Laboratorio de Datos

Clustering: k-medias

Primer Cuatrimestre 2024

Turnos tarde y noche

Facultad de Ciencias Exactas y Naturales, UBA

Aprendizaje Supervisado y No Supervisado

Aprendizaje Supervisado:

- El objetivo es predecir una variable respuesta a partir de variables explicativas.
- Conocemos la respuesta correcta para un conjunto de datos y usamos esos datos para construir el modelo.
- Ejemplos: regresión, clasificación.

Aprendizaje No Supervisado:

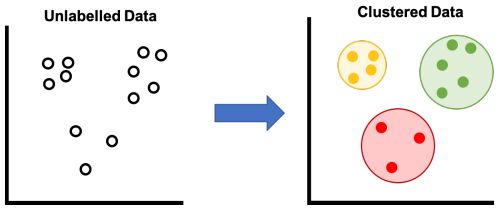
- El objetivo es encontrar patrones o estructuras ocultas en los datos.
- No conocemos o no hay a priori una respuesta correcta.
- Ejemplos: agrupación (clustering), reducción de dimensionalidad.

Clustering

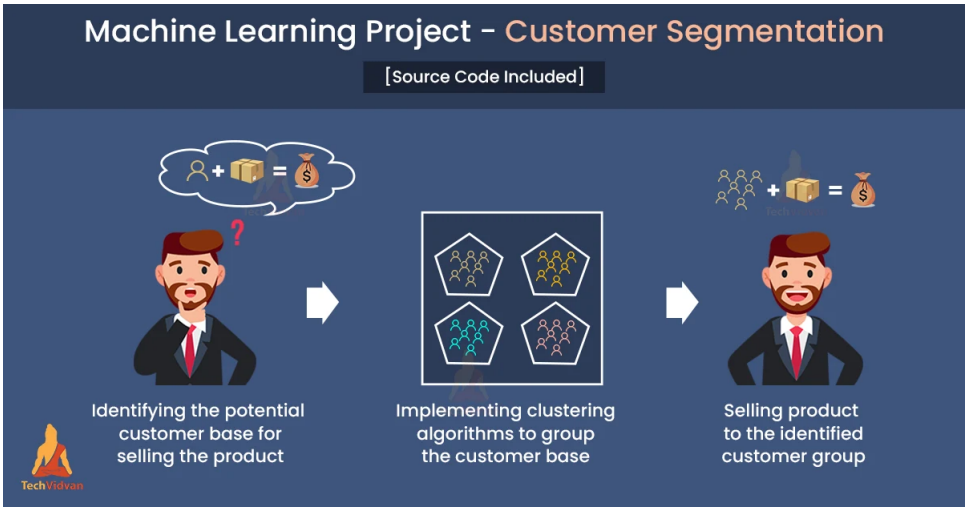
En las últimas clases estudiamos modelos de regresión (aprendizaje supervisado).

Estudiamos ahora métodos de agrupamiento(clustering).

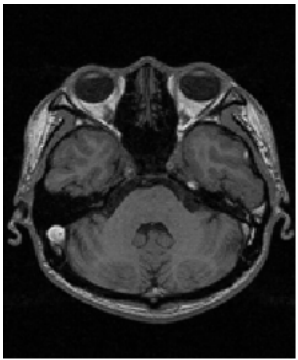
- Clustering es una técnica de aprendizaje no supervisado.
- El objetivo es agrupar datos similares en conjuntos llamados clústeres.
- Aplicaciones: segmentación de mercado, análisis de redes sociales, imagenes médica, etc.



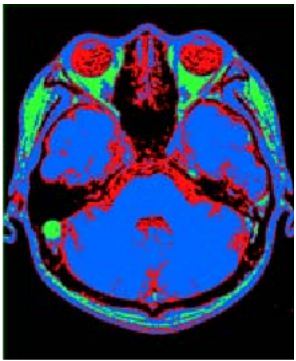
Aplicaciones: segmentación de mercado



Aplicaciones: imágenes médicas



(a)

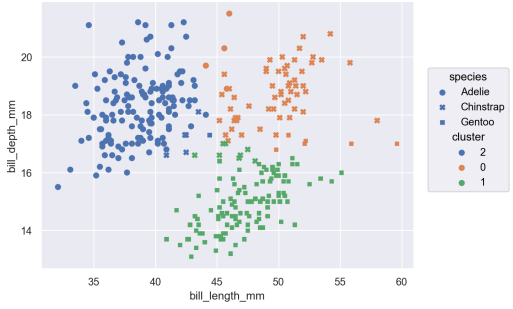


(b)

Clustering y datos cercanos

Si tenemos un conjunto de datos (por ejemplo, un DataFrame), consideramos a cada observación (fila) como un punto en un espacio de dimensión n, donde n es la cantidad de variables (columnas).

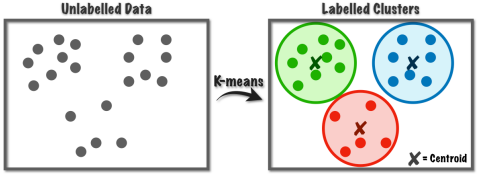
La cantidad de variables puede ser arbitrariamente grande, pero en general vamos a trabajar con 2 variables para poder visualizar los datos en el plano.



Clustering por K-medias (K-means)

Existen diversos algoritmos de clustering. El algoritmo k-medias comenzó a utilizarse alrededor de 1960.

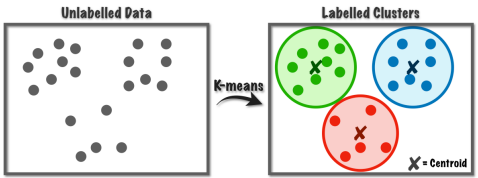
- K-medias es un algoritmo de agrupación que particiona los datos en k clústeres.
- Se definen k centros para los clusters.
- Se supone que cada cluster es un conjunto de datos que están razonablemente bien aproximados por el centro del cluster (el promedio de los valores del cluster).



Idea del funcionamiento de k-medias

- Si fijamos unos centros para los clusters, podemos determinar los clusters tomando los puntos más cercanos a cada centro.
- Si fijamos los clusters, podemos determinar los centros tomando el promedio de todos los puntos del cluster.

Repitiendo estos pasos alternativamente obtenemos un algoritmo conocido como algoritmo k-medias naive.



Algoritmo k-medias (naive)

Comenzamos con un conjunto arbitrario de centros (por ejemplo k datos seleccionados al azar).

Ahora repetimos los siguientes dos pasos hasta "alcanzar convergencia":

1. Actualizamos las etiquetas: para cada punto (dato), buscamos cuál es el centro más cercano a ese punto y lo asignamos a ese cluster.
2. Actualizamos los centros: para cada cluster, recalculamos el centro del cluster como el promedio de todo los datos del cluster (si los puntos tienen varias coordenadas, calculamos el promedio de cada coordenada).

¿Qué significa en este caso alcanzar convergencia?

Algoritmo k-medias (naive)

Comenzamos con un conjunto arbitrario de centros (por ejemplo k datos seleccionados al azar).

Ahora repetimos los siguientes dos pasos hasta "alcanzar convergencia":

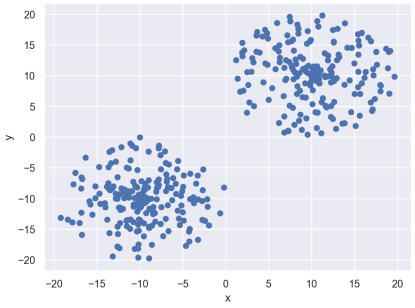
1. Actualizamos las etiquetas: para cada punto (dato), buscamos cuál es el centro más cercano a ese punto y lo asignamos a ese cluster.
2. Actualizamos los centros: para cada cluster, recalculamos el centro del cluster como el promedio de todo los datos del cluster (si los puntos tienen varias coordenadas, calculamos el promedio de cada coordenada).

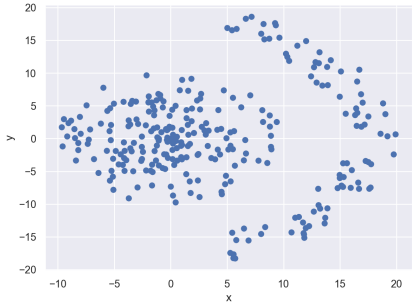
¿Qué significa en este caso alcanzar convergencia?

Que las etiquetas no cambien de un paso al siguiente.

Ejemplos

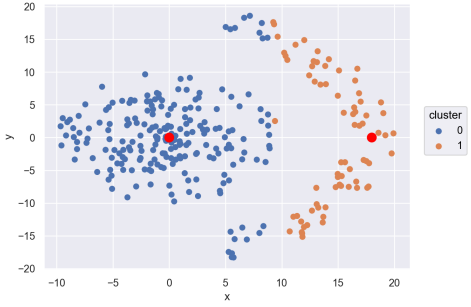
¿En cuál de estos casos, los clusters quedarán bien definidos a partir de sus centros?





Ejemplos

¿En cuál de estos casos, los clusters quedarán bien definidos a partir de sus centros?



Aplicaciones de k-means

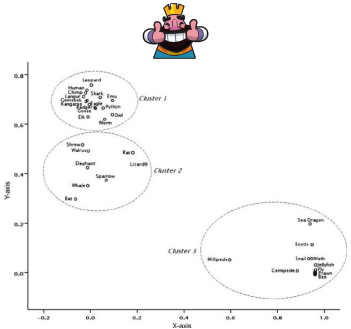
¿En cuáles de estas aplicaciones el método de k-means puede resultar apropiado?

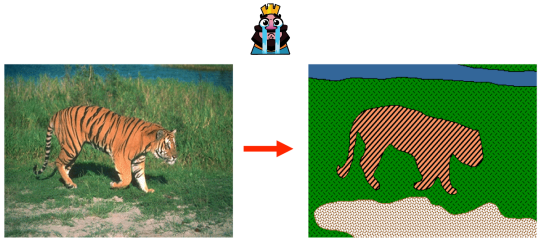
1. Clasificar especies animales según características (peso, largo, longitud de la cola, ...).
2. Identificar distintas componentes de una imagen.

Aplicaciones de k-means

¿En cuáles de estas aplicaciones el método de k-means puede resultar apropiado?

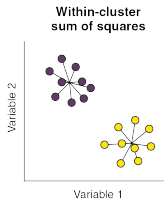
1. Clasificar especies animales según características (peso, largo, longitud de la cola, ...).
2. Identificar distintas componentes de una imagen.





K-medias como problema de optimización

Podemos plantear el problema de clustering por k-medias como un problema de optimización (y el algoritmo k-medias naive es un método simple de optimización).



- Buscamos minimizar la "suma de los errores".
- Consideramos que los errores son las distancias (al cuadrado) de cada punto al centro de su cluster: $\|\mathbf{x} - \boldsymbol{\mu}_i\|^2$.
- Es decir, queremos minimizar la suma de los cuadrados dentro de cada cluster (WCSS o within cluster sum of squares).
- Formalmente queremos elegir clusters $S_1, \ldots, S_k$ y centros $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ que minimicen:

$$WCSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Ventajas y Desventajas del algoritmo k-medias naive

Ventajas:

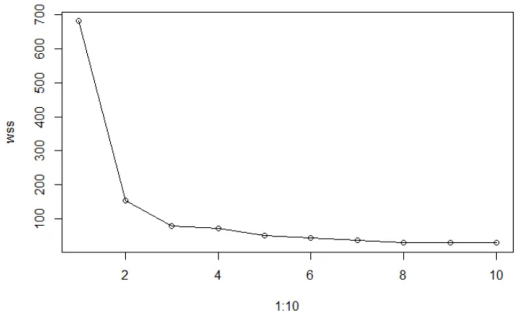
- Simple y fácil de implementar.
- Escalable para grandes conjuntos de datos.

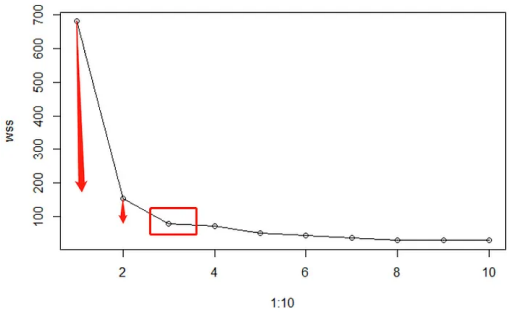
Desventajas:

- Requiere especificar el número de clústeres (k) de antemano.
- Sensible a la elección de los centroides iniciales.
- Puede converger a mínimos locales.

¿Número óptimo de clusters?

- A medida que aumentamos la cantidad de clusters, el error WCSS disminuye.
- Por lo tanto no hay un número óptimo de clusters.
- Podemos mirar la curva y ver cuándo aumentar la cantidad de clusters no reduce significativamente el error WCSS.
- Este proceso suele llamarse "método del codo".





Requerimientos para poder aplicar k-medias

k-medias supone las siguientes propiedades de los datos, y puede fallar si estas propiedades no se cumplen:

1. Los clusters son esféricos e isotrópicos (el mismo radio en todas las direcciones).
2. Los clusters tienen la misma varianza.
3. Todos los clusters tienen aproximadamente la misma cantidad de puntos.