Clustering

Abraham Jain Jiménez

1 Generalidades

- Clustering se refiere a un amplio conjunto de técnicas de agrupación para detectar grupos (llamados clusters) en un dataset.
- En clustering se busca una partición de los datos en diferentes grupos tal que las observaciones en cada uno sean similares.
- En Machine Learning, los algoritmos de clustering son parte de las técnicas de aprendizaje no supervisado, las cuales permiten agrupar datos sin que haya etiquetas previas, pues se basan en similitudes entre ellos.

2 Clasificación como Aprendizaje Supervisado vs Clustering como Aprendizaje No Supervisado

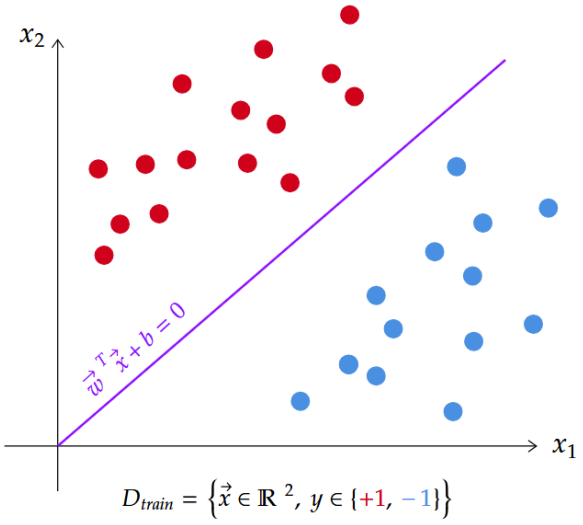
- **Clasificación (S.L):**

  Tenemos un conjunto de datos de entrenamiento

  $$D_{train} = \{\vec{x}_i \in \mathbb{R}^d, y_i \in Y\}_{i=1}^{N}$$

  Por lo que el modelo de aprendizaje es un clasificador $\hat{f}$.

  Consideremos el siguiente ejemplo de clasificación binaria lineal:



La meta de este problema es encontrar una frontera lineal de separación, un hiperplano clasificador.

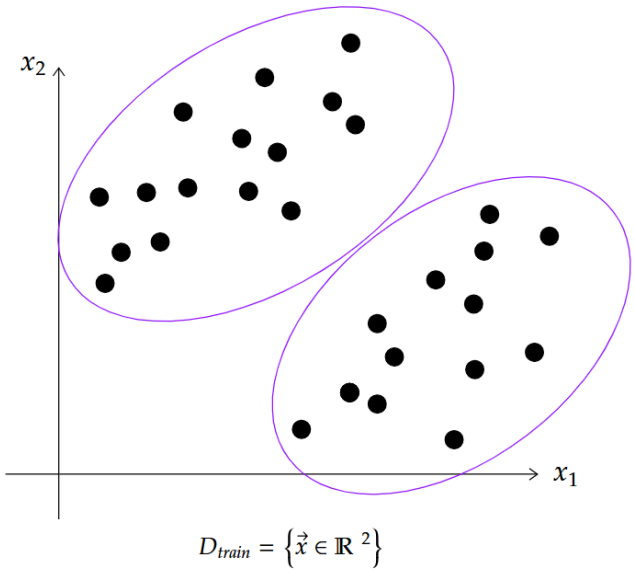
- **Clustering (US.L):**

  Tenemos un conjunto de datos de entrenamiento

  $$D_{train} = \{\vec{x}_i \in \mathbb{R}^d\}_{i=1}^N$$

  Por lo que el modelo de aprendizaje es un asignador a un conjunto de $K \in \mathbb{N}$ clusters, $\{C_K\}$.

  La meta del clustering es aprender estructuras de interés (clusters) en los vectores de características. Por ejemplo, consideremos el siguiente problema:



La meta de este problema es encontrar estructuras de clusters dado un conjunto de vectores de características.

- **Clasificación Lineal vs Clustering:**

  El problema que tiene clustering con respecto a clasificación lineal, es que, por ejemplo en el caso de clasificación binaria, tenemos una respuesta correcta, un datapoint pertenece a la clase roja o pertenece a la azul, y el clasificador dio una predicción bien o mal. En clustering no tenemos esta clase de evaluación, por ejemplo, un conjunto de clusters puede verse como un solo cluster y ambas formas de agrupar son válidas.

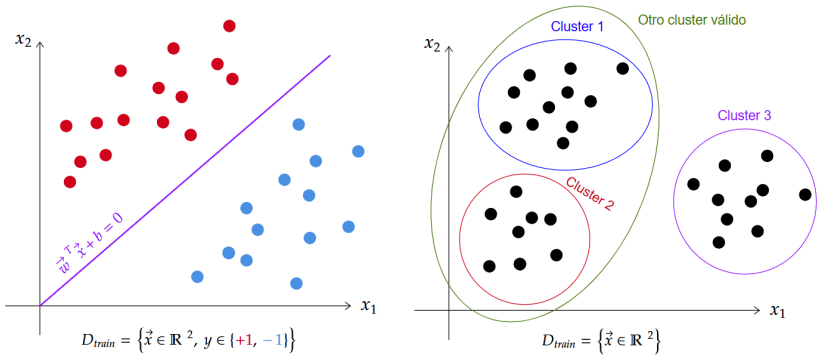


Figure 1: Clasificación Lineal (SL) vs Clustering (USL)