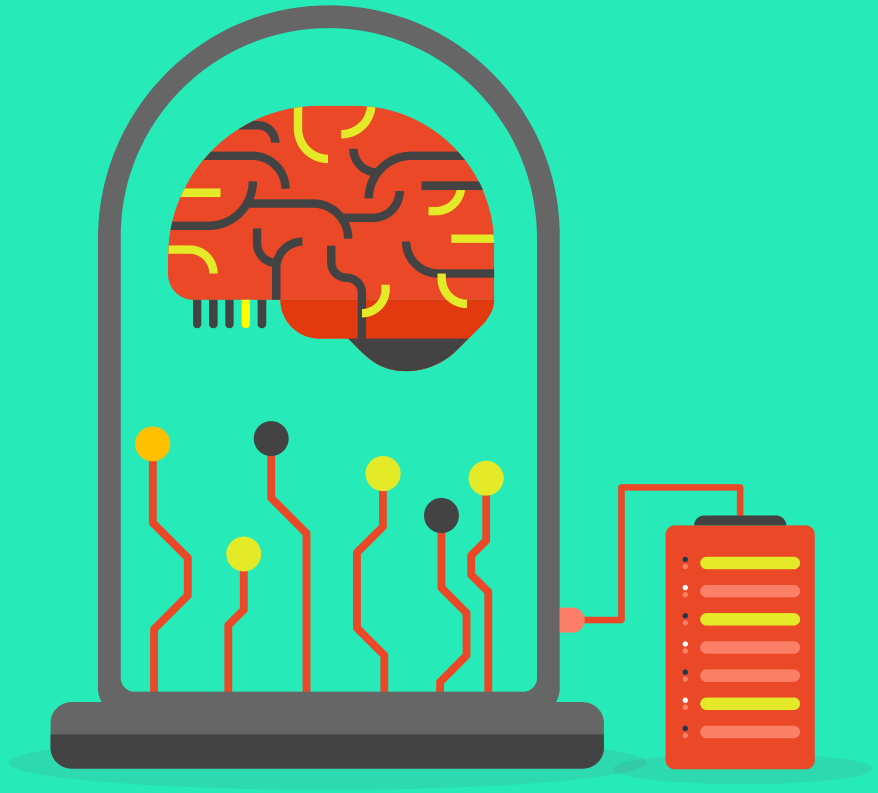
15. Agrupamiento o Clustering



Planteamiento del Problema

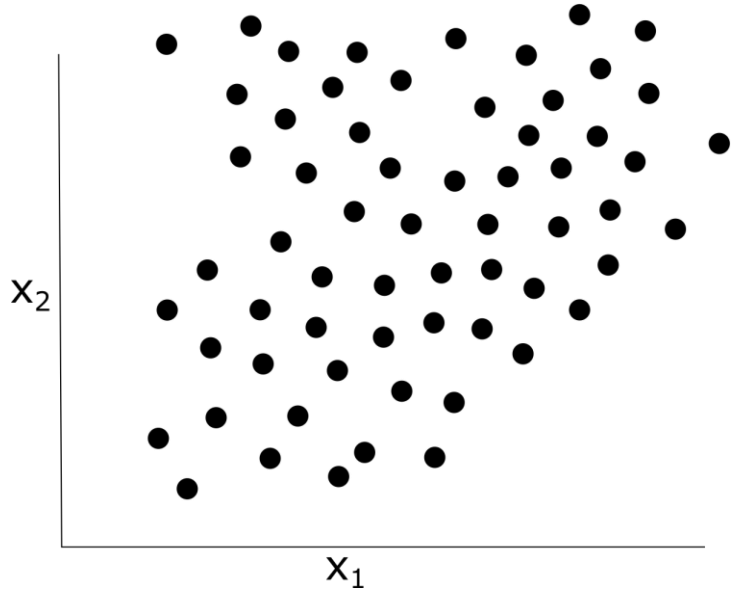
Como se explicó en la Introducción del curso, existen cuatro tipos de aprendizaje en el Machine Learning: **Supervisado**, **No Supervisado**, **Semi Supervisado** y por **Refuerzo**. Todos los modelos de clasificación y regresión que vimos en las secciones anteriores se trataron de modelos **Supervisados**, ya que siempre contamos con las clases o etiquetas (para la clasificación) o bien con el valor numérico deseado (para la regresión).

Sin embargo, no siempre es posible contar con dicha última información. Existen una gran cantidad de problemas en donde se tienen datos o variables sin etiquetas o valores deseados de antemano. Se habla entonces de modelos de aprendizaje **No Supervisados**.

En este tipo de modelos, el objetivo entonces es tratar de encontrar *patrones subyacentes* a los datos que permitan, por ejemplo, clasificar o *agrupar* los mismos.

Planteamiento del Problema

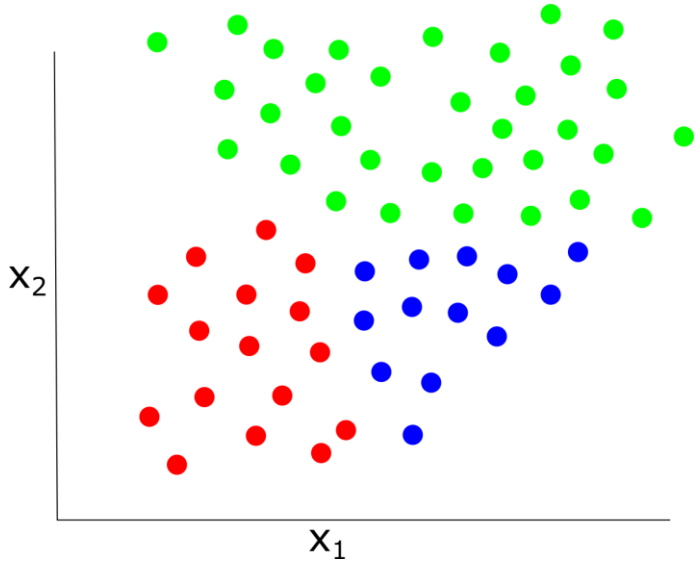
Supongamos que contamos con los siguientes puntos pertenecientes a un conjunto de datos:



A primera vista, no hay mayor información que podemos decir de los datos en cuestión, mas que los mismos están distribuidos de cierta manera.

Planteamiento del Problema

Pero quizá, partiendo de considerar algún tipo de métrica o parámetros de comparación entre los puntos (distancias euclídeas, por ejemplo) podría ocurrir que ciertos datos del conjunto tienen similitudes con otros. Si coloreásemos los datos en función de estas similitudes, podríamos encontrar algo como:



Es decir, en efecto existe un patrón subyacente a los datos.

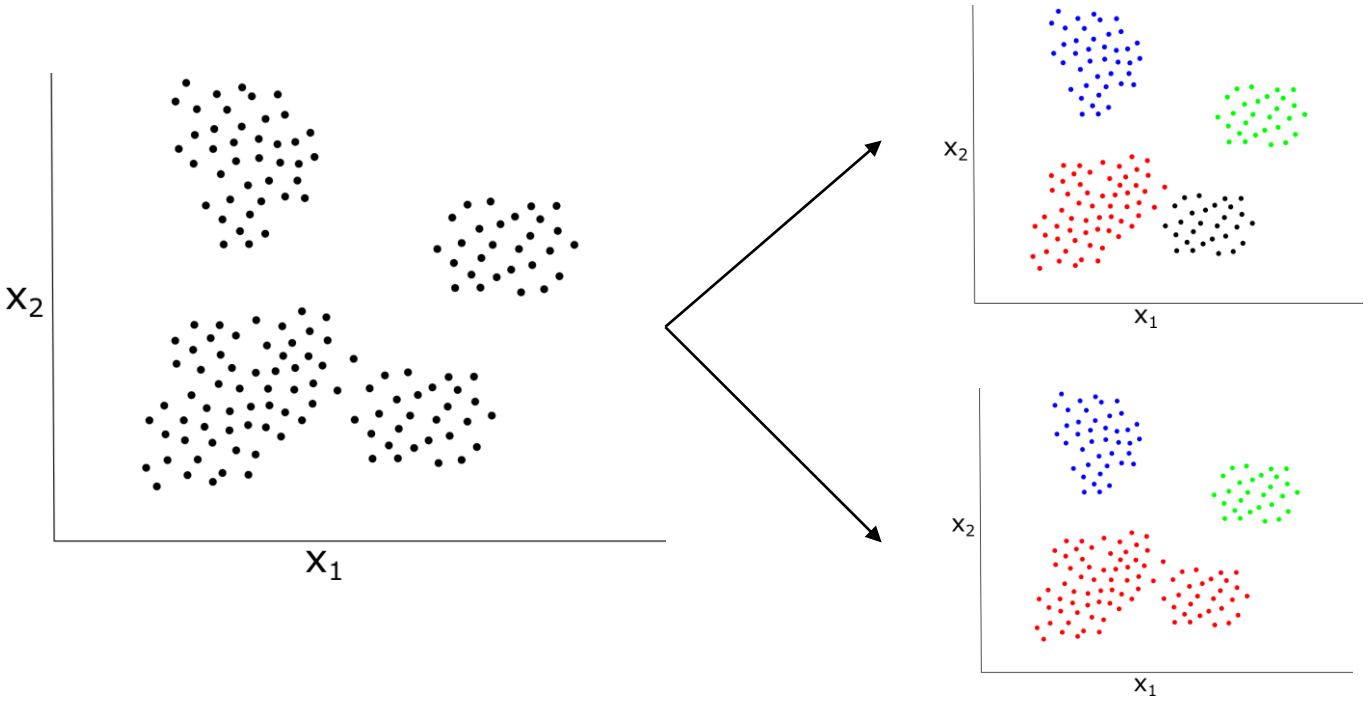
Agrupamiento o Clustering

De manera que, sin necesidad de contar con etiquetas para estos datos, los mismos presentan características que, por algún parámetros de similaridad, es posible agrupar datos similares en regiones o grupos similares. A esto se le conoce, por supuesto, como **Agrupamiento** o **Clustering**.

Una vez agrupados los datos (a partir de un algoritmo), entonces cualquier otro dato nuevo que sea incorporado al conjunto original podrá ser clasificado en función de su similaridad con cada uno de estos grupos obtenidos.

Agrupamiento o Clustering

Los algoritmos de Agrupamiento, se encargan entonces de, a partir de los propios datos de entrenamiento, encontrar de manera automática esos grupos subyacentes:



Agrupamiento o Clustering

Los algoritmos de Agrupamiento son especialmente útiles en problemas como:

- **Segmentación de clientes:** a partir del comportamiento o actividad de clientes o usuarios, segmentarlos a fin de ofrecer servicios o recomendaciones a medida.
- **Análisis de Datos:** encontrar similitudes o patrones entre datos puede ayudar a entender la naturaleza de los mismos.
- **Reducción de dimensionalidad:** es posible codificar datos de grandes dimensiones, en vectores más pequeños que agrupen datos semejantes.
- **Detección de anomalías:** datos que presenten comportamientos que se salen en gran medida de los agrupamientos, pueden considerarse anómalos (gran uso en detección de fraudes bancarios o defectos en fabricación de dispositivos).
- **Motores de búsqueda:** es posible realizar, por ejemplo, búsqueda de imágenes similares a partir de sus agrupamientos naturales.
- **Segmentación de imágenes:** a partir de agrupar pixeles con colores semejantes.

Agrupamiento o Clustering

Sin embargo, no siempre es posible definir lo que es un grupo. Depende del contexto y de la naturaleza de los datos, y algoritmos diferentes capturarán tipos de grupos diferentes, por lo que el uso de los mismos depende mucho del problema en sí mismo y de los agrupamientos que obtengamos para cada uno.

En este curso, estudiaremos e implementaremos los algoritmos de **K-Medios** y el **DBSCAN**.