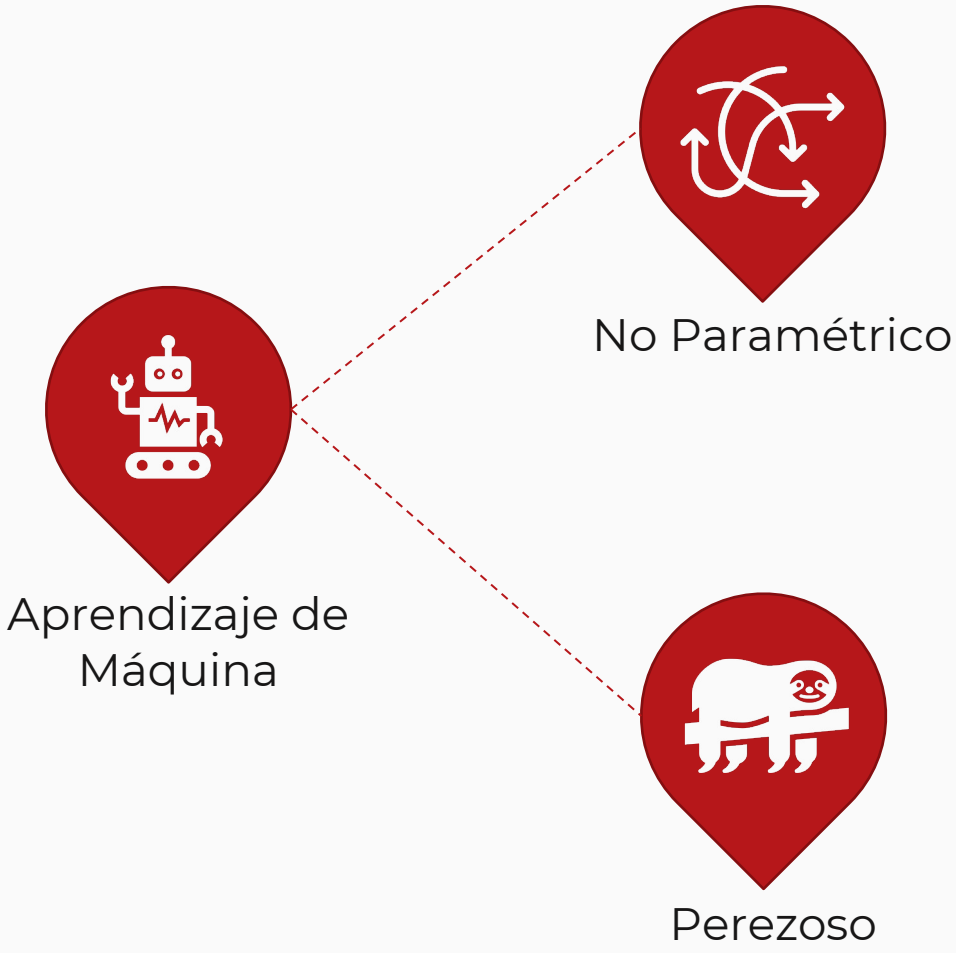
Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

K-Vecinos Cercanos

Mtro. René Rosado González
Director de Programa LTP

K-Vecinos Cercanos

K-Nearest Neighbors (K-NN)



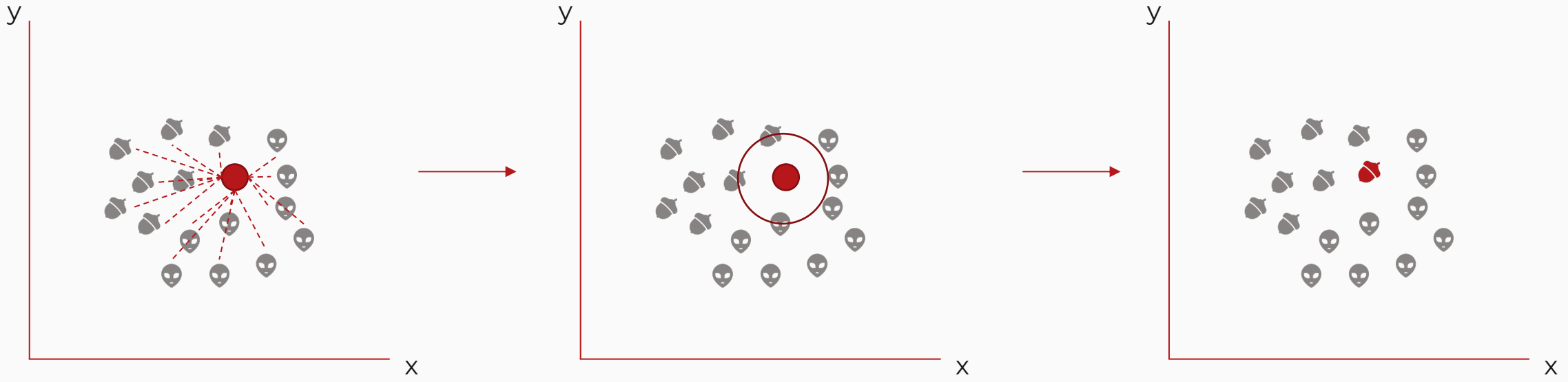
No hay suposiciones a priori sobre la relación entre las observaciones.

Todos los datos de entrenamiento son utilizados en la fase de prueba. Esto hace que el entrenamiento sea más rápido y la fase de prueba más lenta y costosa.

K-Vecinos Cercanos

K-Nearest Neighbors (K-NN)

1. Se calcula la distancia entre el nuevo punto y cada punto de entrenamiento.
2. Se seleccionan los k puntos de datos más cercanos (en función de la distancia).
3. El promedio o mayoría de estos puntos de datos es la predicción final para el nuevo punto.



Distancia Minkowski

Minkowski Distance

La distancia Minkowski es una métrica dentro de un espacio vectorial normado.

Una distancia Minkowski de orden p entre dos puntos está definida como:

$$D(X,Y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Donde p es un número entero

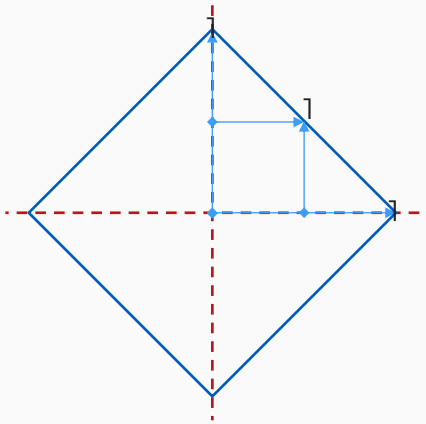
Distancia Minkowski

Minkowski Distance

Manhattan ($p = 1$)

$$\sum_{i=1}^{k} |x_i - y_i|$$

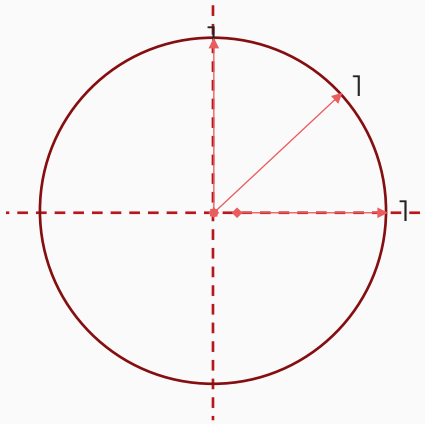
$$|5 - 1| + |4 - 1| = 7$$



Eculidiana ($p = 2$)

$$\sum_{i=1}^{k} \sqrt{(x_i - y_i)^2}$$

$$\sqrt{(5 - 1)^2 + (4 - 1)^2} = 5$$

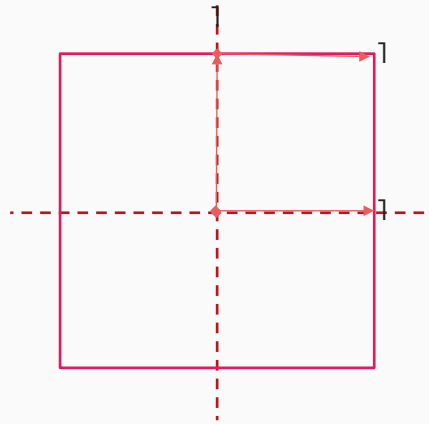


...

Chebyshev ($p = \infty$)

$$\max(|x_j - x_i|, |y_j - y_i|)$$

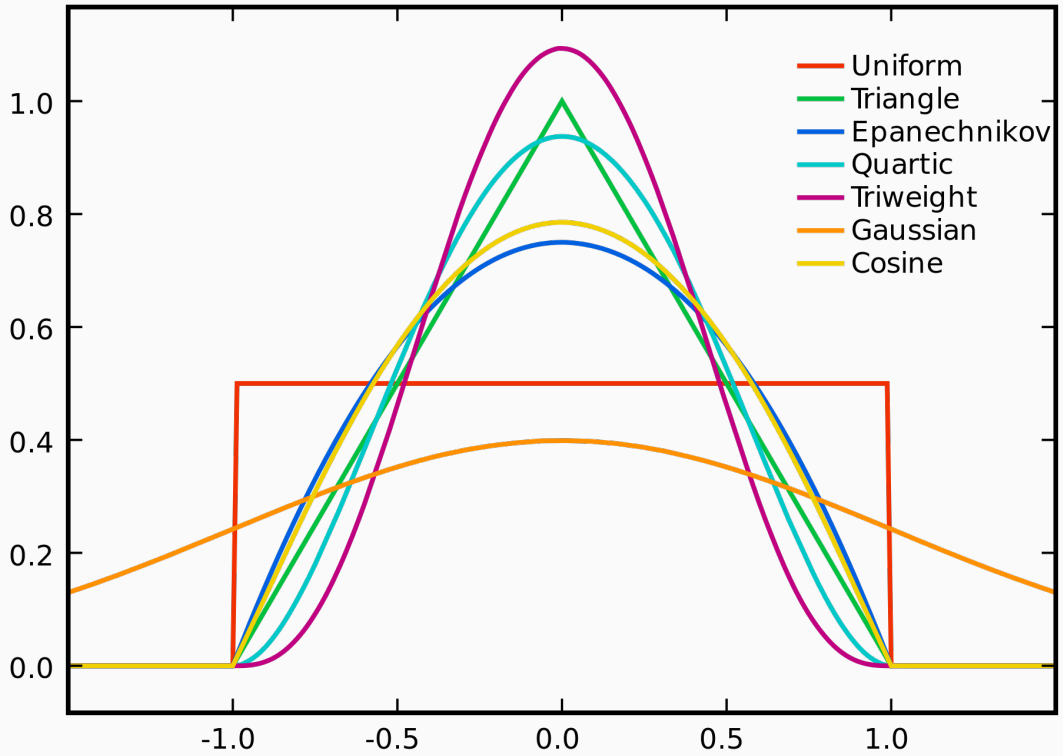
$$\max(5 - 1, 4 - 1) = 9$$



Funciones Núcleo

Kernel Functions

- Es una función de ponderación utilizada en técnicas de estimación no paramétricas.
- Se utilizan para estimar las funciones de densidad de las variables aleatorias o la expectativa condicional de una variable aleatoria.
- En el caso de knn nos sirve como ponderador de la distancia entre los puntos.



K-Vecinos Cercanos

Consideraciones

- Puede ser usado para regresión y clasificación.
- Funciona mejor con una cantidad menor de variables que con una gran cantidad de variables.
- El aumento de la dimensión también conduce al problema del sobreajuste.
- La investigación ha demostrado que en grandes dimensiones la distancia euclidiana ya no es útil.

Un ejemplo