ANÁLISIS MULTIVARIANTE DE DATOS

Métodos de Clasificación y Reducción de datos.
Cluster

Presentación del problema

Manejo de los datos y su representación

Entomólogo
Darwin
Obtención y medición de características.
Agrupar según las características semejantes.

Presentación del problema ¿Qué es agrupar?

$Sea\ S\ un\ conjunto\ de\ n\ datos: \quad S = \{\overline{\boldsymbol{x}}_i\}\ 1 \leq i \leq n,$

$cada\ dato\ tiene\ "l"\ características\ de\ \overline{\boldsymbol{x}}_i = (x_1, x_2, x_3 \ldots x_l)$

$se\ busca\ encontrar\ k\ subconjuntos\ S_k\ tales\ que$

$S_i \cap S_j = \emptyset\ para \quad i \neq j\ \ y\ \ S = \bigcup_{1 \leq i \leq k} S_i \quad \text{ó} \quad S = \bigcup_{i=1}^{k} S_i$

$con\ la\ condición\ que\ los\ datos\ de\ cada\ subconjunto\ \ S_j$

$sean\ parecidos\ entre\ si.$

Manejo de los datos y su representación

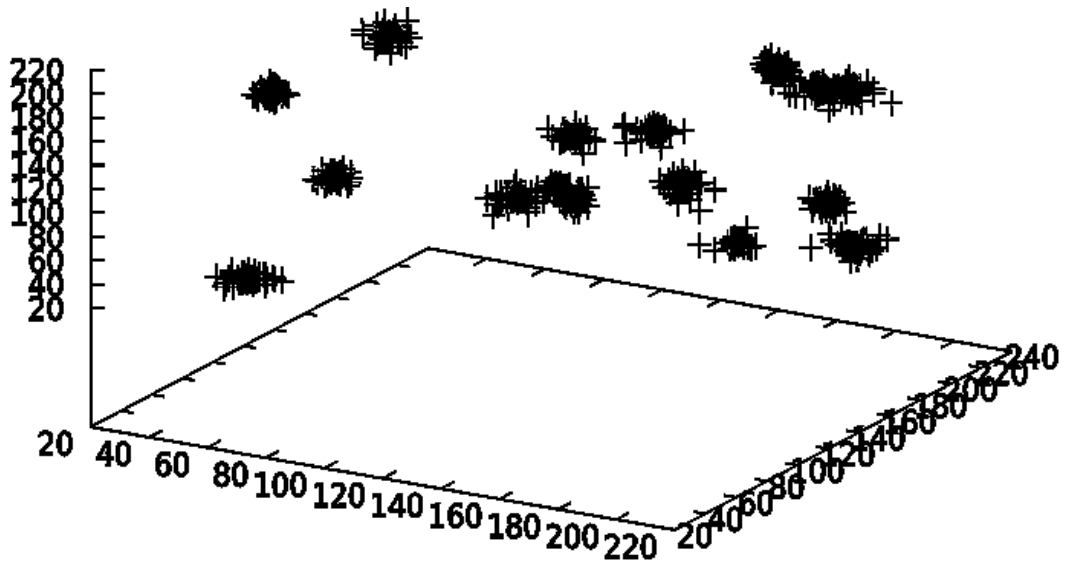
Obtención y Medición de varias características,

Representación. Espacio de características.

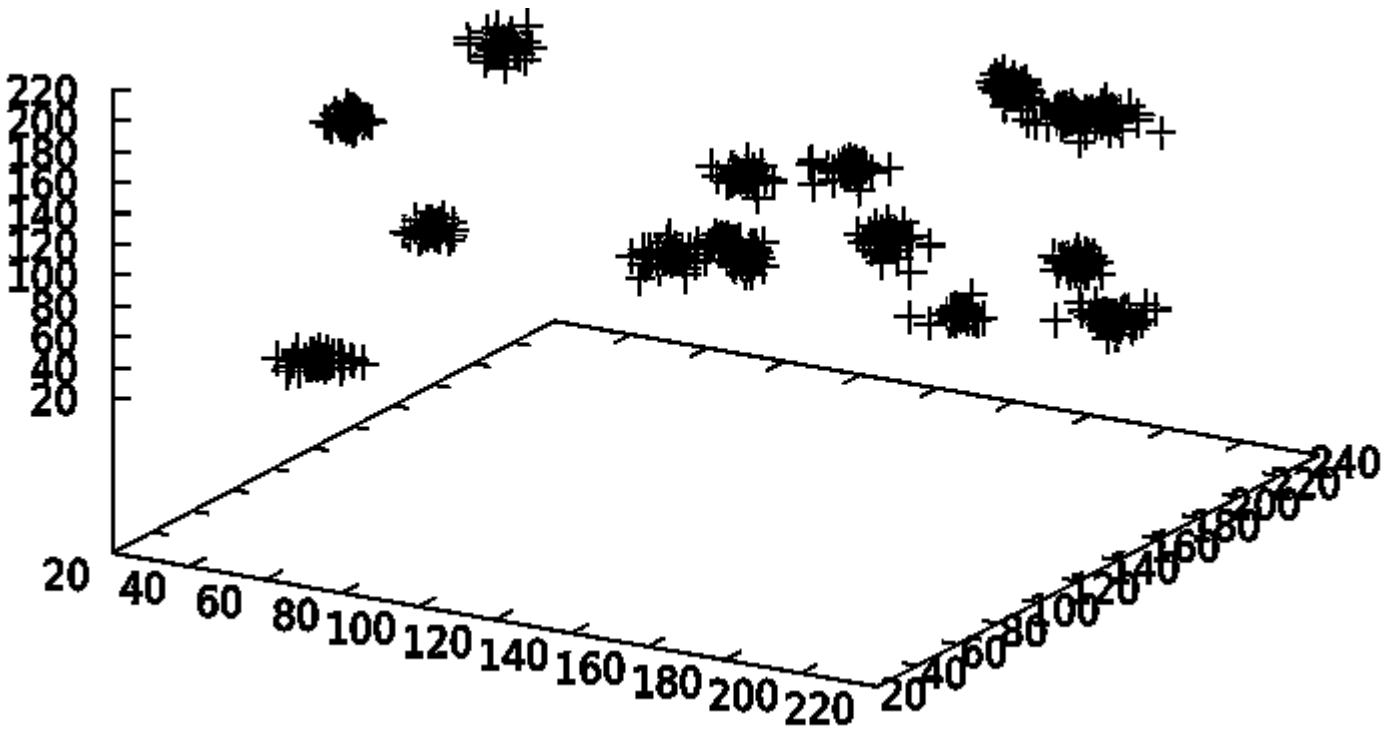
Vector de Características.

Los grupos están próximos entre si.

Medida de similaridad.



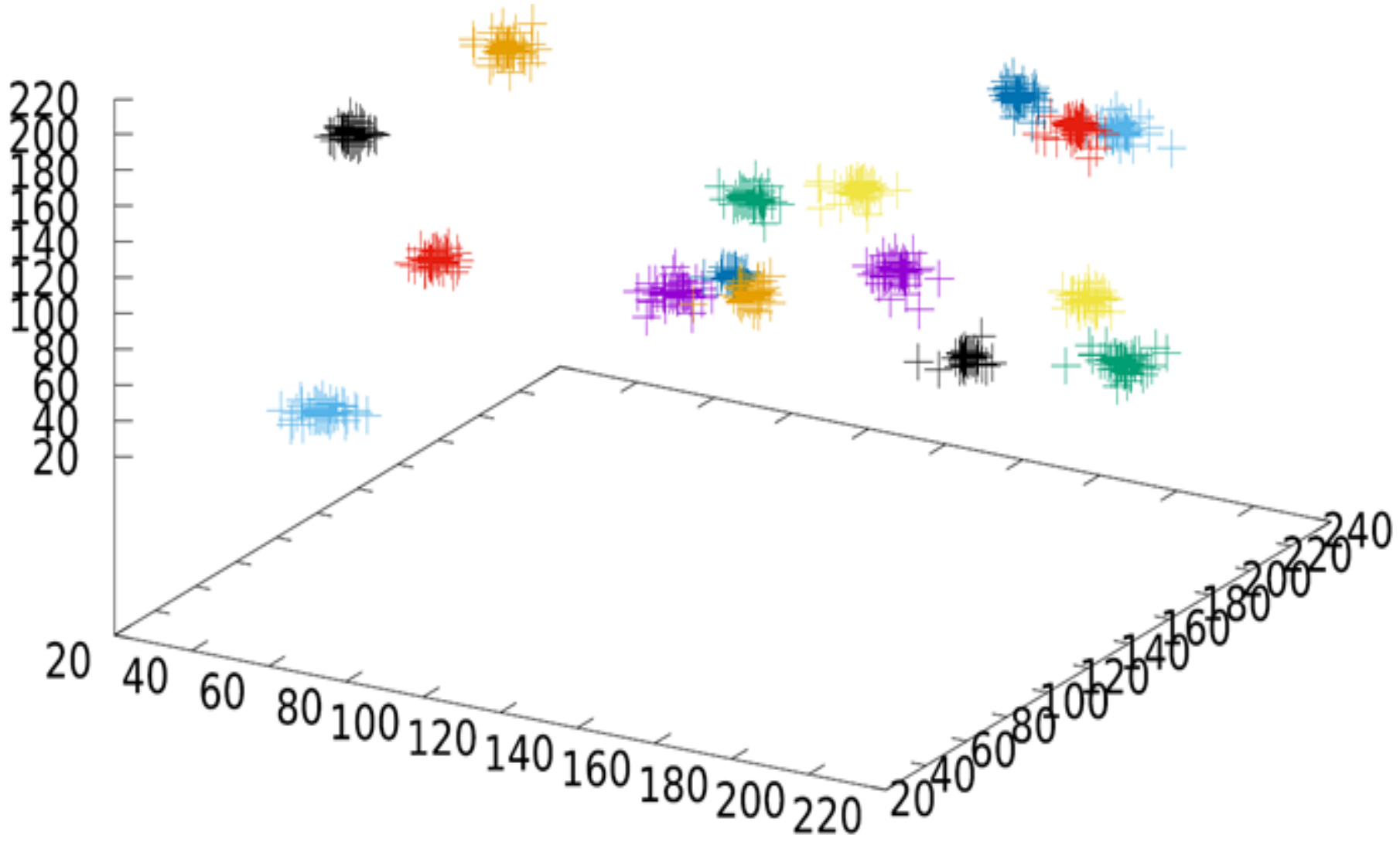
VARIABLES ALEATORIAS VARIAS CARACTERÍSTICAS



Sea un conjunto de datos con tres características.
Cada una de ellas representadas por números naturales entre 0 y 255.
Se los quiere agrupar por semejanza de sus Características.
El problema entonces es encontrar las agrupaciones Naturales por semejanza de características.
O sea clusterizar.

Ejemplo

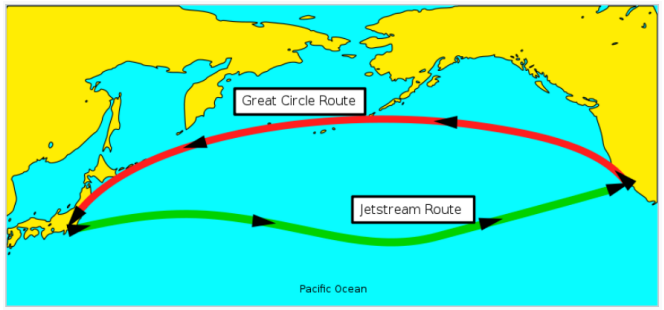
En una representación tridimensional de, por ejemplo un conjunto de datos que tienen tres colores, tres características, podríamos obtener un gráfico parecido a este.



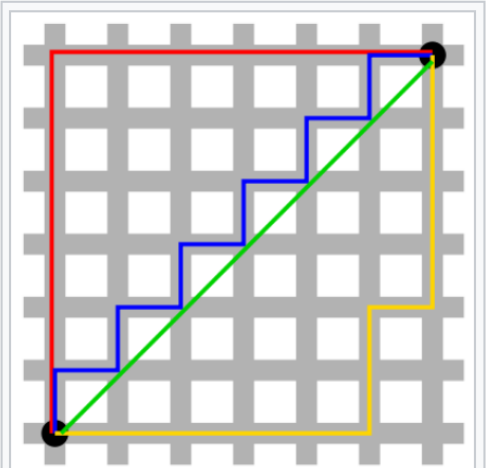
El problema es, dado un conjunto de datos a los que se les mide cierta cantidad de cualidades,
Encontrar subconjuntos disjuntos con objetos parecidos entre sí, con el total de los datos.

Manejo de los datos y su representación

¿Cómo se mide la agrupación.
Medida de similaridad.
Medidas de dis-similaridad.
Distancias
Euclidea, Manhattan, Chebyshev o Minkowski en general,
Mahalanobis, Jaccard , Spearman, Hamming, etc.



Airline routes between Los Angeles and Tokyo approximately follow a direct great circle route (top), but use the jet stream (bottom) when heading eastwards. Note that the shortest route appears as a curve rather than a straight line because this map is a Mercator projection, which does not scale all distances equally compared to the real spherical surface of the Earth.



"Manhattan distance" on a grid

Distancias

Distancias
Manhattan (norma 1),
Euclidea (norma 2),
Minkowski (norma p),
Chebyshev (norma ∞),

Mahalanobis , Jaccard , Spearman, Hamming, etc.

$$D_{Ch}(\bar{X}, \bar{Y}) = \max_{i=1}^{d} |x_i - y_i|$$

For a point $(x_1, x_2, ..., x_n)$ and a point $(y_1, y_2, ..., y_n)$, the **Minkowski distance** of order p (**p-norm distance**) is defined as:

$$\text{1-norm distance} = \sum_{i=1}^{n} |x_i - y_i|$$

$$\text{2-norm distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$$

$$\text{p-norm distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

$$\text{infinity norm distance} = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = \max \left(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n| \right).$$

p need not be an integer, but it cannot be less than 1, because otherwise the triangle inequality does not hold.

Distancias

Datos Normalizados.

Para poder utilizar una norma o distancia, es necesario que todas las coordenadas tengan igual peso, es así que cada característica tiene que ser normalizada. Si no no tiene sentido aplicar, por ejemplo la suma de los cuadrados de la diferencia entre cada característica (Euclideana)

Matriz de similaridad o similitud **S**.

Dada una métrica, la relación entre dos objetos cualesquiera se puede considerar como la distancia en el espacio de las características entre sus respectivos vectores.

La matriz, puede ser $S \in R^{n*n}$ o ser $S \in R^{n*m}$

o ser $S \in N^{n*n}$

o ser $S \in [0;1]^{n*n}$

o ser $S \in [0;256]^{n*n}$ etc.

Los métodos de agrupamiento se pueden clasifiar de distintas formas.

Machine Learning

Machine Learning (ML) o aprendizaje automático es una disciplina de dentro de la Inteligencia Artificial que crea sistemas que le permiten a una computadora aprender automáticamente es decir, que le permite a la máquina predecir comportamiento futuros a partir del estudio de datos.
El término se viene usando desde los años 50’s pero es en los últimos tiempos en que más se ha desarrollado gracias a la capacidad de los computadores y a la gran cantidad de datos que hay para procesar.
Los algoritmos usados en ML pueden subdividirse en dos grandes grupos, Supervisadas (supervised learning SL) y No Supervisadas (unsupervised learning UL).
En los primeros SL, se aprovecha un conocimiento previo de los datos para obtener los resultados, en los segundos UL, no, el sistema trata de buscar patrones que permitan agrupar a los datos.

Machine Learning

Las técnicas de Machine Learning se pueden separar en tres.

1) Clasificación.

2) Regresión.

3) Clustering.

En las primeras dos se trata de hallar una función que asigne etiquetas de pertenencia a las observaciones.
Es importante entrenar a la función para la asignación de etiquetas.
Por esto estas técnicas son Supervisadas.
Las técnicas Supervisadas son costosas tanto para máquinas como para humanos. (Entrenamiento del humano, entrenamiento de la máquina)

Las técnicas No Supervisadas no necesitan asignar etiquetas, Clústering encuentra grupos que son similares.

Machine Learning

- **Técnicas Supervisadas (Supervised Learning).**
  - Comparar las etiquetas halladas con las predichas.
  - Las Predicciones deberán ser parecidas a las reales.
- **Técnicas No Supervisadas (Unsupervised Learning).**
  - Es más complejo porque no hay etiquetas reales para comparar.
- **Técnicas Semi-Supervisadas**
  - En la práctica nos podemos hallar con un conjunto de objetos no etiquetados
  - Y otro conjunto de datos etiquetados.
  - Se puede agrupar por métodos de clustering. Y con otras observaciones etiquetadas conocemos la performance del método.

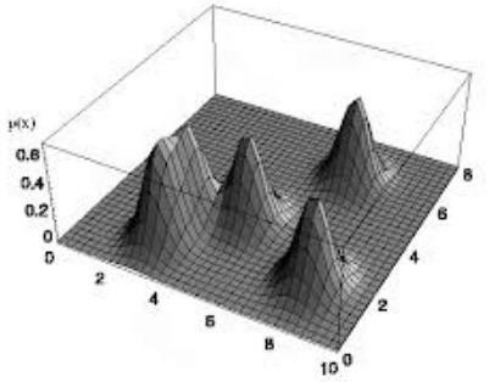
Indices de Performance del Modelo

Cualquiera sea el método de agrupamiento, se puede medir su performance ya sea comparándolo con datos reales o no.

Se define **Eficiencia (Accuracy) y Error**.

$$Eficiencia = \frac{Instancias\_correctamente\_clasificadas}{Total\_de\_instancias\_clasificadas}$$

$$Error = 1 - Eficiencia$$



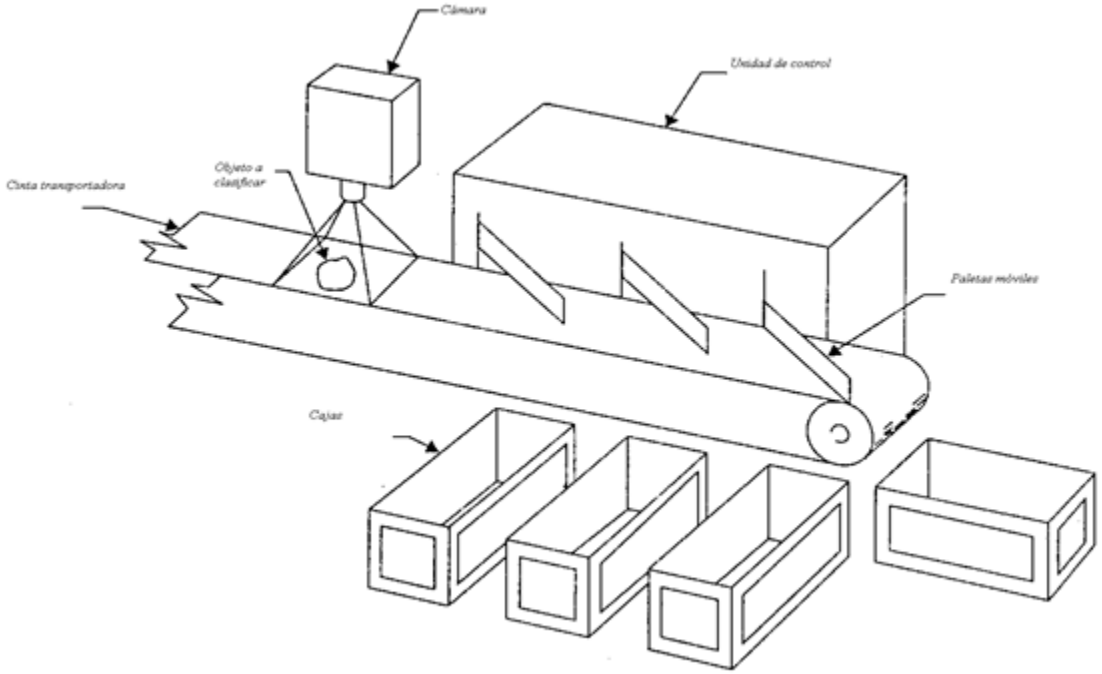
Problema de Clasificación

Tenemos una cinta transportadora sobre la cual un sistema mide características de objetos y un brazo actúa separándolos.

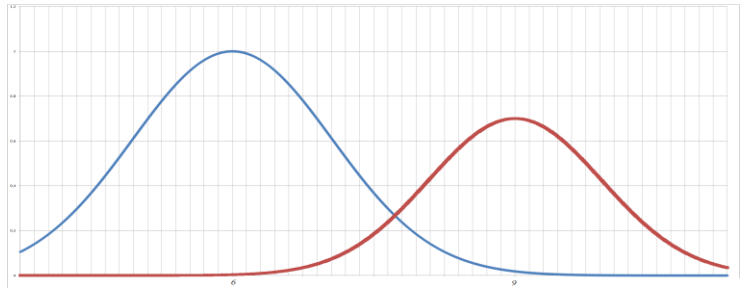
Implementamos un algoritmo que toma las decisiones.

¿Cómo medimos la performance del Método?

Una manera es haciendo pasar muchos objetos y ver cómo responde.

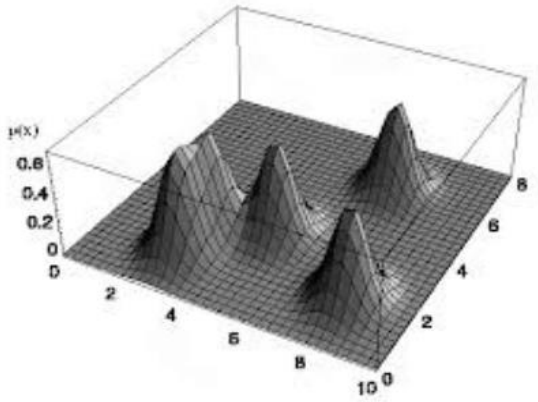


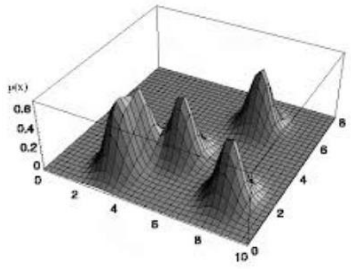
Sistema Clasificador de frutas

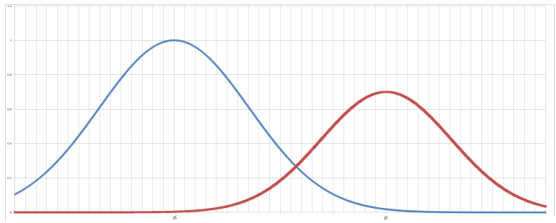


Indices de Performance del Modelo

A manera de ejemplo supongamos que estamos clasificando naranjas y mandarinas con una máquina.
Como conocimiento previo que usamos el diámetro y el grado de rojo de las frutas.
Sabemos que el valor esperado del diámetro de las naranjas es de (9 ± 3) cm y que el de las mandarinas es de (6 ± 3) cm.
Además sabemos que el rojo esperado de las naranjas es (100 ±20) y el de las mandarinas es (150 ±20).
Normalizando estos valores tendremos todos los datos en el mismo rango de valores (por ej [0;1].

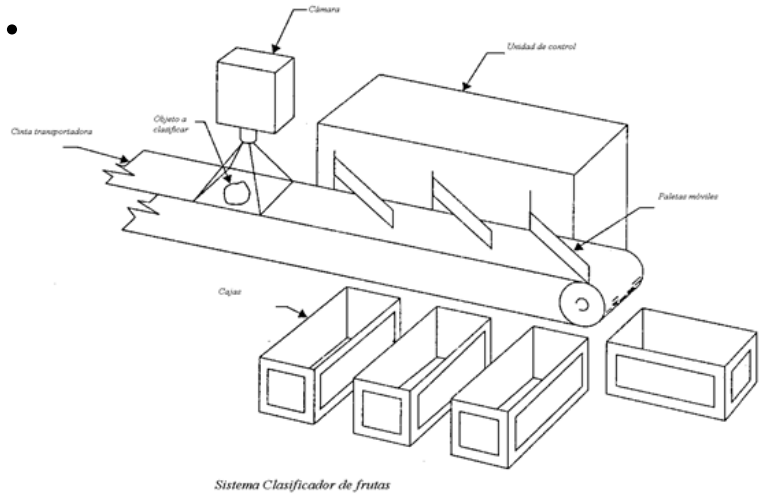






Indices de Performance del Modelo

Supongamos que clasificamos cien frutas, y que 60 salen caen en el cajón de naranjas y 40 en el de mandarinas.
Una manera de saber como funcionó el algoritmo es revisar las 60 naranjas y las 40 mandarinas estudiar los aciertos y dar la eficiencia del método.
Supongamos que el sistema acertó en 48 naranjas y 32 mandarinas, el total de datos acertados es (48+32), así la eficiencia del método es (48+32)/100= 0,8 un 80%.
No siempre este número es útil.



Sistema Clasificador de frutas

Indices de Performance del Modelo

Por eso se estudia la Matriz de Confusión.

La predicción de naranjas fue de 60 casos positivos y de 40 negativos, pero en realidad los verdaderos positivos fueron 48 y los falsos positivos fueron 12.Y si el algoritmo acertó 32 mandarinas (falso para naranjas) hubo 8 verdaderas naranjas entre las predicciones negativas o sea hubo 8 falsos negativos.

La Matriz de Confusión queda como
Verdaderos Positivos: 48
Verdaderos Negativos: 32

		Predicción	
		60	40
Veracidad	Naranjas Verdaderas	48	8
	Falsas Naranjas Mandarinas Verdaderas	12	32

Indices de Performance del Modelo

Matriz de Confusión

VP, VN, FP y FN

V+, V-, F+ y F-

		Predicción	
		P	N
Veracidad	V	VP	FN
	F	FP	VN

indican la frecuencia de aparición de cada una de estas situaciones (VP, V+ : verdadero positivo)

Indices de Performance del Modelo

Matriz de Confusión

		Predicción	
		P	N
Verdadero o Falso	V	VP	FN
	F	FP	VN

$$Precisión = \frac{VP}{VP + FP}$$

La precisión en la clasificación de naranjas es del 48/(48+12)=0,8 un 80%

		Predicción	
		60	40
Veracidad	V	48	8
	F	12	32

$$Recall = \frac{VP}{VP + FN}$$

El recall en la clasificación de naranjas será de 48/(48+8)=0,857 un 85,7%

Índices de Performance del Modelo

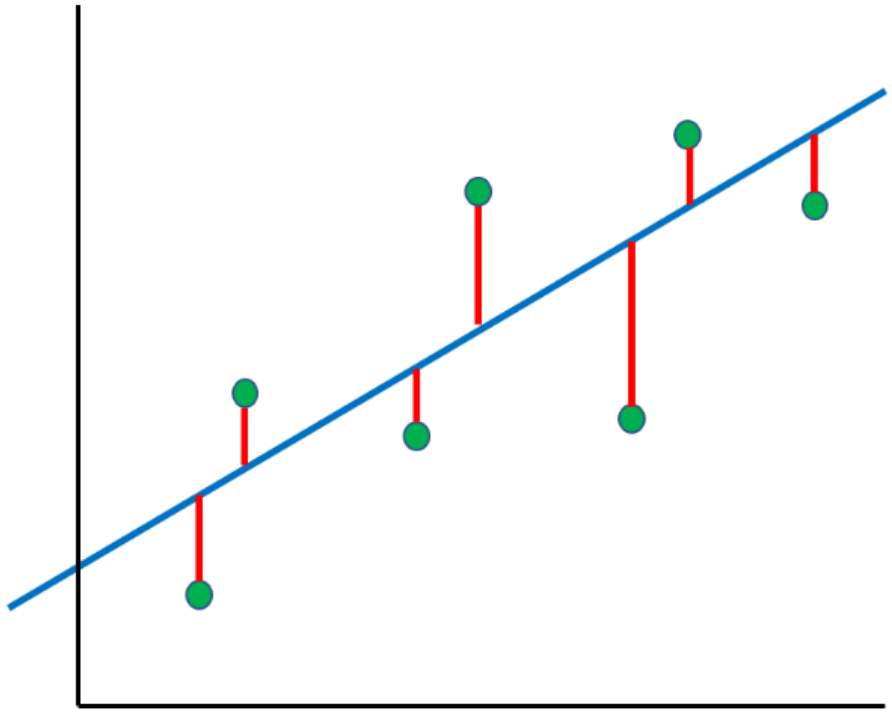
Regresión: Error Cuadrático Medio (RMS)

$$Error\_Cuadrático\_Medio = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (y_i - \bar{y}_i)^2}$$

$y_i : i - ésima\_observación$

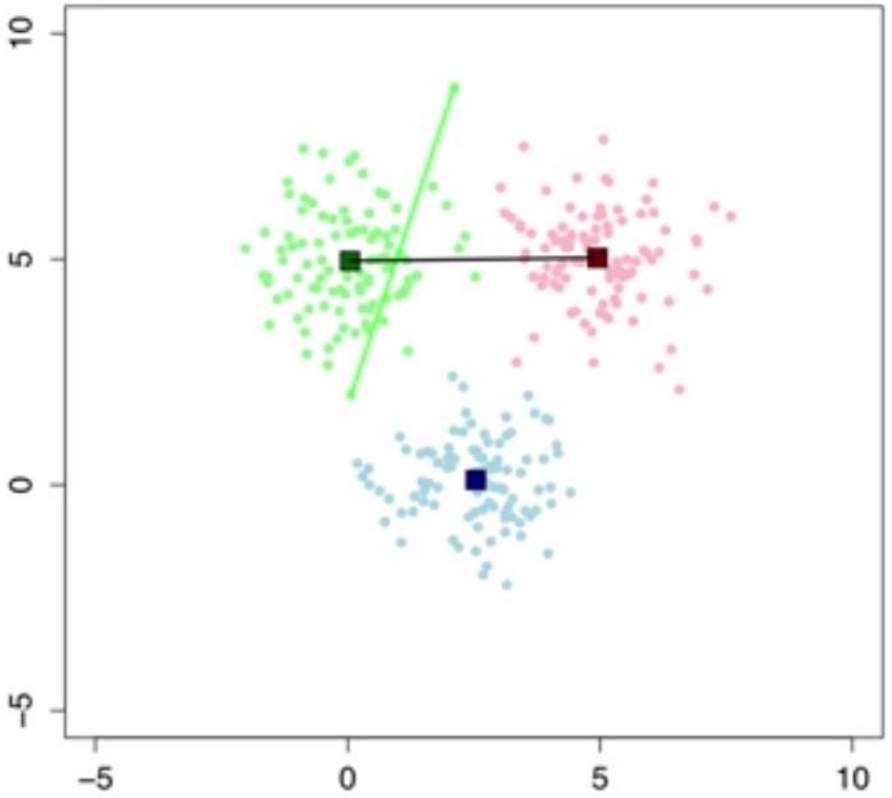
$\bar{y}_i : Valor\_predicho\_por\_la\_regresión\_para\_i$

$N : Cantidad\_de\_Observaciones$



Indices de Performance del Modelo

Índice de DUNN



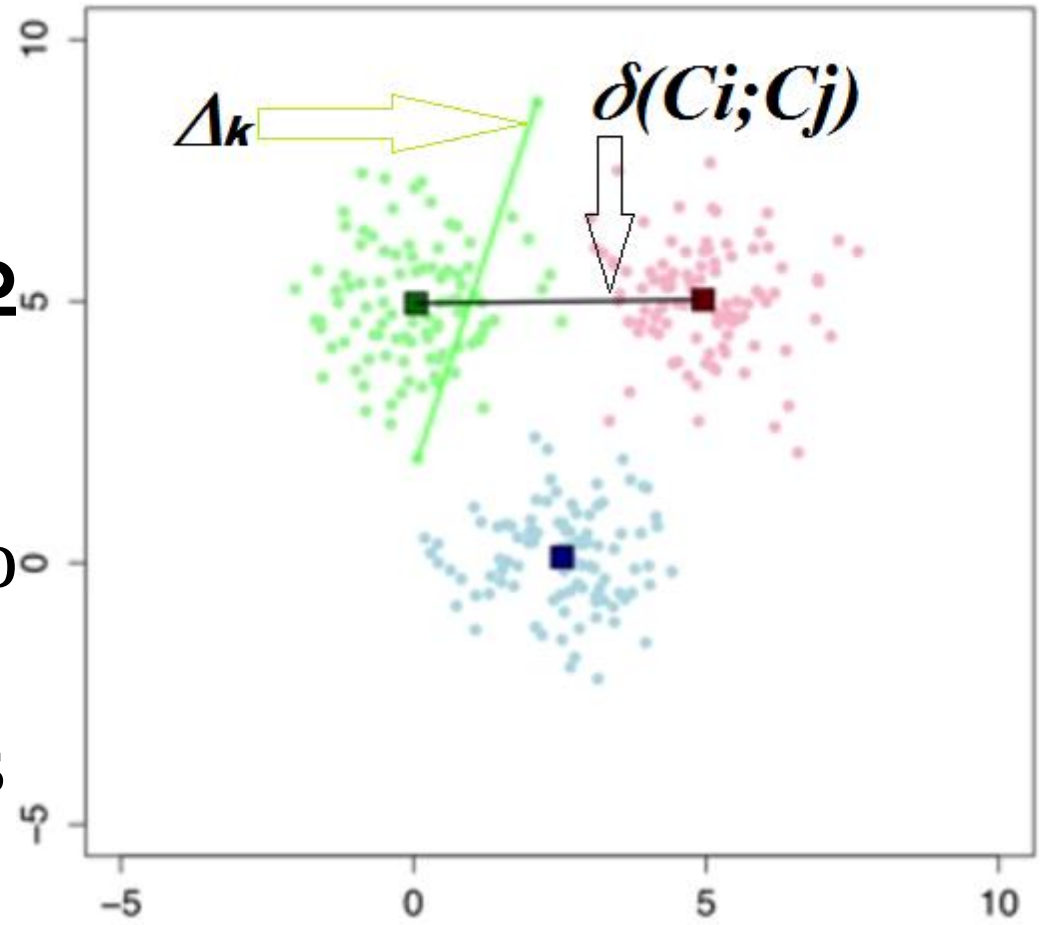
Este índice se usa para evaluar los algoritmos de clústering.

Se basa en los datos aportados por el resultado de la clusterización sin compararlo con datos reales.
Identifica cuán compactos son los clusters y cuán bien separados están.
Cuanto mayor sea el índice de Dunn, mejor es la clusterización.

Indices de Performance del Modelo

Índice de DUNN

Estudiemos los puntos dentro del espacio de características. Sea max Δ_k la mayor distancia entre la posición de dos objetos pertenecientes a un mismo cluster.



Sea C_i el centroide del cluster (i) llamamos min $\delta(C_i;C_j)$ a la menor distancia entre clusters.

$$Índice\_de\_Dunn = \frac{min\delta(C_i; C_j)}{max\Delta_k}$$

Otros métodos

Los métodos de clusterización pueden dividirse en tres grandes grupos:

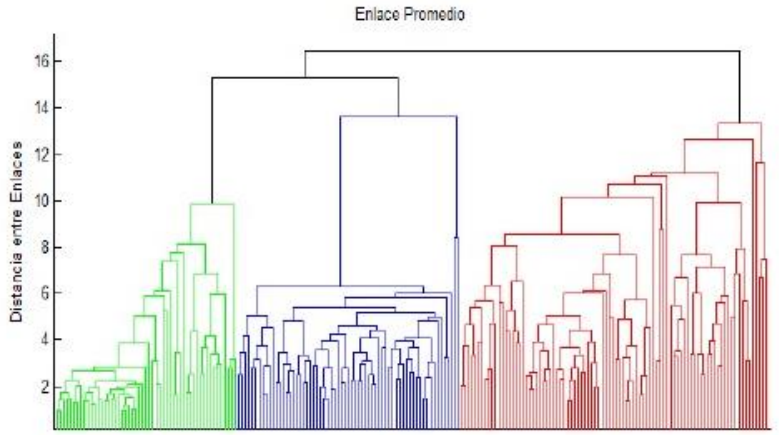
1) Clustering Jerárquico.

2) Clustering Basado en Particiones.

3) Clustering basado en rejillas (Grid Based).

A continuación se hará una breve descripción de los métodos mencionados.

Clustering Jerárquico



Los algoritmos de clusterización Jerárquica buscan en forma recursiva clusters anidados, ya sea en forma aglomerativa o en forma divisiva.

La forma aglomerativa empieza con cada punto en su propio cluster y busca los pares más parecidos entre sí. Los reemplaza por uno que represente a ambos y en el paso siguiente hace lo mismo hasta encontrar los clústeres jerárquicamente.

Es una estrategia Bottom-up.

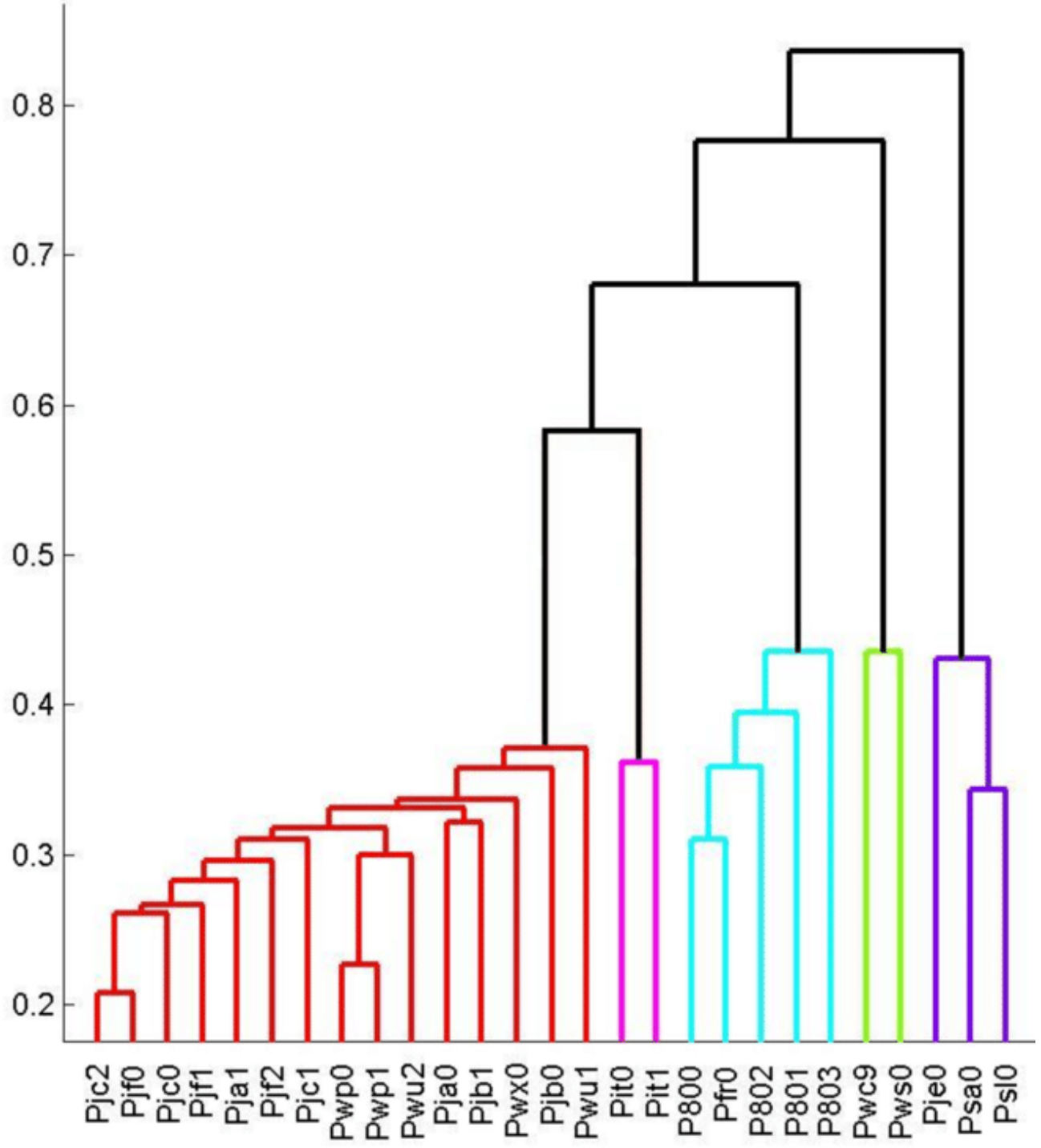
La forma divisiva comienza con todos los puntos en un solo cluster y divide recursivamente en más pequeños hasta encontrar los clústeres, conformando una estrategia top-down.

AGNES & DIANA.

CURE, GENIE & DENPEHC.

Clustering Jerárquico

En cualquiera de las estrategias, el principal problema es elegir el nivel en el cual se detiene el proceso.



Clustering Basado en Particiones.

Los algoritmos basados en particiones del espacio de características, encuentran todos los clústeres al mismo tiempo como una partición de los datos.

La partición está basada en funciones de similaridad o de dis-similaridad entre los puntos.

El método de partición más popular es K-Means y sus variantes. Otro método particional más moderno es AFFINITY PROPAGATION y sus variantes.

Clustering Basado en Particiones K-Means.

Minimiza el error cuadrático medio entre un punto medio empírico y los puntos de ese cluster.

EL acierto de K-Means es minimizar a suma del error cuadrático sobre todos los “k” clusters.

K-Means puede converger hacia un mínimo local, aunque estudios recientes muestran que puede converger hacia un mínimo global cuando los clusters están bien separados.

K-Means comienza con una partición inicial de “k” clusters y asigna cada uno de los puntos a uno de esos clusters minimizando el error cuadrático.

Como el error cuadrátoc siempre disminuye con el aumento de los clusters, sólo se puede buscar este mínimo fijando la cantidad de clusters.

Clustering Basado en Particiones K-Means.

Los pasos principales de K-Means son:

1) Inicializar el número de clusters, la función de distancia, el criterio de finalización de la iteración y el centroide de cada cluster. Los centroides pueden ser especificados manualmente o al azar.

2) Para cada punto, calcular su distancia (con la función de distancia elegida) a cada centroide y re-asignar ese punto al centroide más cercano.

3) Calcular el nuevo centroide de cada cluster como el centro de masa (centro geométrico) de dicho cluster.

4) Repetir los pasos 2 y 3 hasta que se cumpla la condición de finalización de iteraciones.

Clustering Basado en Particiones K-Means.

Las mayores ventajas de K-Means son:

1) Alcanza buenos resultados.
2) Es muy simple de implementar.
3) La complejidad espacial es O(N).

1) Las mayores desventajas son:

1) Depende fuertemente de los parámetros de inicialización.
2) Depende mucho de los outliers (Valores Atípicos).
3) La cantidad de clusters y el criterio de finalización debe ser definido previamente).
4) La complejidad de K-Means es O(n.k.l) siendo “n” la cantidad de puntos, “k” la cantidad de clases y “l” la cantidad de iteraciones.

Clustering Basado en Particiones K-Means.

Para mejorar K-Means se han creado muchos métodos. Ejemplos de estos son:

1) ISODATA.
2) FORGY.
3) FUZZY C-MEANS.
4) K-MEDIODS.
5) DBSCAN.
6) K-MEANS++.

Clustering Basado en Particiones
AFFINITY PROPAGATION (AP).

La idea principal de Affinity Propagation es localizar puntos que estén en el centro de concentraciones de puntos enel espacio de características.

Llama a estos puntos Ejemplares.

AP no depende de las inicializaciones ni requiere conocer de antemano la cantidad de clusters a encontrar.

El algoritmo trabajo pasando mensajes entre cada punto buscando información para determinar cuales son los puntos ejemplares.

La matriz de similitud S(n;n) con “n” la cantidad de puntos contiene los S(i;j) la información acerca de la similitud entre los puntos “i” y “j”. La diagonal se inicializa con ceros y se actualiza durante la ejecución. Se suele usar la menos el error cuadrático $S(i,j)= -(D_E(i;j))^2$ como medida de similitud.

Clustering Basado en Particiones
AFFINITY PROPAGATION (AP).

Para actualizar S(j;j) se usan otras dos matrices R(n;n) y A(n;n), la matriz de Resposabilidad y Afinidad.

R(i;j) cuantifica cuan adecuado resulta xj para ser ejemplar de xi (algo así como que xj se vende como ejemplar).

En tanto que A(i;j) cuantifica cuan apropiado resulta ser xi para elegir a xj para que sea su ejemplar (Aqui es xi quien elije a su posible ejemplar).

La complejidad espacial de este método es O(N2) haciendo muy dificultoso de ser usado en problemas de big-data.

Clustering basado en rejillas (Grid Based)

Los métodos basados en rejillas difieren de los demás algoritmos en que se enfocan en los valores en un entorno a los puntos y no en los puntos específicamente. En general, un clústering basado en rejillas consiste en los siguientes pasos:

1) Se crea la estructura de rejilla, por ejemplo una celda hipercubica de lado “d”
2) Se calcula la densidad para cada celda.
3) Se ordena de acuerdo a las densidades.
4) Se identifican los centroides de cada cluster.
5) Se revisa la vecindad de cada celda para incorporar, o no las celdas vecinas.

Clustering basado en rejillas (Grid Based)

Los métodos basados en rejillas también conocidos como métodos basados en densidad, son muy usados en grandes espacios multidimensioanles en los cuales los clusters son regiones más densas que sus regiones vecinas. Las mayores ventajas de los métodos basados en rejillas son su apreciable baja complejidad computacional, especialmente para conjuntos de datos muy grandes y su tolerancia a los outliers.

Ejemplos de algoritmos basados en regillas son STING y CLIQUE.

Uso de Histogramas en Clustering.

El uso de histogramas para clusterizar se basa en en dividir el espacio de características en bines o buckets (baldes). En cada bin, la distribución se supone que es uniforme. Y la agrupación de bines permite encontrar modas asociadas a los clusters.
A los histogramas en espacios multidimensionales se los ha llamado clásicamente como hiperhistogramas,
y se los ha construido proyectando los datos sobre cada dimensión (eje en dicho espacio) creando un histograma por dimensión o directamente particionando el espacio de características en hiperbines y usando algún método basado en rejillas para agrupar los hiperbines en clusters.
Para que el histograma muestre información que se pueda interpretar, es imprescindible hallar un adecuado tamaño del bin.