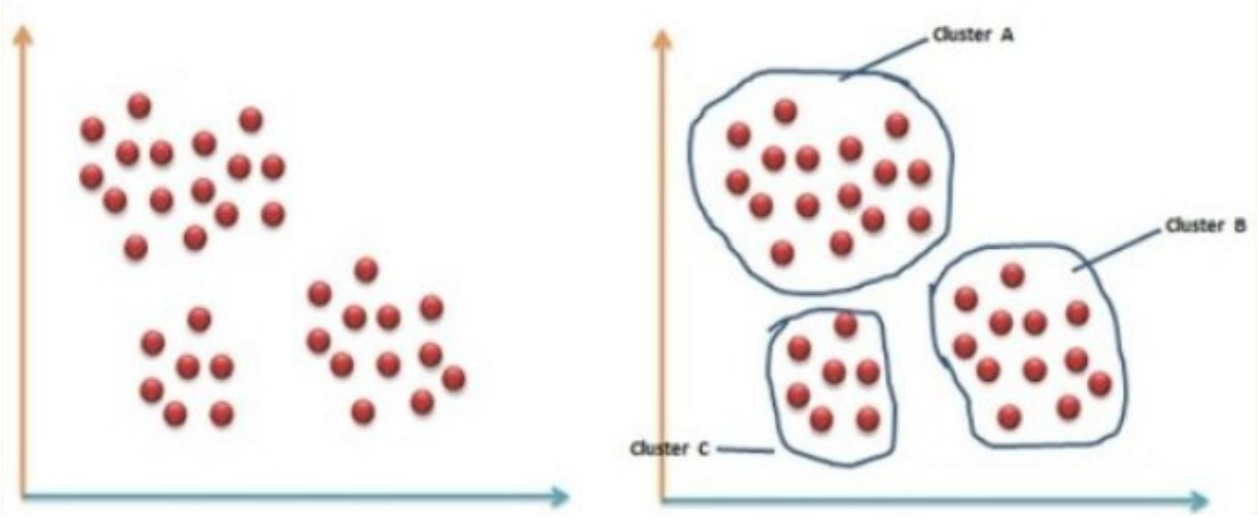
Clustering y aprendizaje no supervisado

En el aprendizaje no supervisado no tenemos una variable objetivo como lo hicimos en la clasificación y la regresión. En lugar de decirle a la máquina "Predecir Y para nuestros datos X", estamos preguntando "¿Qué puede decirme acerca de X?" Cosas que le pedimos a la máquina que nos diga acerca de X puede ser "¿Cuáles son los seis mejores grupos que podemos hacer Fuera de X? "O" ¿Cuáles son las tres características que ocurren con más frecuencia en X? "

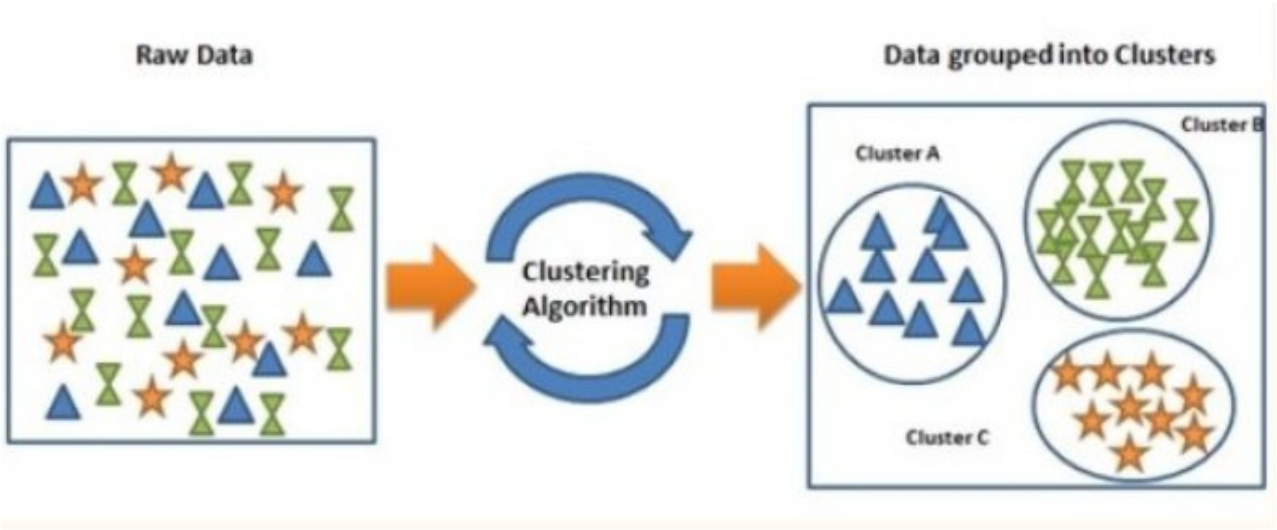
Clustering es un tipo de aprendizaje sin supervisión que forma automáticamente grupos de cosas similares. Es como una clasificación automática. Puede agrupar casi cualquier cosa, y mientras más similares estén los elementos en el clúster, mejores son sus clústeres. En este capítulo, vamos a estudiar un tipo de algoritmo de agrupamiento llamado k-means. Se llama kmeans porque encuentra k clústeres únicos, y el centro de cada clúster es la media de los valores en ese clúster.

Clustering a veces se llama clasificación sin supervisión porque produce el mismo resultado que la clasificación, pero sin tener clases predefinidas. Con el análisis de clúster estamos tratando de poner cosas similares en un clúster y cosas diferentes en un clúster diferente. Esta noción de similitud depende de una medida de similitud.El tipo de medida de similitud utilizado depende de la aplicación.

El objetivo principal de la técnica de agrupamiento es encontrar grupos similares u homogéneos en los datos que se llaman clusters. La forma en que se hace esto es: las instancias de datos que son similares o, en resumen, están cerca una de la otra, se agrupan en un clúster y las instancias que son diferentes se agrupan en un clúster diferente. El siguiente diagrama muestra una representación de puntos de datos en un gráfico, y cómo los clústeres se marcan (en aquí, es por intuición pura) por los tres grupos naturales:



Por lo tanto, un cluster se puede definir como una colección de objetos que son similares entre sí y diferentes de los objetos de otro grupo. El siguiente diagrama muestra el proceso de agrupación:



Clustering es la tarea de particionar el conjunto de datos en grupos, llamados clusters. El objetivo es dividir los datos de tal manera que los puntos dentro de un solo grupo son muy similares y los puntos en los diferentes clústeres son diferentes.

De forma similar a los algoritmos de clasificación, los algoritmos de agrupación asignan (o predicen) un número a cada punto de datos, indicando a qué grupo corresponde un punto particular.

El **Clustering** es una tarea que consiste en agrupar un conjunto de objetos (no etiquetados) en subconjuntos de objetos llamados Clusters. Cada Cluster está formado por una colección de objetos que son similares (o se consideran similares) entre sí, pero que son distintos respecto a los objetos de otros Clusters.

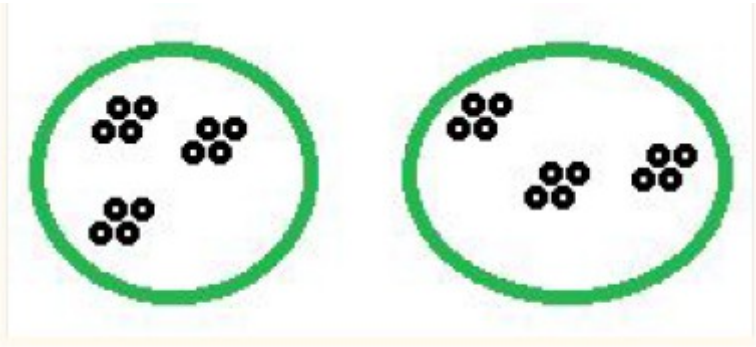
En el campo del ML, el Clustering se enmarca dentro del aprendizaje no supervisado; es decir, que para esta técnica solo disponemos de un conjunto de datos de entrada, sobre los que debemos obtener información sobre la estructura del dominio de salida, que es una información de la cual no se dispone.

Es importante no confundir el Clustering con los problemas de Clasificación. Las técnicas de Clasificación se enmarcan dentro del aprendizaje supervisado porque para cada dato tenemos información sobre sus variables de entrada y de salida; es decir, cada dato u objeto está etiquetado. Sin embargo para aquellos casos en los que no disponemos de la salida de cada dato y queramos agrupar estos objetos en grupos similares, debemos de aplicar alguna de las técnicas de Clustering para saber la procedencia de estos datos.

El objetivo principal es poder determinar la similitud o diferencia de los elementos que conforman nuestra información y en base a esto podemos tomar decisiones.

En muchas ocasiones la noción de los cluster en nuestra información no está bien definida, veamos un ejemplo, tomando en cuenta la siguiente imagen:

Si quisiéramos agrupar los círculos en la misma podríamos tener los siguientes resultados (entre muchos otros):



o bien

uno mas

Como podemos apreciar, depende mucho de la interpretación que deseemos darle a nuestra información. Usualmente una buena idea es conocer los datos antes siquiera de intentar agruparlos, posterior a ello se deben de probar diversos algoritmos de agrupamiento para así poder determinar cual es el más indicado.

Aprendizaje no supervisado

El aprendizaje no supervisado es un paradigma en el aprendizaje automático donde construimos modelos sin etiquetas.

Estos algoritmos se utilizan cuando queremos encontrar subgrupos dentro de conjuntos de datos utilizando alguna métrica de similitud.

Uno de los métodos más comunes es la agrupación. Lo usamos principalmente para el análisis de datos donde queremos encontrar clusters en nuestros datos. Estos grupos se encuentran generalmente utilizando cierto tipo de medida de similitud como la distancia euclidiana. El aprendizaje sin supervisión se utiliza ampliamente en muchos campos, tales como minería de datos, imágenes médicas, análisis de mercado de valores, visión por computadora, segmentación de mercado, y así sucesivamente

El algoritmo k-means es uno de los algoritmos de agrupamiento más populares. Este algoritmo se utiliza para dividir los datos de entrada en k subgrupos usando varios atributos de los datos. El agrupamiento se logra utilizando una técnica de optimización donde se intenta minimizar la suma de cuadrados de distancias entre los puntos de datos y el centroide correspondiente del grupo.

http://www.onmyphd.com/?p=k-means.clustering&ckattempt=1