MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl

IIC-2433

T-SNE y UMAP

Stochastic Neighbor Embedding (SNE)

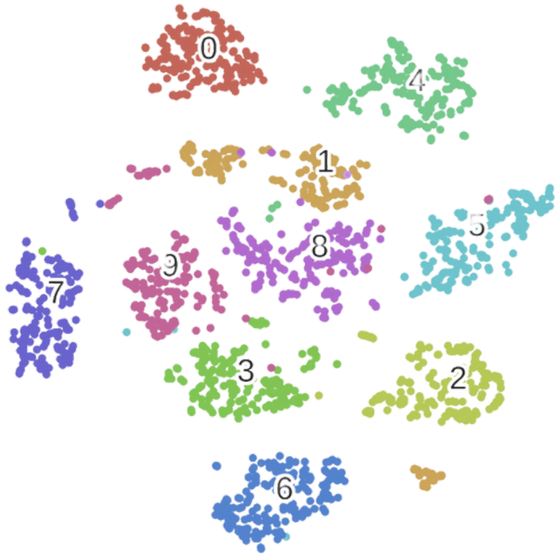
Objetivo

Proyectar datos en 2 o 3 dimensiones para visualización

Problema

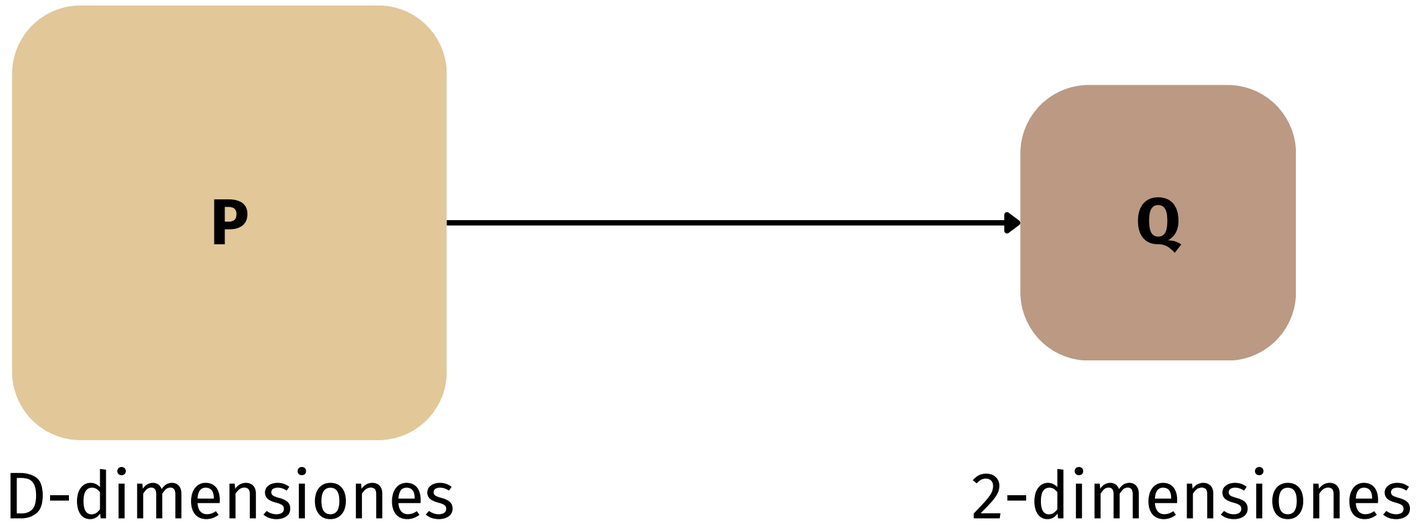
Dados puntos de alta dimensión $x_i \in \mathbb{R}^D (i = 1, \ldots, n)$, t-SNE busca encontrar representaciones que conserven vecindades locales $y_i \in \mathbb{R}^d$

Esto nos lleva al siguiente problema. Como hacer que dos distribuciones de similitud ***P*** (en alta dimensión) y ***Q*** (en baja dimensión) sean lo más parecidas posible



Stochastic Neighbor Embedding (SNE)

Visualización que **preserva distancias** del espacio original

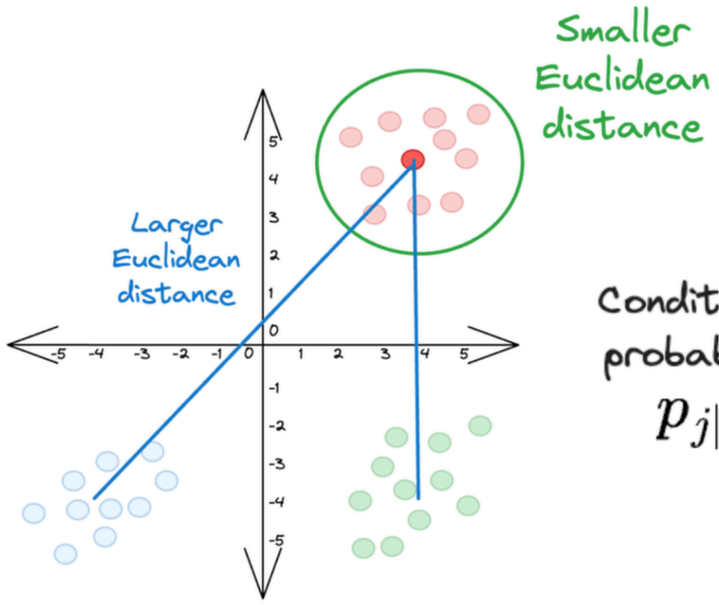


SNE: Similitud en alta dimensión (P)

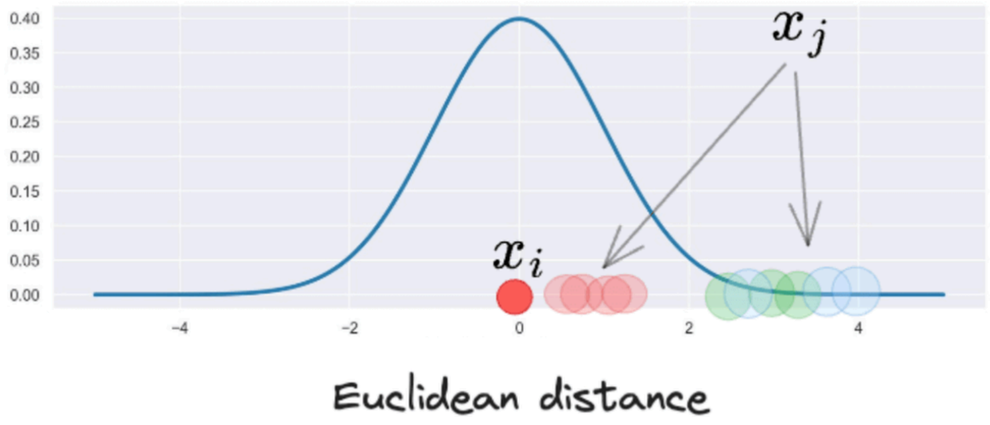
Para cada punto, definimos una distribución condicional sobre sus vecinos:

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-||x_i - x_k||^2 / 2\sigma_i^2\right)},$$

σ_i es el ancho de la gaussiana del punto i



Conditional probability $p_{j|i}$



SNE: Similitud en baja dimensión (Q)

Para cada punto, definimos una distribución condicional sobre sus vecinos:

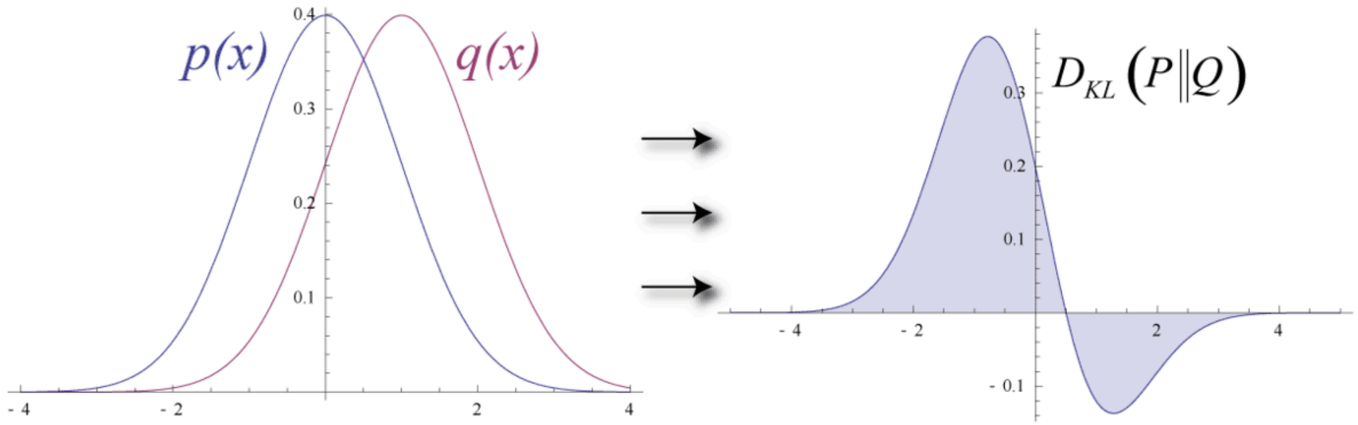
$$q_{j|i} = \frac{\exp\left(-||y_i - y_j||^2\right)}{\sum_{k \neq i} \exp\left(-||y_i - y_k||^2\right)},$$

En baja dimensión no hay sigma, el embedding debe ser coherente en escala para todos los puntos → se fija un σ global

Divergencia KL (Kullback-Leibler)

Determina en qué medida una distribución de probabilidad se desvía de otra distribución de referencia

$$\mathrm{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$



Divergencia KL en SNE

Determina en qué medida una distribución de probabilidad se desvía de otra distribución de referencia

$$\mathcal{C} = \sum_i \mathrm{KL}(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log\left(\frac{p_{j|i}}{q_{j|i}}\right)$$

Debemos minimizar este “costo” de forma que la diferencia entre ambas distribuciones sea la menor posible:

$$\frac{\partial \mathcal{C}}{\partial y_i} = 2 \sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}\right)\left(y_i - y_j\right)$$

Model Complexity

Entropía: Mide la **incertidumbre promedio** o la cantidad de información esperada en una distribución de probabilidad.

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

$$\text{Perp}(P_i) = 2^{H(P_i)}$$

```
tsne = TSNE(n_components=2, perplexity=30)
X_embedded = tsne.fit_transform(X_scaled)
```

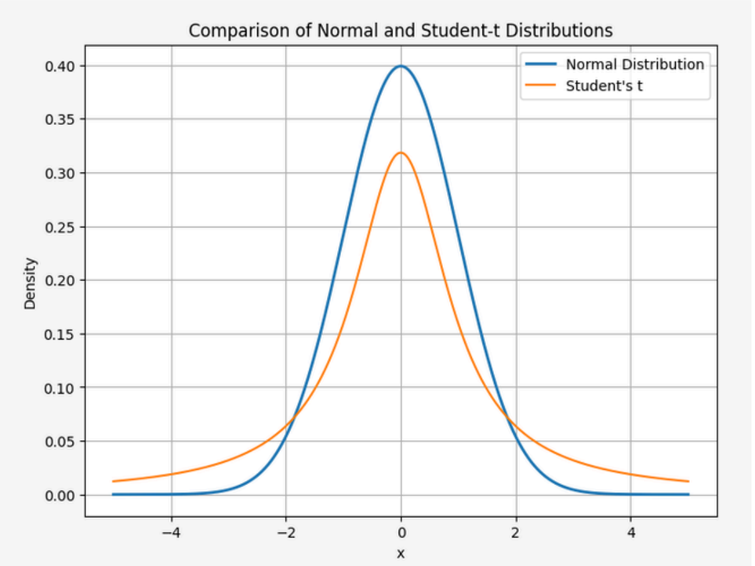
- Se fija una perplejidad objetivo **Perp > 0** (típicamente 5–50)
- Se halla cada σ_i tal que $\text{Perp}(P_i) \approx$ la perplejidad objetivo.

t-SNE

t-SNE (2008, van der Maaten & Hinton) nace como una mejora directa de SNE. Dos diferencias claves:

- En baja dimensión cambia a una **t Student** con 1 grado de libertad

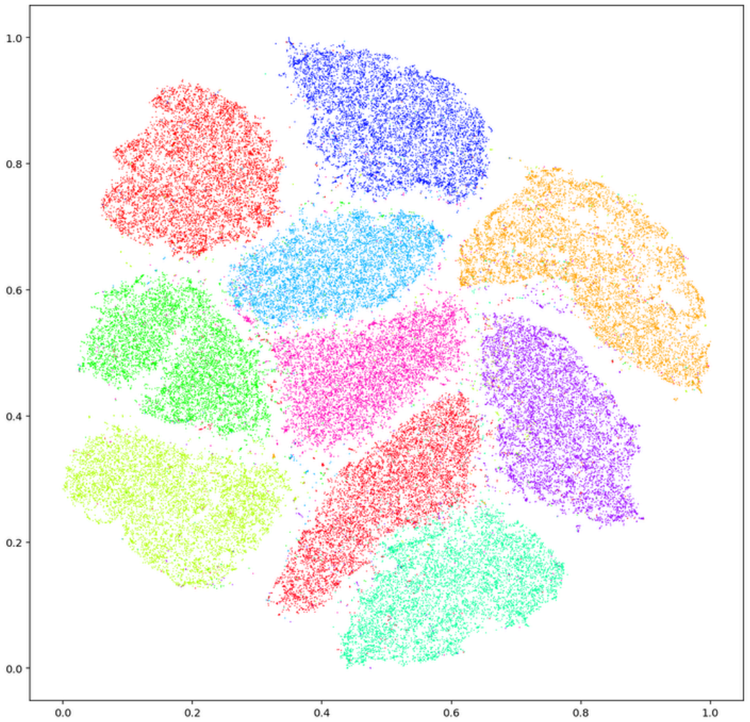
$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq \ell} (1 + \|y_k - y_\ell\|^2)^{-1}}$$



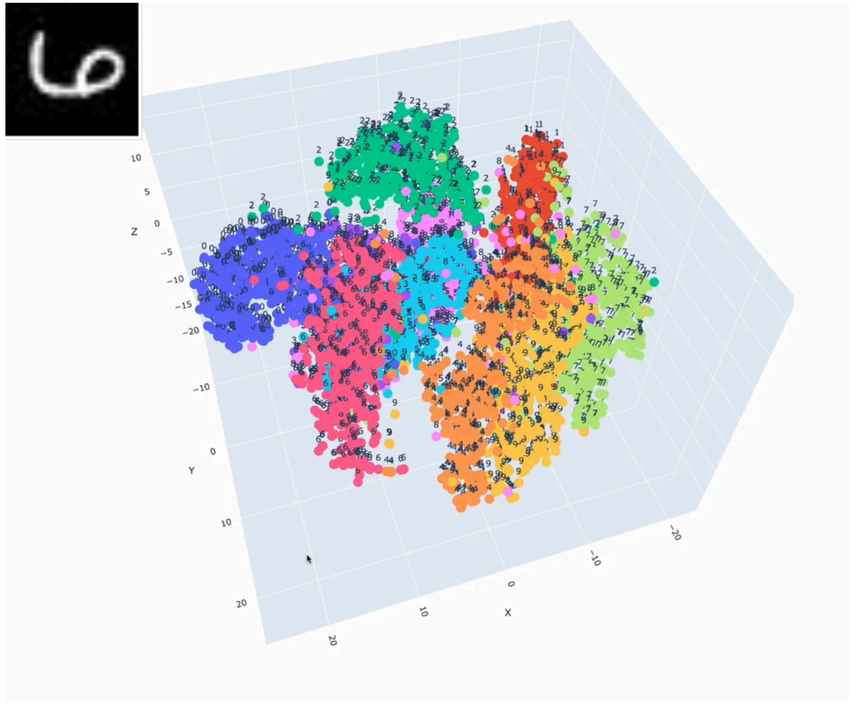
- Define distribuciones simétricas

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

t-SNE



t-SNE 2 dimensiones MNIST

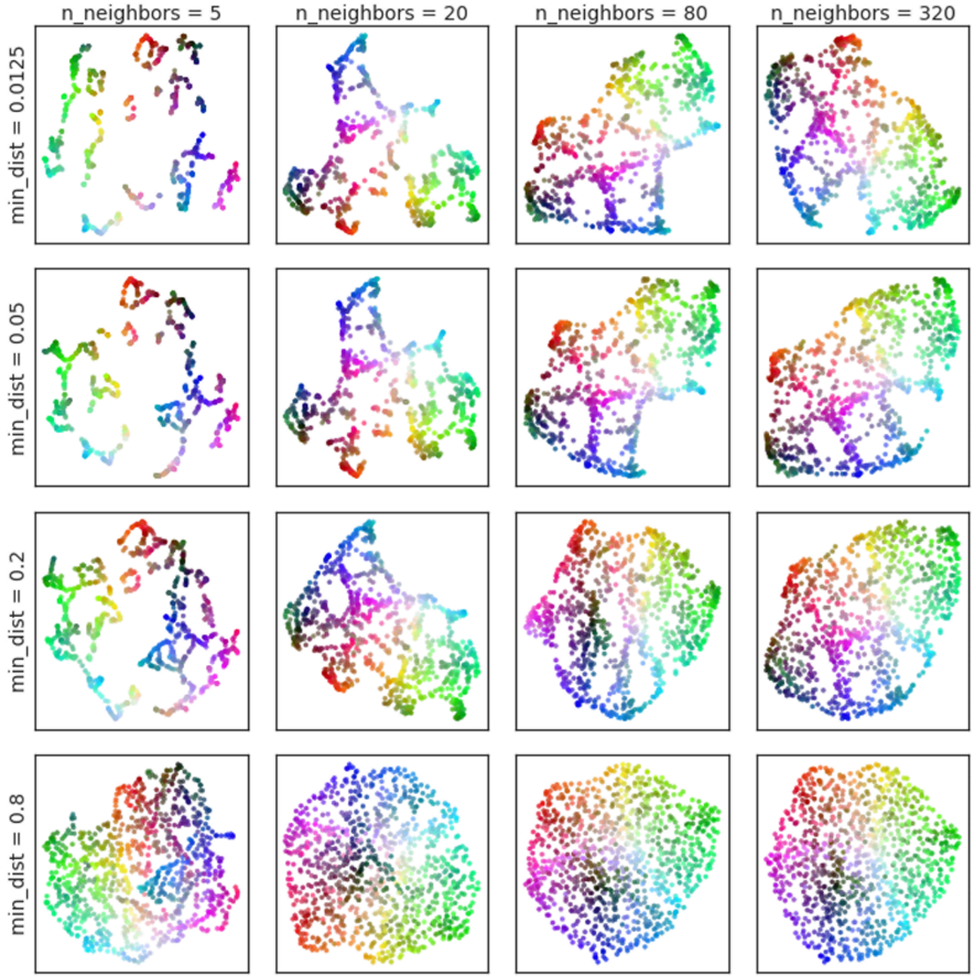


t-SNE 3 dimensiones MNIST

Uniform Manifold Approximation and Projection (UMAP)

Es un algoritmo de reducción de dimensionalidad parecido a t-SNE, pero con una teoría más sólida por debajo:

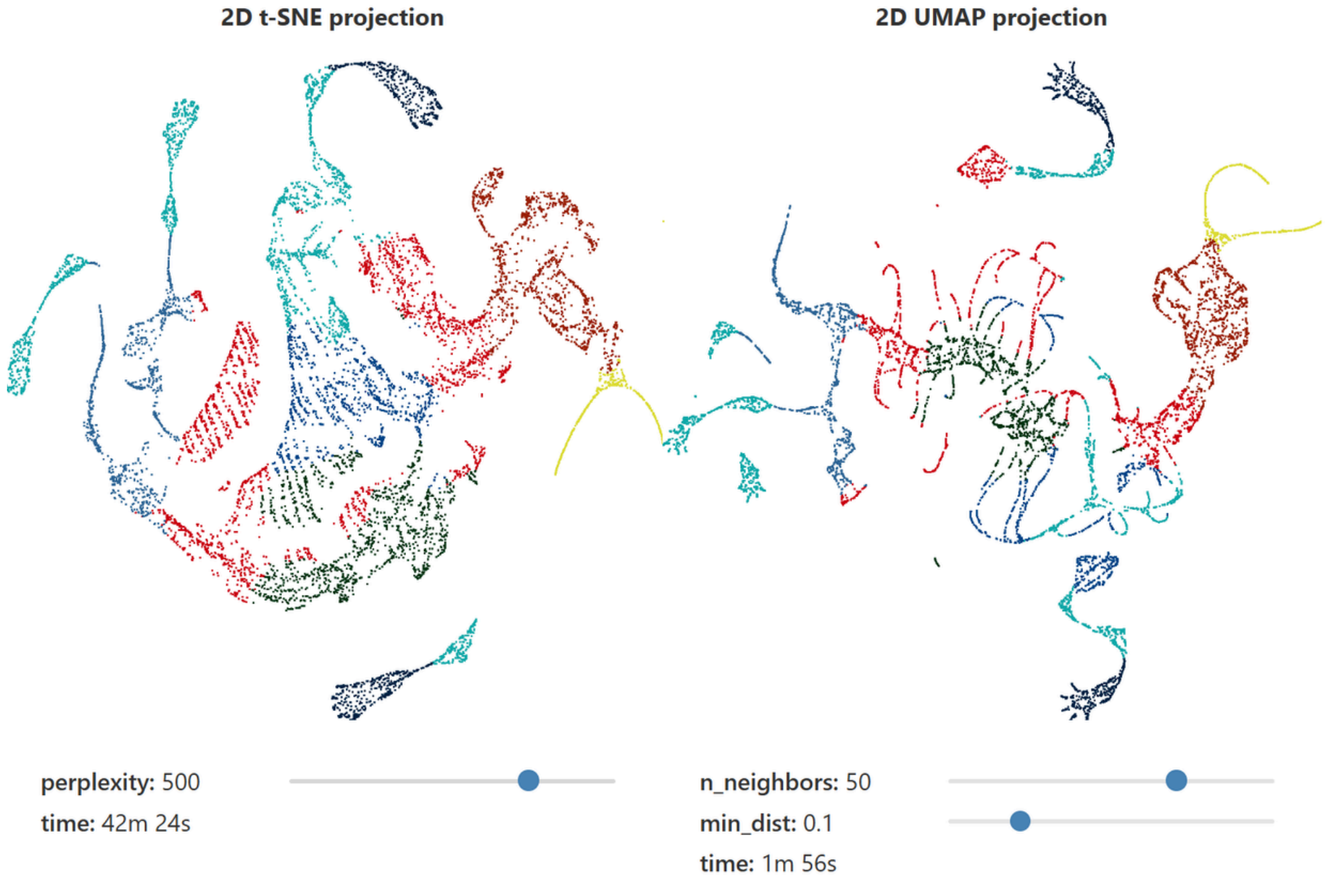
- **Supuesto de variedad:** los datos de alta dimensión X viven en una variedad (manifold) de dimensión mucho más baja dentro de $\mathbb{R}^D$.
- **Aproximación fuzzy-topológica:** la estructura de vecinos se representa con un grafo difuso (fuzzy graph) que codifica las relaciones de cercanía entre puntos.



Uniform Manifold Approximation and Projection (UMAP)

Diferencias con t-SNE:

- **Modelo teórico:** Se basa en geometría de variedades y topología algebraica
- **Estructura global:** Preserva mejor la forma global de los datos.
- **Parámetros clave:**
  - **n_neighbors:** vecinos considerados para aproximar métrica local
  - **min_dist:** separación entre puntos cercanos.

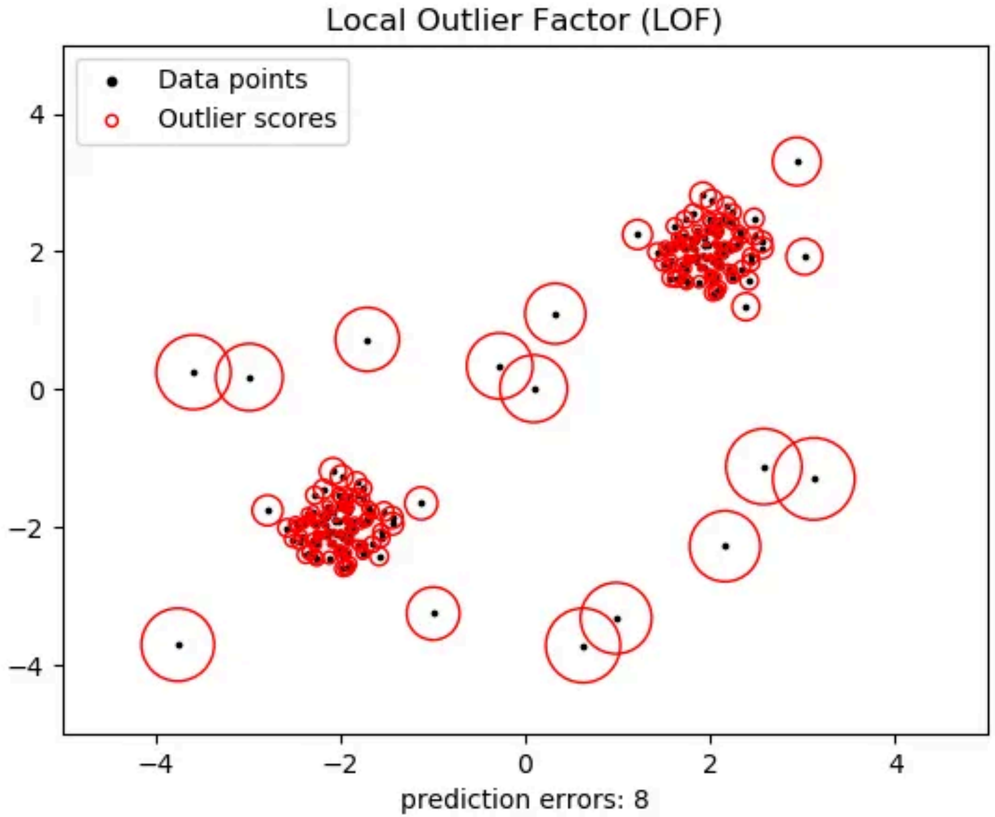


Local Outlier Factor

Local Outlier Factor (LOF)

El método LOF detecta outliers (valores atípicos) basado en densidad. Comparar la densidad local de un punto con la densidad de sus vecinos más cercanos:

- Si un punto está en una región mucho menos densa que la de sus vecinos, se considera un outlier.
- Si la densidad es similar a la de sus vecinos, se considera un punto normal.



Local Outlier Factor vs IQR

El método IQR de la primera clase es un método **global y unidimensional.** Mientras que LOF es **multidimensional y basado en densidad local**

IQR:

- Los datos son univariados (solo una variable)
- Método rápido, simple y robusto
- Los outliers son valores muy lejanos al rango central

LOF:

- Datos son multivariados
- Detectar outliers locales