

ESTUDIO COMPARATIVO DEL ALGORITMO K-MEANS PARA DISTINTAS DISTANCIAS EN ESPACIOS METRICOS

Trabajo Fin de Grado

Grado en Matemáticas

Autor: *Marcos Crespo Díaz*

Tutor: *María Jesús Algar Díaz*



Universidad
Rey Juan Carlos

Escuela Técnica Superior
Ingeniería Informática

Contenido



- 1 **Introducción**
 - Objetivos
 - Contexto del K-Means
 - El algoritmo K-Means
- 2 **Espacios métricos y distancias**
 - Espacios métricos
 - Funciones de distancia
- 3 **Estudio comparativo del K-Means**
 - Aspectos generales
 - Dataset simple
 - Dataset con outliers
 - MNIST
- 4 **Conclusiones**



Objetivos del TFG

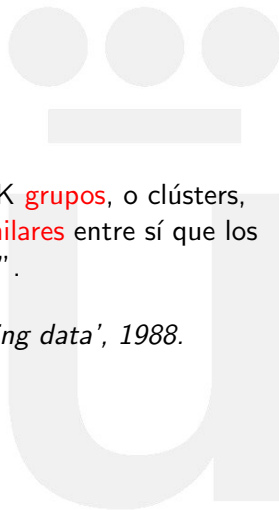
- Intuición y fundamentos matemáticos del *K-Means*.
- Introducción a los espacios métricos y funciones de distancia.
- Exponer el estudio comparativo en distintos espacios métricos.



El K-Means en el contexto de la ciencia de datos



K-Means

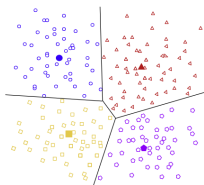


'... determinar una **partición** de los datos en K **grupos**, o clústers, tales que los datos de un clúster sean más **similares** entre sí que los datos de clústers diferentes”.

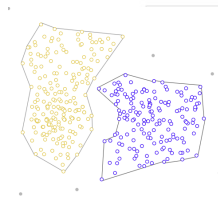
Jain and Dubes, 'Algorithms for clustering data', 1988.



K-Means



(a)



(b)

Figura: Diferentes formas de agrupar datos sobre el plano

K-Means. Centroide

Definition

Dado un conjunto de datos $\mathbf{C} \subset \mathbb{R}^n$, se llama **centroide** $\hat{c} \in \mathbb{R}^n$ al punto que satisface:

$$\hat{c} = \sum_{\mathbf{x} \in \mathbf{C}} \frac{\mathbf{x}}{|\mathbf{C}|} \quad (1)$$



K-Means. Problema

Podemos hablar de la varianza dentro de cada clúster C_k como

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^n (x_{ij} - x_{i'j})^2 \quad (2)$$

Queremos que dentro de cada clúster la varianza sea mínima.
Resolver el siguiente **problema de optimización**:

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K W(C_k) \right\} \quad (3)$$



El algoritmo de K-Means

- ❶ Inicializar C_i con $i \in \{1, \dots, k\}$ conjuntos de \mathbb{R}^n vacíos, que serán nuestros clústers
- ❷ Tomar k elementos de \mathbf{C} aleatoriamente como centroides \hat{c}_i con $i \in \{1, \dots, k\}$
- ❸ Calcular la $d(\hat{c}_i, \mathbf{x})^1 \forall \mathbf{x} \in \mathbf{C}, i \in \{1, \dots, k\}$.
- ❹ Para cada $\mathbf{x} \in \mathbf{C}$, asignar a \mathbf{x} al C_i cuya $d(\hat{c}_i, \mathbf{x})$ sea menor (elemento 'más similar')
- ❺ Iterar hasta que la asignación de clúster no cambie:
 - Para cada clúster C_i , recalcular su centroide.(1)
 - Asignar a cada $\mathbf{x} \in \mathbf{C}$, el clúster C_i cuya $d(\hat{c}_i, \mathbf{x})$ sea menor

¹ $d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$ (Distancia euclídea).



K-Means. Convergencia

Lemma

Sean $x^1, x^2, \dots, x^m \in \mathbb{R}^2$, con $m \geq 1$ puntos. Sea $\hat{c} = \frac{1}{m} \sum_{i=1}^m x^i$ su centroide, y sea $z \in \mathbb{R}^2$ un punto arbitrario en el espacio 2-dimensional. Entonces:

$$\sum_{i=1}^m \|x^i - z\|^2 \geq \sum_{i=1}^m \|x^i - \hat{c}\|^2.$$



Espacio Métrico

Definition

Sea X un conjunto no vacío y d una función de valor real definida sobre $X \times X$ tal que para $a, b \in X$:

- ❶ $d(a, b) \geq 0$, y $d(a, b) = 0$ si, y sólo si, $a = b$;
- ❷ $d(a, b) = d(b, a)$; y
- ❸ $d(a, c) \leq d(a, b) + d(b, c)$, para toda a, b y c en X (desigualdad triangular).

Entonces d es llamada métrica sobre X , (X, d) es llamado **espacio métrico** y $d(a, b)$ se conoce como la **distancia** entre a y b .^a

^aS. Morris, *Topología sin dolor*. 1989



Distancias

Sea $A = (a_1, a_2, \dots, a_n)$ y $B = (b_1, b_2, \dots, b_n)$ puntos de \mathbb{R}^n :

Definition

Se define la **distancia euclídea** como:

$$d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (4)$$

Definition

Se define la **distancia de Manhattan** como:

$$d(A, B) = |b_1 - a_1| + |b_2 - a_2| + \dots + |b_n - a_n| \quad (5)$$



Distancias

Sea $A = (a_1, a_2, \dots, a_n)$ y $B = (b_1, b_2, \dots, b_n)$ puntos de \mathbb{R}^n :

Definition

Se define la **distancia de Chebysev** como:

$$d(A, B) = \max(|b_1 - a_1|, |b_2 - a_2|, \dots, |b_n - a_n|) \quad (6)$$

Definition

Se define la **distancia de Minkowski** como:

$$d(A, B) = [|b_1 - a_1|^p + |b_2 - a_2|^p + \dots + |b_n - a_n|^p]^{1/p} \quad (7)$$



Aspectos generales

Realizaremos diferentes pruebas:

- 1 Base de datos simple
- 2 Base de datos simple con outliers
- 3 MNIST



Metodología

	Predicted Condition	
	Positive	Negative
	Actual Condition	
	<i>True Positive (TP)</i> <i>hit</i>	<i>False Negative (FN)</i> <i>Underestimation,</i> <i>type II error</i>
	<i>False Positive (FP)</i> <i>Overestimation,</i> <i>type I error</i>	<i>True Negative (TN)</i> <i>Correct rejection</i>

Cuadro: Matriz de confusión

$$precision = \frac{TP}{TP + FP} \quad (8)$$



Dataset simple. Resultados

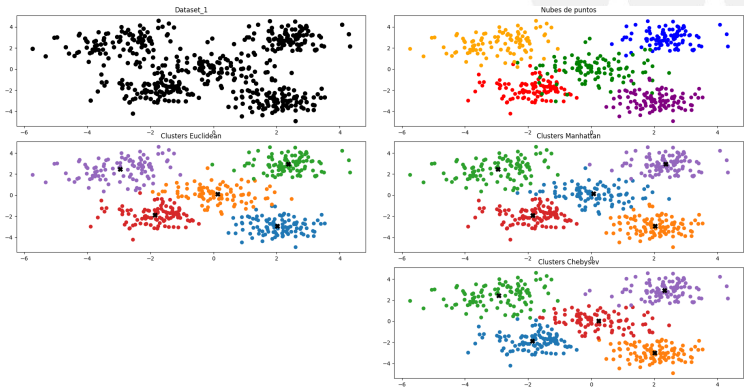


Figura: Resultados datos simples



Dataset con outliers. Resultados

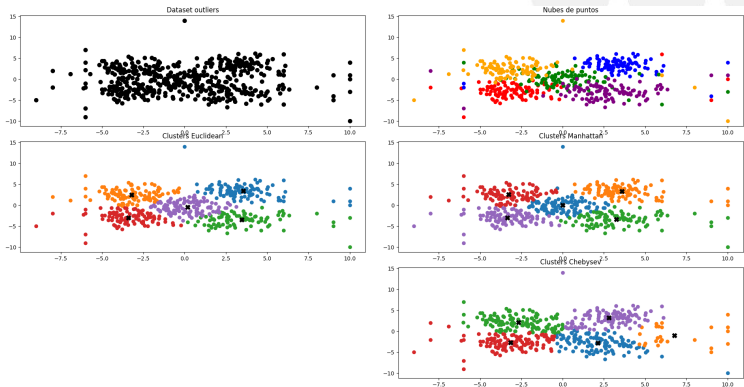


Figura: Resultados datos con atípicos



Datasets simples. Comparación

	Euclídea	Manhattan	Chebysev
Simple	95,4 %	95,6 %	95 %
Outliers	84,9 %	86,03 %	78,68 %

	Euclídea	Manhattan	Chebysev
Simple	10 it, 0.26s	6 it, 0.06s	8 it, 0.08s
Outliers	19 it, 0.47s	13 it, 0.17s	14 it, 0.19s

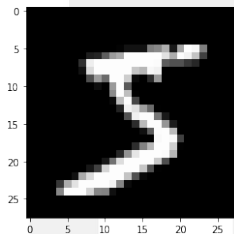


MNIST

MNIST (*Modified National Institute of Standards and Technology*)

```
label,1x1,1x2,1x3,1x4,1x5,1x6,1x7,1x8,  
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

(a)



(b)

Figura: Base de datos MNIST.

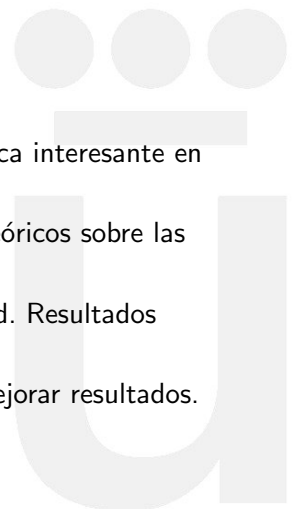


MNIST. Resultados

	Euclídea	Manhattan	Chebysev
precisión	53,24 %	39,115 %	30,88 %
ejecución	4 it, 723s	4 it, 511s	4 it, 564s



Conclusiones

- 
- 1 Espacios métricos: Herramienta topológica interesante en ciencia de datos.
 - 2 Verificación de la mayoría de aspectos teóricos sobre las distancias.
 - 3 MNIST: Espacio de gran dimensionalidad. Resultados aceptables.
 - 4 Posibles mejoras y futuras líneas para mejorar resultados.



ESTUDIO COMPARATIVO DEL ALGORITMO K-MEANS PARA DISTINTAS DISTANCIAS EN ESPACIOS METRICOS

Trabajo Fin de Grado

Grado en Matemáticas – Curso 2022-2023

Autor: *Marcos Crespo Díaz*

Tutor: *María Jesús Algar Díaz*



Universidad
Rey Juan Carlos

Escuela Técnica Superior
Ingeniería Informática