Módulo 3

3.3 Aprendizaje no supervisado

*Por **Rafael Alcalá Fernández***
Catedrático de la Universidad de Granada. Instituto Andaluz Interuniversitario en *Data Science and Computational Intelligence* (DaSCI)

1. ¿QUÉ ES EL APRENDIZAJE NO SUPERVISADO?

Como se ha visto en la cápsula anterior, en el área del aprendizaje supervisado, normalmente tenemos acceso a p características (variables de entrada) $X_1, X_2, \dots, X_p$ sobre un conjunto de n observaciones (instancias), y una respuesta Y (variable de salida) asociada a las mismas n instancias. El objetivo es entonces predecir Y usando $X_1, X_2, \dots, X_p$.

Sin embargo, no siempre se dispone de una respuesta asociada. O incluso si se dispone de ella, podríamos tener interés en descubrir otro tipo de asociaciones. En estos casos, se pueden utilizar una serie de técnicas que se engloban dentro de lo que se llama aprendizaje no supervisado. El término *no supervisado* hace referencia a que este aprendizaje **no se basa en la existencia de una respuesta previamente conocida**. Es decir, no se supervisa el aprendizaje, como un maestro haría con un estudiante, proporcionando la respuesta correcta después de cada intento fallido del estudiante (véase la Figura 1).

El **aprendizaje no supervisado** comprende un conjunto de herramientas estadísticas destinadas al **entorno en el que solo tenemos (o usamos) un conjunto de variables $X_1, X_2, \dots, X_p$ sobre un conjunto de n instancias**. No estamos interesados en la predicción, porque no tenemos una variable de salida asociada Y. Más bien, el objetivo es descubrir cosas interesantes sobre las distintas mediciones $X_1, X_2, \dots, X_p$:

- ¿Hay alguna manera informativa de representar los datos?
- ¿Podemos descubrir subgrupos entre las instancias?
- ¿Sería posible encontrar relaciones de interés entre las propias variables?

El aprendizaje no supervisado se refiere a un conjunto diverso de técnicas para responder preguntas como las anteriores. Entre las más conocidas se encuentran el *clustering* y las *reglas de asociación*, aunque también son parte del aprendizaje no supervisado técnicas como el *análisis de componentes principales* (técnica que vimos en la Sección 2.3. de la cápsula 2 del módulo 2). En este MOOC introduciremos las dos primeras.

Aprendizaje no supervisado

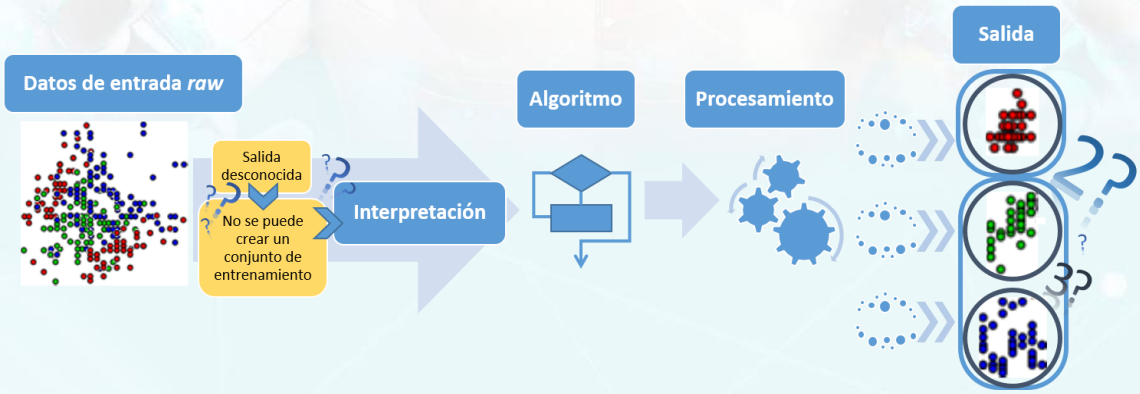

Figura 1. Representación esquemática del aprendizaje no supervisado como contrapunto al supervisado, que si dispone de salida conocida. Figura diseñada para este MOOC

2. LA DIFICULTAD DEL APRENDIZAJE NO SUPERVISADO VS. SUPERVISADO

El aprendizaje supervisado es un área con objetivos bien definidos. Por ejemplo, si se le pide predecir una salida binaria para un conjunto de datos, tiene un conjunto extremadamente extenso de herramientas muy bien desarrolladas a su disposición (como la regresión logística, el análisis discriminante lineal, los árboles de clasificación, las máquinas de vectores de soporte y más) junto con una comprensión clara de cómo evaluar la calidad de los resultados obtenidos (mediante validación cruzada, validación en un conjunto de pruebas independiente, métricas de estimación directa del error, etc.).

Por el contrario, **el aprendizaje no supervisado suele ser mucho más desafiante**. Su aplicación tiende a ser más subjetiva y no existe un único objetivo claro para el análisis, como la predicción de una variable de salida. El aprendizaje no supervisado a menudo se realiza como parte de un análisis exploratorio de datos. Además, puede ser muy difícil evaluar los resultados obtenidos a partir de métodos de aprendizaje no supervisados, ya que no existe un mecanismo universalmente aceptado para realizar validaciones cruzadas o validar resultados en un conjunto de datos independiente. La razón de esta diferencia es simple. Si ajustamos un modelo predictivo utilizando una técnica de aprendizaje supervisado, entonces es posible verificar nuestro trabajo al ver qué tan bien nuestro modelo predice la variable de salida Y en las instancias no utilizadas para ajustar el modelo. Sin embargo, en el aprendizaje no supervisado, no hay forma directa de verificar nuestro trabajo porque no sabemos la respuesta verdadera: el problema no está supervisado.

3. IMPORTANCIA Y APLICABILIDAD DEL APRENDIZAJE NO SUPERVISADO

Las técnicas para el aprendizaje no supervisado son cada vez más importantes en multitud de áreas. Un sitio *web* de compras en línea podría intentar identificar grupos de compradores con historiales similares de navegación y compras, así como artículos que sean de particular interés para los compradores dentro de cada grupo. Igualmente, también podrían querer buscar relaciones causales entre la compra de unos artículos y otros. Luego, se puede mostrar preferencialmente a un comprador individual los artículos en los que es más probable que esté interesado, según los historiales de compra de compradores similares. Un motor de búsqueda podría elegir qué resultados de búsqueda mostrar a un individuo en particular en función del historial de clics de otros individuos con patrones de búsqueda similares. O se podrían mostrar los artículos que es más probable que compre tras haber realizado ciertas compras con anterioridad según las relaciones causales aprendidas.

Igualmente, acercándonos al área de la bioinformática, un investigador del cáncer podría evaluar los niveles de expresión génica en 100 pacientes con cáncer de mama. Luego, podría buscar subgrupos entre las muestras de cáncer de mama o entre los genes, para obtener una mejor comprensión de la enfermedad. Igualmente, se podrían buscar relaciones causales entre la activación de unos genes y otros (que podrían ser considerados marcadores de interés), e incluso entre la activación de ciertos grupos de genes y el desarrollo de alguno de los tipos de cáncer de mama. Estas tareas de aprendizaje estadístico, y muchas más, se pueden realizar a través de técnicas de aprendizaje no supervisadas.

4. CLUSTERING

El *clustering* es un conjunto de técnicas de *aprendizaje no supervisado* cuyo objetivo es la **identificación de grupos en los datos**. Un grupo (o *cluster*) es un conjunto de **objetos (instancias) que se parecen entre sí**. De este modo, el objetivo de un algoritmo de *clustering* es agrupar los objetos disponibles de forma que los objetos dentro del mismo grupo sean similares entre sí, y diferentes a los objetos de otros grupos.

Otra posible explicación, teniendo en cuenta que ya hemos visto qué es la clasificación en la cápsula anterior, es que el *clustering* consiste en realizar una **clasificación sobre un conjunto de datos sin conocer previamente las clases**. Ya no es sólo que no se dispone de la información relativa a la clase a la que puede pertenecer la instancia, sino que ni siquiera se conoce si existen clases ni cuantas. Por eso, se intenta agrupar por similitud en grupos que, *a priori*, son desconocidos y suficientemente distintos de otros grupos como para poder afirmar que dicho grupo representa un *cluster*, es decir, una entidad o posible clase.

Resulta casi intuitivo pensar que un problema de *clustering* es *a priori* más difícil de resolver que uno de clasificación, ya que en el caso de la clasificación al menos se conocen las clases del

conjunto de datos. En general, dicha afirmación es cierta. Si se conocen las clases será más fácil que los datos queden bien clasificados (o agrupados por clases). No obstante, el descubrimiento de información que proporcionan las técnicas de *clustering* suele ser de altísima utilidad cuando los *clusters* obtenidos están bien claros (incluso cuando únicamente se tiene un *cluster* claramente identificado, si resulta que es el grupo de mayor interés que estábamos buscando).

Existen numerosas funciones para medir similitud o distancia entre objetos. El uso de una u otra depende del tipo de variables y del problema. Además de la definición de similitud/distancia que se proponga, existen multitud de algoritmos de *clustering* que pueden proporcionarnos *clusters* diferentes incluso para un mismo problema.

Presentaremos algunas de las medidas de similitud/distancia más conocidas junto con la metodología asociada a algunos de los algoritmos más conocidos en el Módulo 6 de este MOOC. En la Figura 2, se muestra la evolución (convergencia) del conocidísimo algoritmo *k-means* en un ejemplo con $K = 3$ *clusters*, que inicialmente se generan centrados en tres instancias escogidas de manera aleatoria. Antes de visualizar la animación, fíjese en la distribución de datos e imagine cómo deberían de quedar los *clusters* al final.

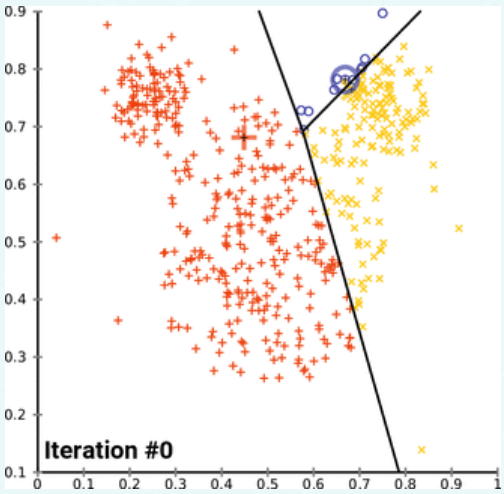

Figura 2. Convergencia del algoritmo *k-means*
(**mueva el ratón sobre la imagen hasta que aparezca el enlace, haga clic y pulse permitir para ver la animación en el navegador**). Dominio Público (Wikipedia).

5. REGLAS DE ASOCIACIÓN

Las reglas de asociación son utilizadas para **identificar y representar dependencias entre los elementos o valores de un conjunto de datos** en el que, igualmente, no conocemos la clase a la que pertenecen. Estas reglas son definidas como expresiones del tipo:

$$A \rightarrow C,$$

donde *A* y *C* son conjuntos de elementos que verifican $A \cap C = \emptyset$. *A* se conoce como el antecedente de la regla y *C* como el consecuente de la regla. Estas reglas representan que cuando en una instancia del conjunto de datos aparecen los elementos de *A*, con una alta probabilidad también aparecen los elementos de *C* en esa misma instancia.

Las reglas de asociación fueron inicialmente utilizadas para detectar asociaciones entre los productos que compraban los consumidores en un supermercado. En ese caso particular:

- Los productos del supermercado son los elementos entre los que queremos encontrar asociaciones, denominados **ítems**.
- A un conjunto de *k* productos se denomina ***itemset***. Más concretamente se les denomina ***k-itemset***, donde k es el número de ítems que lo componen.
- Y a cada una de las ventas del supermercado se le denomina **transacción**.

Al analizar estos conjuntos de datos podemos extraer reglas como:

Pañales → Toallitas,

indicando que cada vez que se compran pañales también se compran toallitas. Este tipo de reglas pueden ayudar al dueño a tomar decisiones sobre qué ofertas realizar y cómo distribuir los productos en el supermercado para aumentar las ventas, mejorar la calidad del servicio proporcionado e incrementar el grado de satisfacción de los clientes. Si bien la regla mostrada parece evidente, también pueden aparecer otras como *Pañales → Cervezas* que en principio podrían no ser tan evidentes. Así, tenemos una dualidad entre la aparición de reglas que sólo representan una manera de confirmar que el algoritmo está funcionando correctamente (información ya conocida) y la aparición de reglas que representan el verdadero descubrimiento (las que realmente buscamos).

Las reglas de asociación son comúnmente evaluadas haciendo uso de las medidas clásicas de **soporte** (frecuencia con la que la regla se cumple en el conjunto de datos) y **confianza** (que indica en cuántas transacciones del conjunto de datos en las que aparece el antecedente también aparece el consecuente de la regla). De manera general, los algoritmos tratan de obtener reglas con alta confianza cumpliendo con un mínimo soporte. No obstante, a día de hoy, existen muchas medidas de calidad para seleccionar y clasificar las reglas en función de su potencial interés para el usuario: *Lift*, *leverage*, *conviction*, etc. En el Módulo 6, se presentan las distintas alternativas en cuanto a

medidas de calidad y la metodología de los algoritmos más conocidos, así como la manera de ejecutarlos en Python.

6. CARACTERÍSTICAS DE LOS DATOS QUE INFLUYEN EN EL APRENDIZAJE

Al igual que en el caso del aprendizaje supervisado, un tema importante a discutir es si resulta posible determinar la cantidad de datos que es óptima para realizar un correcto aprendizaje, así como la relación entre el número de instancias y el número de variables que las representan. Hay que tener en cuenta que en este caso también se presentan los mismos problemas que en el caso del aprendizaje supervisado, e incluso agravados por los mismos motivos indicados en la sección 2.1 de esta cápsula, ya que el aprendizaje no supervisado suele presentar más dificultades:

- **Sobre el número de instancias**: Igualmente, es necesario tener un mínimo número de instancias (datos), que como ya se ha dicho suele depender también del número de variables. Nos podemos guiar por los valores indicados para el aprendizaje supervisado en la sección 2.1 de la cápsula anterior. No obstante, ya sabemos que no hay una respuesta exacta.
- **Sobre las variables de entrada**: En este caso todas son variables de entrada, por lo que no se busca una relación de dependencia con respecto a una variable de salida. No obstante, también tendremos la existencia de variables "*confounding*" que pueden causar agrupamientos de tipo espurios, que en el caso de las reglas de asociación no representen verdadera causalidad o que en el caso del *clustering* puedan hacer pensar que existen más grupos de los que realmente son. Igualmente la "dimensionalidad" puede ser un problema para la obtención de buenos resultados, por lo que muchas veces se aplica pre-procesamiento para intentar reducirla.
- **Sobre el sesgo de los datos**: Aparte del problema de las variables "*confounding*", motivadas normalmente por un sesgo en los datos, el aprendizaje no supervisado también se ve gravemente afectado por los desequilibrios en la distribución de cualquiera de las variables. Será muy difícil el poder agrupar o asociar si no se tiene una muestra representativa de todos por posibles casos o situaciones en nuestros datos.

Finalmente, puesto que en el caso del *clustering* necesitamos calcular la similitud para formar los clusters, la **heterogeneidad de las variables** también puede representar un serio problema. Si los atributos son numéricos y son directamente comparables, el cálculo de la similitud o distancia es realmente sencillo. Cuando los ejemplos contienen atributos complejos y heterogéneos, las cosas se vuelven más complicadas. Usualmente se utiliza la distancia euclídea, pero cuando tenemos variables categóricas como por ejemplo el *sexo*, no se puede calcular directamente. Se suelen asignar valores numéricos (hombre=0, mujer=1 por ejemplo), pero sólo se puede si únicamente tenemos dos valores o si hay una relación ordinal entre los valores categóricos. Por ejemplo, "prepuber"=1, "puber"=2 y "adulto"=3.

En este sentido también hay que tener en cuenta factores de escala. Si se están calculando distancias, los datos deberían de estar normalizados a un mismo rango. Usualmente se aplica una normalización de los valores al rango [0,1].

Debido a todo lo anterior, es muy importante prestar atención al análisis exploratorio de los datos, y realizar un preprocesamiento o preparación del conjunto de datos para facilitar las tareas de aprendizaje.

7. EVALUACIÓN DE LOS RESULTADOS DEL APRENDIZAJE NO SUPERVISADO

Al igual que en el caso del aprendizaje supervisado, un tema importante a discutir es si los resultados obtenidos tienen la calidad suficiente. En este caso, no tiene sentido aplicar una validación cruzada, ni ningún tipo de validación basada en el conocimiento de ningún tipo de valores de salida. La pregunta a responder en este caso sería: ¿es la información que hemos descubierto fiable? En nuestro caso, desde el punto de vista de la Bioinformática, sería equivalente a preguntar: ¿está la información que hemos descubierto soportada por los datos y a su vez tiene fundamentación biológica?

Cada vez que se aplica un *clustering* o un aprendizaje de reglas de asociación sobre un conjunto de datos, encontraremos *clusters* y reglas, respectivamente. Particularmente en el caso de las reglas de asociación se pueden llegar a obtener incluso cientos o miles para algunos problemas. Pero realmente queremos saber si los clusters o reglas que se han encontrado representan verdaderos subgrupos en los datos, o si son simplemente el resultado de la agrupación o asociación del ruido. Por ejemplo, si pudiésemos disponer de un conjunto independiente de datos, ¿esos datos también mostrarían el mismo conjunto de *clusters*, o seguirían dando el mismo soporte a las reglas?

Esta es una pregunta difícil de responder. Por ejemplo, existen varias técnicas para asignar un *p-valor* a un *cluster* con el fin de evaluar si hay más evidencia para dicho *cluster* de lo que cabría esperar debido al azar. Igualmente, como ya hemos indicado también hay métricas de calidad para las reglas de asociación. Sin embargo, no ha habido consenso sobre un único mejor enfoque, ya que, en cualquier caso, la verdadera evaluación suele estar ligada a una comprobación/validación realizada por los expertos en el área correspondiente. En Bioinformática, en la gran mayoría de casos, los hallazgos suelen guiar la investigación en laboratorio, y es ahí cuando verdaderamente se comprueba si la información descubierta es o no fiable.

REFERENCIAS BIBLIOGRÁFICAS

- **L. Geng, H. Hamilton.** Interestingness measures for data mining: a survey. ACM Computing Surveys 38:3 (2006) 1–32.
- **P-N. Tan, M. Steinbach, A. Karpatne, V. Kumar.** Introduction to Data Mining (2ª ed.). Pearson, 2019.
- **Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.** An Introduction to Statistical Learning with Applications in R. Springer, 2013 (Capítulos 2 y 10).
- **P. Berkhin.** A Survey of Clustering Data Mining Techniques, Springer, Berlin, 2006.
- **R. Agrawal, T. Imielinski, A. Swami,** Mining association rules between sets of items in large databases. ACM SIGMOD International Conference on Knowledge Discovery and Data Mining, Washington DC (USA), 1993, 207-2016.
- **C. Zhang, S. Zhang**. Association Rule Mining: Models and Algorithms. Lecture Notes in Computer Science 2307, Springer-Verlag, Berlin, 2002.

REFERENCIAS ADICIONALES

- **M.R. Berthold, Ch. Borgelt, F. Höppner, F. Klawonn**. Guide to Intelligent Data Analysis, Springer-Verlag, 2010.
- **F. Berzal, I. Blanco, D. Sanchez, M. Vila.** Measuring the accuracy and interest of association rules: a new framework. Intelligent Data Analysis 6:3 (2002) 221–235.
- **K.J. Cios, W. Pedrycz, R.W. Swiniarski, L.A. Kurgan**. Data mining: A knowledge discovery approach, Springer, Boston, 2007
- **M. Zaki, W. Meira.** Data Mining and Machine Learning: Fundamental Concepts and Algorithms (2ª ed.). Cambridge University Press, 2020.