generales.

A background image with an orange tint showing a man and a woman working on laptops. The man on the left is smiling and has his hand near his face, while the woman on the right is also smiling. A decorative pattern of blue dots is located to the left of the text.

Como comparar modelos.

Comparación entre modelos.

- En este punto estamos suponiendo que el test set y training set para ambos algoritmos es el mismo.
- Para elegir el mejor algoritmo no hay una receta fija, pues depende del problema.
 - Si queremos bajos falsos negativos elegimos el que tenga el mejor recall.
 - Si queremos bajos falsos positivos elegimos el que tenga mayor precision.
 - Pero no es claro que sea bueno enfocarse solo en una medida.

A background image of a man and a woman working on laptops. The man on the left is wearing glasses and has a blue dot pattern on his hand. The woman on the right is also wearing glasses and smiling. The entire image has an orange overlay.

Como mejorar modelos.

Y si todo anda mal?

- Si a pesar de todos nuestros esfuerzos, en el momento de la validación o cuando implementamos la solución propuesta por el algoritmo, el resultado es deficiente:
 - Conseguir mas datos.
 - Probar con menos atributos.
 - Conseguir mas atributos.
 - Añadir atributos polinomiales (la edad al cuadrado, altura al cubo, raíz cuadrada del ingreso, etc).



Modelo Regresión Logística.



Regresión Logística

- Aprendizaje supervisado.
- La etiqueta u objetivo debe ser categórico (ej: Pago o No Pago; Buen, Regular o Mal desempeño; etc), o estar entre 0 y 1.
- Este algoritmo estima las probabilidades de pertenencia a un grupo (Probabilidad de que tenga un Buen, Regular o Mal Desempeño).

The background image shows a man and a woman in an office setting, both wearing glasses and smiling while working on laptops. The man is on the left, and the woman is on the right. The entire image has an orange tint. Overlaid on the man's face is a diagram of a k-means cluster, consisting of six blue dots arranged in a hexagonal pattern.

k-means.

k-means

- Aprendizaje no supervisado.
- Este algoritmo nos permite segmentar nuestros datos, encontrando un numero deseado a priori de Centroides.
- Cada centroide representa, un “tipo de dato” el cual no es siempre fácil darle una interpretación.

Como funciona?

