Introducción al Aprendizaje Automático

Intuición, Algoritmos y Mejores Prácticas

David Omar Flores Chávez

Escuela Superior de Cómputo
davidomarfch@gmail.com

18 de octubre de 2017

Resumen

Este artículo pretende resaltar la importancia de la intuición de los algoritmos más utlizados en el aprendizaje automático. También se proporcionará una pequeña guía de visualización de datos para realizar las exploraciones iniciales en cada problema. El objetivo, es que más allá de saber cómo implementarlos, el lector sepa elegir qué algoritmo es el apropiado para un conjunto de datos en particular, ahorrando tiempo y recursos técnicos y humanos.

1. Introducción

El Aprendizaje Automático, Aprendizaje de Máquinas o *Machine Learning* es una rama de la inteligencia artificial que pretende otorgar a las computadoras la **capacidad de aprender**.

El término fue acuñado en 1959 por el informático Arthur M. Samuel, y lo definió como *«el campo de estudio que le otorga a las computadoras la habilidad de aprender, sin ser programadas de manera explícita.»* [1]

Tom Mitchell, en el año de 1997, propone una definición más formal para el término: *«Una computadora aprende de la experiencia E, con respecto a una tarea T, y una medida de rendimiento P, si su rendimiento en la tarea T, siendo medido por P, mejora con la experiencia E»* [2]

2. Tipos de Algoritmos

Los algoritmos de aprendizaje pueden clasificarse en tres categorías principales, tomando como criterio las características de los datos con que se entrenan:

- Aprendizaje supervisado
- Aprendizaje no-supervisado
- Aprendizaje por refuerzo

2.1. Aprendizaje Supervisado

Permite inferir información a partir de un conjunto de datos etiquetados[1].

Cada muestra consiste de

(a) **una entrada**, y

(b) **una salida esperada.**

Este tipo de aprendizaje analiza los datos de entrenamiento, aprende de ellos, y produce una **función** que puede usarse para *mapear* nuevas muestras.

Ejemplos de posibles aplicaciones de estos algoritmos, pueden ser:

[1] Grupo de muestras cuyas características han sido etiquetadas para crear datos más significativos.

- Aprendiendo del tamaño, velocidad de crecimiento (entrada), y el tipo (salida) de miles de tumores, predecir si un tumor diferente será benigno o maligno.
- Aprendiendo del tamaño, estilo arquitectónico, número de baños (entrada), y costo (salida) de miles de casas, predecir un nuevo costo de una casa diferente.

2.2. Aprendizaje No Supervisado

Permite descubrir -tanto si existe, como si no-, alguna estructura oculta dentro de un conjunto de datos.

Dado que la característica principal de este tipo de aprendizaje es la utilización de datos no etiquetados, es imposible conocer la validez de la estructura descubierta por el algoritmo.

Ejemplos de posibles aplicaciones de estos algoritmos, pueden ser:

- Analizando una base de datos de miles de clientes, poder generar distintos segmentos de mercado, y describir las cualidades de cada uno.
- Analizando una enorme cantidad de fotos, separarlas todas, utilizando como criterio de separación la cara de las personas que aparecen en las fotos.

3. Regresión

Los problemas de regresión son útiles en la predicción de **valores reales**.

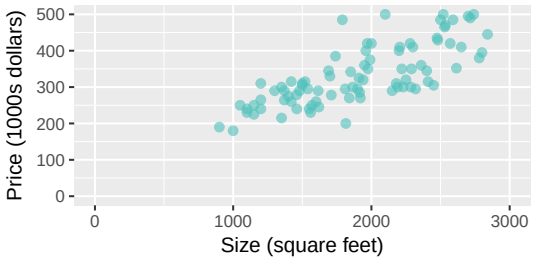
Comencemos con un ejemplo motivador para la predicción del costo de viviendas.

Usaremos un conjunto de datos ficticio que contiene

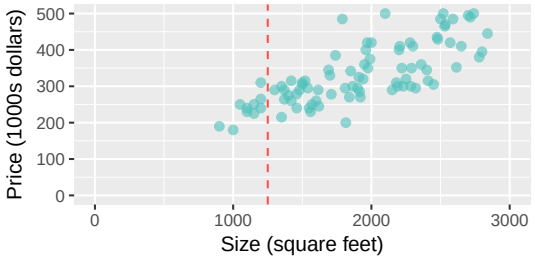
(a) el tamaño de una vivienda en pies cuadrados, y

(b) el precio de la vivienda en miles de dólares

Si intentamos visualizar la distribución de los datos, tendremos la siguiente gráfica:

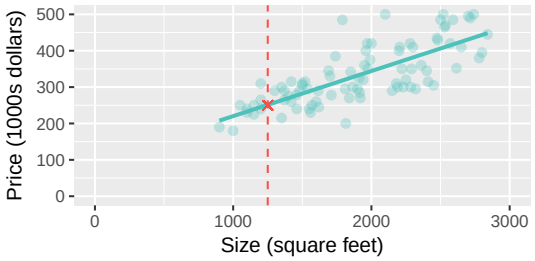


Ahora, imaginemos que tenemos una vivienda de 1250 ft^2 y nos interesa tener una idea de cuál es el precio que podría alcanzar en el mercado.



Para hacerlo, podemos intentar **ajustar un modelo** a la gráfica, y usarlo para hacer nuestras predicciones.

Ajustando una línea recta, podemos estimar que una vivienda de 1250 ft^2, tiene un costo de $ USD 250, 000.



Este es un ejemplo de **aprendizaje supervisado**: cada *evento* de nuestro conjunto de datos cuenta con la *salida esperada*: el precio de cada vivienda.

3.1. Nomenclatura y notación

Más formalmente, en el aprendizaje supervisado, tenemos un conjunto de datos llamado **conjunto de entrenamiento**.

El conjunto de entrenamiento contiene m ejemplos (χ, y), donde χ es una variable de entrada, e y una variable de salida.

Ambos valores son vectores de dimensión m.

$$\chi, y \in \mathbb{R}^m$$

Para referirnos a un elemento i dentro del conjunto, utilizamos la notación $\chi^{(i)}$ e $y^{(i)}$ siempre que $1 \leq i \leq m$.

Es importante que al usar esta notación, se entienda que no se trata de exponenciación.

$$\chi^{(i)} \neq \chi^i$$

Cuando la variable de entrada tiene **más de una característica**, χ es una matriz de $m \times n$ dimensiones,

$$\chi \in \mathbb{R}^{m \times n}$$

donde n es el número de características de cada entrada.

Nótese que si $n = 1 \implies \chi \in \mathbb{R}^{m \times 1}$, por lo que χ es un vector de dimensión m.

Para referirnos a una característica j específica, utilizamos la notación χ_j siempre que $1 \leq j \leq n$.

Apliquemos la notación mencionada para tratar de explorar el conjunto de datos de la siguiente tabla.

Cuadro 1: *Conjunto de entrenamiento.*

Ejemplo	Tamaño	N. Hab.	N. Baños	Costo
1	1160	2	1	468
2	1620	4	1.75	385
3	990	3	1	210
4	2753	3	2.5	1135
5	1980	2	1.75	585
6	1000	3	1	204

Sabemos que nuestro conjunto de datos cuenta con una variable de entrada χ, y una variable de salida y.

La variable de entrada χ tiene las características de:

1. área,
2. número de habitaciones,
3. número de baños

El tamaño del conjunto es $m = 6$, y el número de características de χ es $n = 3$, por lo tanto,

$$\chi \in \mathbb{R}^{6 \times 3}, y \in \mathbb{R}^6$$

Y los valores de cada variable son:

$$\chi = \begin{bmatrix} 1160 & 2 & 1 \\ 1620 & 4 & 1{,}75 \\ 990 & 3 & 1 \\ 2753 & 3 & 2{,}5 \\ 1980 & 2 & 1{,}75 \\ 1000 & 3 & 1 \end{bmatrix}, y = \begin{bmatrix} 468 \\ 385 \\ 210 \\ 1135 \\ 585 \\ 204 \end{bmatrix}$$

Si queremos referirnos al tercer ejemplo,

$$(\chi^{(3)}, y^{(3)}) = (\begin{bmatrix} 990 & 3 & 1 \end{bmatrix}, 210)$$

Si queremos referirnos a la primer característica,

$$\chi_1 = \begin{bmatrix} 1160 \\ 1620 \\ 990 \\ 2753 \\ 1980 \\ 1000 \end{bmatrix}$$

Y si queremos saber el número de baños del 5to ejemplo,

$$\chi_3^{(5)} = 1{,}75$$

En resumen:

χ = Variable de entrada
y = Variable de salida
m = Número de ejemplos de entrenamiento
n = Número de características deχ
$\chi^{(i)}$ = i-ésimo ejemplo
χ_j = j-ésima característica

Referencias

[1] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229, July 1959.

[2] Tom Michael Mitchell. *Machine Learning*. McGraw-Hill Series in Computer Science. WCB/McGraw-Hill, Boston, MA, 1997.