KIBERNUM
IT Academy

Módulo 6 – Aprendizaje de Máquina No Supervisado

Qué es Clustering

Especialización en Ciencia de Datos

Contenido

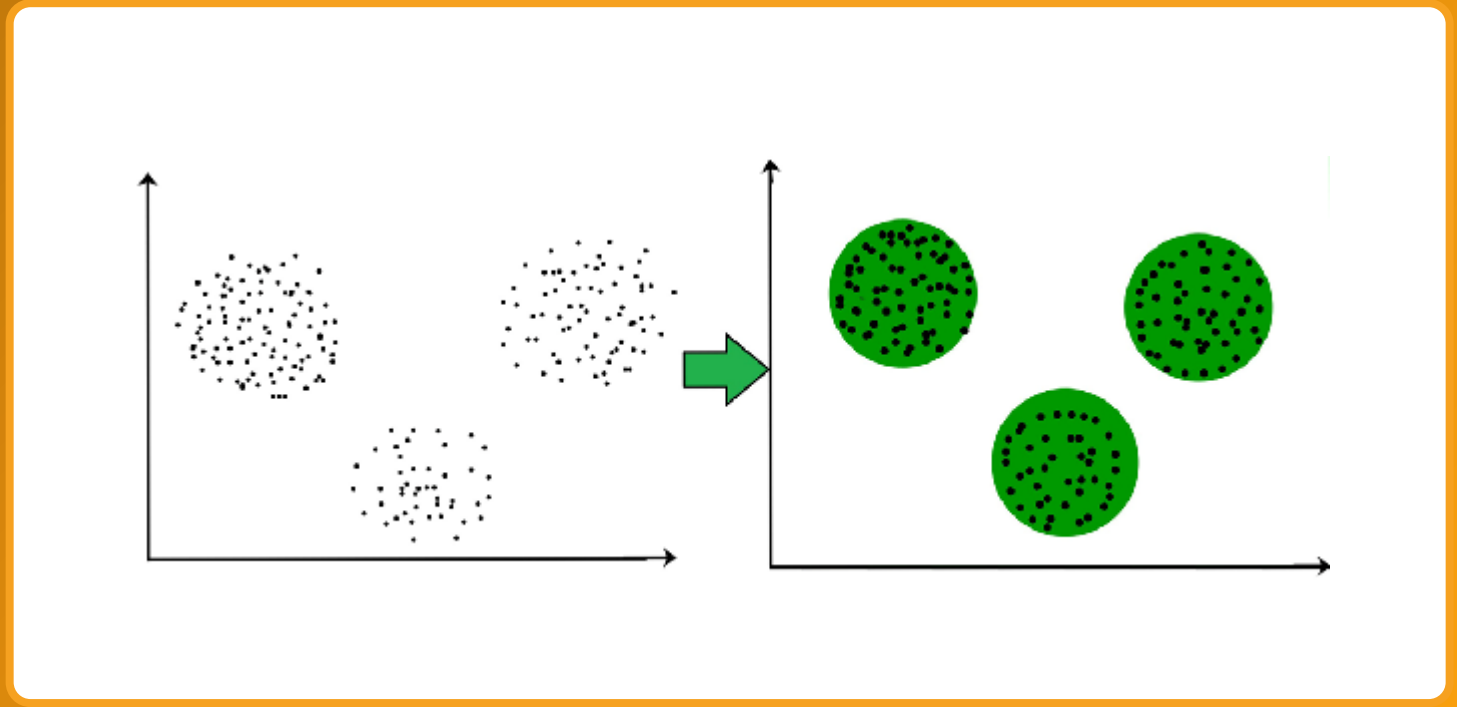
- Describir los conceptos básicos de clustering.
- Identificar los algoritmos de clustering más utilizados.

KIBERNUM

Análisis de Clustering

¿Qué es Clustering?

Clustering, también conocido como agrupamiento, es una técnica de aprendizaje no supervisado en la que se utilizan algoritmos para identificar grupos o clústeres de objetos o datos similares. El objetivo principal del clustering es agrupar objetos similares juntos y separar objetos diferentes en grupos distintos, sin tener una clasificación previa de los datos.



El algoritmo de clustering, funciona al analizar las características de los datos y buscando patrones y similitudes en los valores. Estos patrones se utilizan para agrupar los objetos o datos en clusters o grupos.

Corresponde a técnicas de Machine Learning no-supervisado en donde a partir de datos no etiquetados, se le asigna una etiqueta.

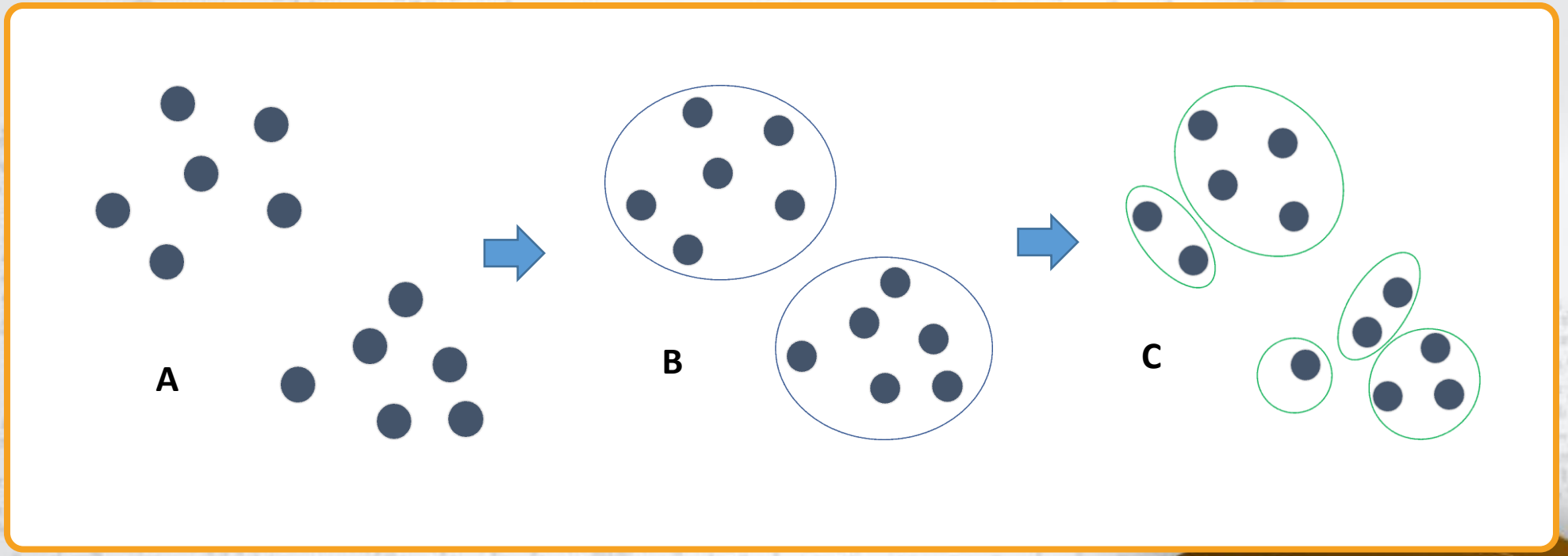
¿Qué es Clustering?

El clustering, al ser un aprendizaje no supervisado, **no tiene respuestas correctas**. Esto hace que la evaluación de los grupos (cantidad y significado) sea un tanto subjetiva.

KIBERNUM

¿Qué es Clustering?

En el siguiente ejemplo, alguna persona podría decir que hay dos clusters, sin embargo, alguien también podría argumentar la existencia de 5 clusters. No hay una respuesta correcta o incorrecta, **va a depender de la interpretación que se haga**.

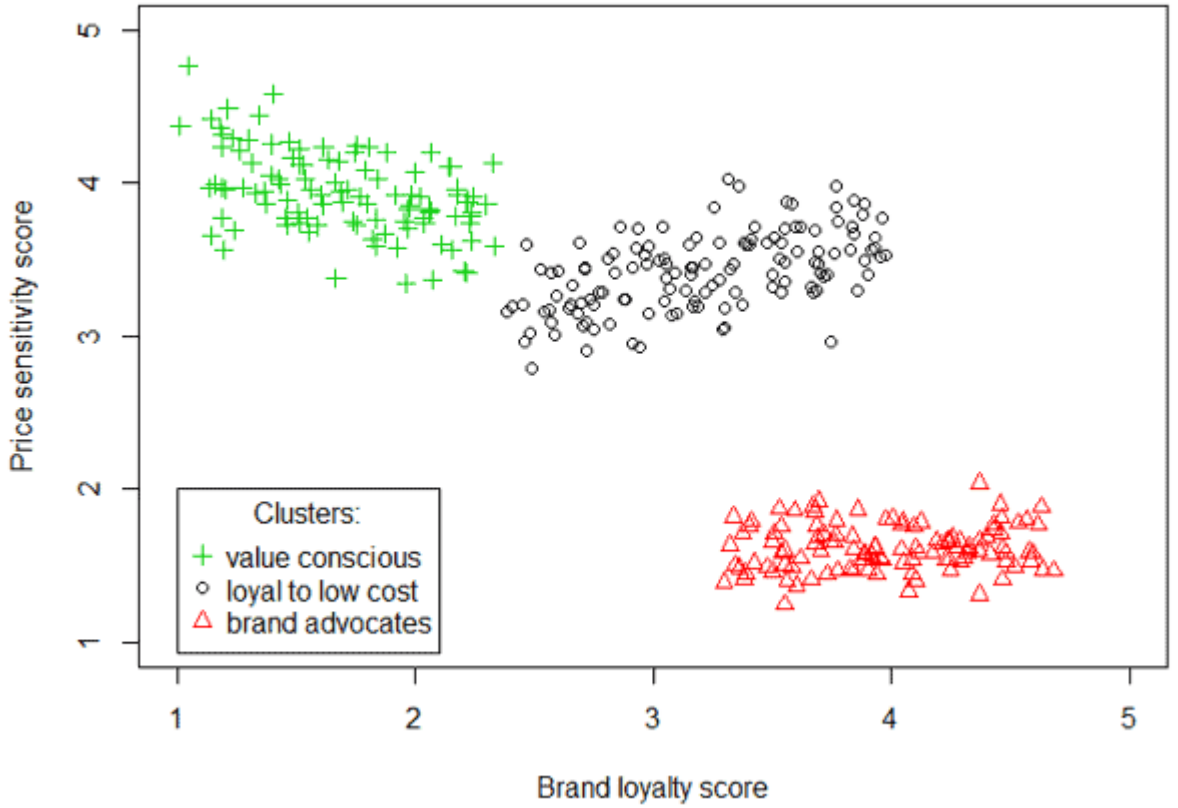


KIBERNUM

Ejemplos de Clustering

Segmentación de Clientes

- Preocupados por el precio (verde): no son leales a la marca y son muy sensitivos al precio.
- Leales a precios bajos (negro): son leales a la marca, pero sólo si es barato.
- Defensores de la marca (rojo): son leales a la marca sin importar demasiado el precio.



KIBERNUM

Algoritmos de Clustering

Algunos de los algoritmos de clustering más populares son k-means, agrupamiento jerárquico, agrupamiento espectral, agrupamiento basado en densidad, entre otros. Cada uno de estos algoritmos tiene sus propias fortalezas y debilidades, por lo que la elección del algoritmo adecuado dependerá del tipo de datos y de los objetivos del análisis.

Nombre del método	Parámetros	Escalabilidad	Caso de uso	Geometría (métrica utilizada)
K-Means	número de grupos	Muy grande n_samples, mediano n_clusterscon código MiniBatch	De uso general, tamaño de grupo uniforme, geometría plana, no demasiados grupos, inductivo	Distancias entre puntos
Propagación por afinidad	amortiguación, preferencia de muestra	No escalable con n_samples	Muchos grupos, tamaño de grupo desigual, geometría no plana, inductivo	Graficar la distancia (por ejemplo, gráfico del vecino más cercano)
Mean-shift	banda ancha	No escalable conn_samples	Muchos grupos, tamaño de grupo desigual, geometría no plana, inductivo	Distancias entre puntos
Clustering espectral	número de grupos	Mediano n_samples, pequeñon_clusters	Pocos grupos, incluso tamaño de grupo, geometría no plana, transductivo	Graficar la distancia (por ejemplo, gráfico del vecino más cercano)
Clustering Jerárquico Ward	número de grupos o umbral de distancia	Grande n_samplesyn_clusters	Muchos grupos, posiblemente limitaciones de conectividad, transductivos.	Distancias entre puntos
Clustering aglomerativo	número de grupos o umbral de distancia, tipo de vínculo, distancia	Grande n_samplesyn_clusters	Muchos grupos, posiblemente limitaciones de conectividad, distancias no euclidianas, transductivas.	Cualquier distancia por pares
DBSCAN	tamaño del vecindario	Muy grande n_samples, medianon_clusters	Geometría no plana, tamaños de conglomerados desiguales, eliminación de valores atípicos, transductivo	Distancias entre puntos más cercanos

KIBERNUM

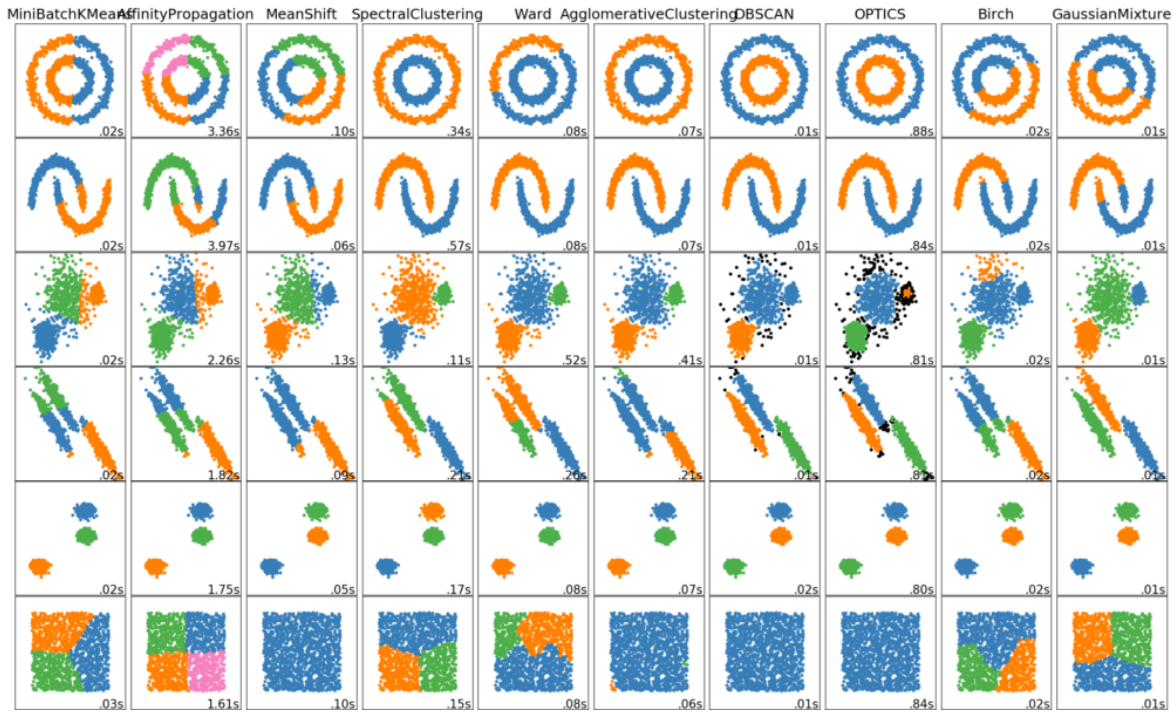
Algoritmos de Clustering

Más algoritmos. Información obtenida de la documentación oficial de la librería scikit-learn.

Nombre del método	Parámetros	Escalabilidad	Caso de uso	Geometría (métrica utilizada)
HDBSCAN	membresía mínima del clúster, vecinos mínimos de puntos	largo n_samplesmedianon_ clusters	Geometría no plana, tamaños de conglomerados desiguales, eliminación de valores atípicos, transductivo, jerárquico, densidad de conglomerados variable	Distancias entre puntos más cercanos
OPTICS	membresía mínima del clúster	muy grande n_samples, granden_clusters	Geometría no plana, tamaños de conglomerados desiguales, densidad de conglomerados variable, eliminación de valores atípicos, transductivo	Distancias entre puntos
Mezclas gaussianas	muchos	No escalable	Geometría plana, buena para estimación de densidad, inductiva.	Distancias de Mahalanobis a los centros
ABEDUL	factor de ramificación, umbral, agrupador global opcional.	Grande n_clustersyn_sampl es	Gran conjunto de datos, eliminación de valores atípicos, reducción de datos, inductivo	Distancia euclidiana entre puntos
Bisectriz de K-medias	número de grupos	Muy grande n_samples, medianon_clusters	De uso general, tamaño de grupo uniforme, geometría plana, sin grupos vacíos, inductivo, jerárquico	Distancias entre puntos

Fuente: https://scikit-learn.org/stable/modules/clustering.html

Algoritmos de Clustering



La comparación de los algoritmos de clustering está hecha en función a los parámetros que necesitan, su escalabilidad, caso de uso y geometría (métrica usada). Algunas preguntas útiles pueden ser:

- ¿Tengo una idea del número de grupos que quiero encontrar?, ¿o prefiero que el algoritmo lo encuentre?
- ¿Tengo muchísimos datos? En este caso, deberemos tener en cuenta la escalabilidad del algoritmo.

KIBERNUM

Aplicaciones del Clustering

Las técnicas de clusterización tienen una amplia variedad de aplicaciones en diferentes campos, entre las que se incluyen:

1. **Marketing y publicidad**: La clusterización se utiliza para segmentar a los consumidores en diferentes grupos basados en sus preferencias, hábitos de compra, comportamientos y otras variables. Esto ayuda a las empresas a personalizar su publicidad y marketing para llegar a los consumidores de manera más efectiva.

2. **Biología**: En biología, la clusterización se utiliza para agrupar células en diferentes tipos o grupos basados en su expresión genética. Esto ayuda a los científicos a entender mejor la biología celular y a identificar posibles tratamientos para enfermedades.

Aplicaciones del Clustering

3. **Análisis de redes sociales**: La clusterización se utiliza para agrupar a los usuarios de las redes sociales en diferentes comunidades o grupos basados en sus intereses y relaciones sociales. Esto ayuda a las empresas a entender mejor a su audiencia y a diseñar estrategias de marketing más efectivas.

4. **Finanzas**: La clusterización se utiliza para agrupar diferentes activos financieros en diferentes clases de activos, como acciones, bonos y materias primas. Esto ayuda a los inversores a diversificar sus carteras y a minimizar el riesgo.

5. **Imagen y visión por computadora**: La clusterización se utiliza para segmentar y clasificar imágenes en diferentes grupos basados en su contenido, como el color, la textura y la forma. Esto ayuda a las máquinas a entender mejor las imágenes y a realizar tareas como la detección de objetos y la clasificación de imágenes.

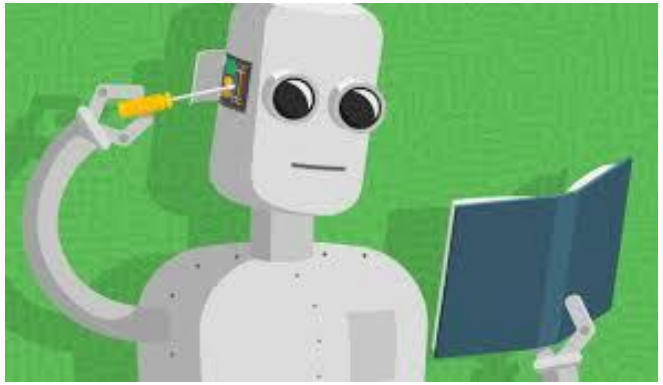
Tarea de Clustering

Ejemplo: "Segmentar clientes en subconjuntos símiles"

- **Experiencia:**
Datos de variables sociodemográficas, comportamiento de compra, comportamiento de pago.

- **Tarea:**
Generar grupos homogéneos de clientes.

- **Performance:**
Similitud de los clientes en cada grupo.



"Se dice que un computador aprende de la *experiencia E*, con respecto a una *tarea T* y una medida de *performance P*, si su performance en T, medido por P, mejora con la experiencia E."

(*Tom Mitchell*, 1998)

KIBERNUM

Dudas y consultas

KIBERNUM

Fin Presentación

KIBERNUM