

# Descripción

Este operador realiza agrupación utilizando el algoritmo *k-means*. Agrupar grupos Ejemplos juntos que son similares entre sí. Como no es necesario ningún atributo *de etiqueta*, la agrupación en clústeres se puede utilizar con datos sin etiquetar y es un algoritmo de aprendizaje automático no supervisado.

El algoritmo k-means determina un conjunto de  $k$  grupos y asigna a cada ejemplo un grupo exacto. Los grupos constan de ejemplos similares. La similitud entre Ejemplos se basa en una medida de distancia entre ellos.

Un grupo en el algoritmo k-means está determinado por la posición del centro en el espacio  $n$ -dimensional de los  $n$  atributos del conjunto de ejemplos. Esta posición se llama centroide. Puede, pero no tiene por qué, ser la posición de un ejemplo de conjuntos de ejemplos.

El algoritmo k-medias comienza con  $k$  puntos que se tratan como el centroide de  $k$  grupos potenciales. Estos puntos de inicio son la posición de  $k$  ejemplos extraídos aleatoriamente del conjunto de ejemplos de entrada o están determinados por la heurística k-means++ si *determinar buenos valores iniciales* se establece en verdadero.

Todos los ejemplos se asignan a su grupo más cercano (el más cercano se define por el *tipo de medida*). A continuación, se recalculan los centroides de los grupos promediando todos los ejemplos de un grupo. Los pasos anteriores se repiten para los nuevos centroides hasta que los centroides ya no se mueven o se alcanzan *los pasos máximos de optimización*. Tenga en cuenta que no se garantiza que el algoritmo k-means converja si el tipo de medida no se basa en el cálculo de la distancia euclidiana (porque el recálculo de los centroides mediante el promedio supone un espacio euclidiano).

El procedimiento se repite en tiempos *máximos de ejecución*, cada vez con un conjunto diferente de puntos de inicio. Se entrega el conjunto de conglomerados que tiene la suma mínima de distancias al cuadrado de todos los ejemplos a sus centroides correspondientes.

## Diferenciación



### k-Medoides

En el caso del algoritmo k-medoides, el centroide de un grupo siempre será uno de los puntos del grupo. Ésta es la principal diferencia entre el algoritmo k-medias y k-medoides.




### k-Medias (Kernel)

Kernel k-means utiliza núcleos para estimar distancias entre ejemplos y grupos. Debido a la naturaleza de los núcleos, es necesario sumar todos los ejemplos de un grupo para calcular una


distancia. Por lo tanto, este algoritmo es cuadrático en número de ejemplos y no devuelve un modelo de clúster centroide (por el contrario, el operador K-Means devuelve un modelo de clúster centroide).

## Aporte


-  entrada de conjunto de ejemplo (*tabla de datos*)

Este puerto de entrada espera un conjunto de ejemplos.

## Producción

-  modelo de clúster (*modelo de clúster centroide*)

Este puerto ofrece el modelo de clúster. Contiene la información sobre qué ejemplos forman parte de qué grupo. También almacena la posición de los centroides de los clusters. Puede ser utilizado por el operador de aplicación del modelo para realizar la agrupación especificada en otro conjunto de ejemplos. El operador de modelos de grupo también puede agrupar el modelo de clúster con otros modelos de clúster, modelos de preprocesamiento y modelos de aprendizaje.

-  conjunto agrupado (*tabla de datos*)

Se agrega un atributo 'id' con la función especial 'Id' al conjunto de ejemplos de entrada para distinguir los ejemplos. Dependiendo del atributo de agregar clúster y de los parámetros de agregar como etiqueta, también se agrega un atributo 'clúster' con la función especial 'Clúster' o 'Etiqueta'. El conjunto de ejemplos resultante se entrega en este puerto de salida.

# K-medias (núcleo)(Núcleo de estudio RapidMiner)

---

## Sinopsis

Este operador realiza agrupación en clústeres utilizando el algoritmo k-means del kernel. La agrupación se ocupa de agrupar objetos que son similares entre sí y diferentes a los objetos que pertenecen a otros grupos. Kernel k-means utiliza kernels para estimar la distancia entre objetos y grupos. K-means es un algoritmo de agrupamiento exclusivo.

## Descripción

Este operador realiza agrupación en clústeres utilizando el algoritmo k-means del kernel. K-means es un algoritmo de agrupamiento exclusivo, es decir, cada objeto se asigna precisamente

a uno de un conjunto de grupos. Los objetos de un grupo son similares entre sí. La similitud entre objetos se basa en una medida de la distancia entre ellos. Kernel k-means utiliza kernels para estimar la distancia entre objetos y grupos. Debido a la naturaleza de los núcleos, es necesario sumar todos los elementos de un grupo para calcular una distancia. Por lo tanto, este algoritmo es cuadrático en número de ejemplos y no devuelve un modelo de clúster centroide al contrario del operador K-Means. Este operador crea un atributo de clúster en el conjunto de ejemplos resultante si el parámetro *Agregar atributo de clúster* está establecido en verdadero. La agrupación se ocupa de agrupar objetos que son similares entre sí y diferentes a los objetos que pertenecen a otros grupos. La agrupación es una técnica para extraer información de datos sin etiquetar. La agrupación puede ser muy útil en muchos escenarios diferentes, por ejemplo, en una aplicación de marketing, podemos estar interesados en encontrar grupos de clientes con un comportamiento de compra similar.

## Agrupación aglomerativa(Núcleo de estudio RapidMiner)

---

### Sinopsis

Este operador realiza agrupación aglomerativa, que es una estrategia ascendente de agrupación jerárquica. Este operador admite tres estrategias diferentes: enlace único, enlace completo y enlace medio. El resultado de este operador es un modelo de conglomerado jerárquico, que proporciona información de distancia para trazar como un dendrograma.

### Descripción

La agrupación aglomerativa es una estrategia de agrupación jerárquica. La agrupación jerárquica (también conocida como agrupación basada en conectividad) es un método de análisis de conglomerados que busca construir una jerarquía de conglomerados. La agrupación jerárquica se basa en la idea central de que los objetos están más relacionados con objetos cercanos que con objetos más lejanos. Como tal, estos algoritmos conectan 'objetos' (o ejemplos, en el caso de un conjunto de ejemplos) para formar grupos en función de su distancia. Un grupo se puede describir en gran medida por la distancia máxima necesaria para conectar partes del grupo. A diferentes distancias se formarán distintos clusters, que pueden representarse mediante un dendrograma, lo que explica de dónde proviene el nombre común de

'clustering jerárquico': estos algoritmos no proporcionan una única partición del conjunto de datos, sino que proporcionan una extensa jerarquía de grupos que se fusionan entre sí a determinadas distancias. En un dendrograma, el eje y marca la distancia a la que se fusionan los grupos, mientras que los objetos se colocan a lo largo del eje x para que los grupos no se mezclen.

Las estrategias de agrupamiento jerárquico generalmente se dividen en dos tipos:

- Aglomerativo: este es un enfoque ascendente: cada observación comienza en su propio grupo y los pares de grupos se fusionan a medida que uno asciende en la jerarquía.
- Divisivo: este es un enfoque de arriba hacia abajo: todas las observaciones comienzan en un grupo y las divisiones se realizan de forma recursiva a medida que uno desciende en la jerarquía.

La agrupación jerárquica es toda una familia de métodos que se diferencian en la forma en que se calculan las distancias. Además de la elección habitual de funciones de distancia, el usuario también debe decidir el criterio de vinculación que utilizará, dado que un grupo consta de múltiples objetos, existen múltiples candidatos para calcular la distancia. Las opciones populares se conocen como agrupación de enlace simple (el mínimo de distancias de objetos), agrupación de enlace completo (el máximo de distancias de objetos) o agrupación de enlace promedio (también conocida como UPGMA, 'Método de grupo de pares no ponderados con media aritmética').

El algoritmo forma grupos de abajo hacia arriba, de la siguiente manera: inicialmente, coloque cada ejemplo en su propio grupo. Entre todos los grupos actuales, elija los dos grupos con la distancia más pequeña. Reemplace estos dos grupos con un nuevo grupo, formado mediante la fusión de los dos originales. Repita los dos pasos anteriores hasta que solo quede un clúster en el grupo.

La agrupación se ocupa de agrupar objetos que son similares entre sí y diferentes a los objetos que pertenecen a otros grupos. Es una técnica para extraer información de datos sin etiquetar y puede resultar muy útil en muchos escenarios diferentes, por ejemplo, en una aplicación de marketing, podemos estar interesados en encontrar grupos de clientes con un comportamiento de compra similar.