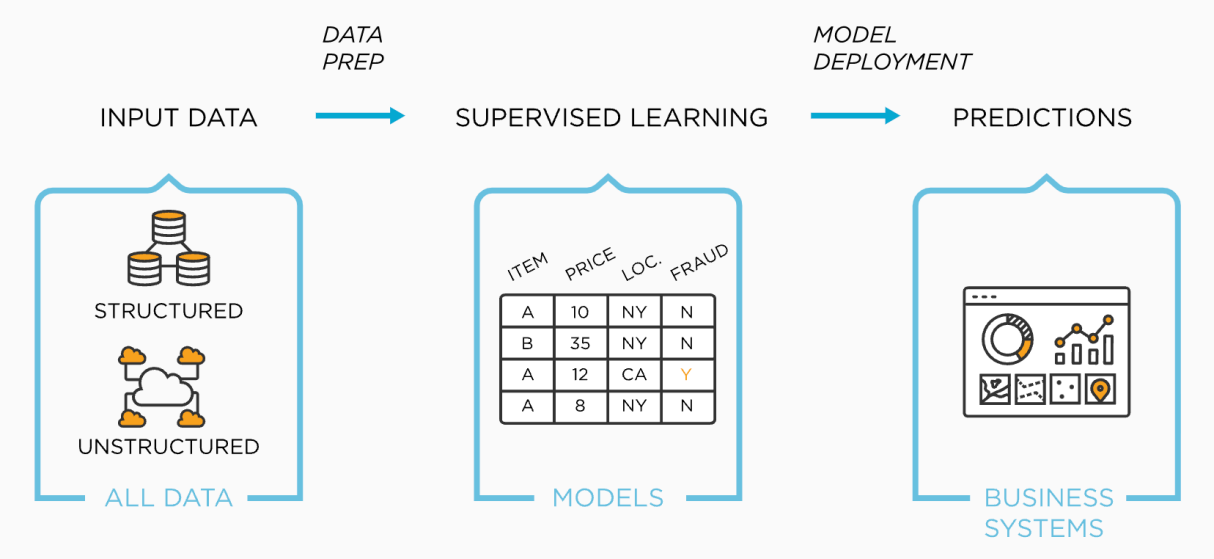
Clustering in R

Ing. Jairo Salazar

Aprendizaje supervisado vrs aprendizaje no supervisado



El aprendizaje supervisado es una rama de Machine Learning , un método de análisis de datos que utiliza algoritmos que aprenden iterativamente de los datos para permitir que las maquinas encuentren **información escondida** sin tener que programar de manera explícita dónde buscar. El aprendizaje supervisado es uno de los tres métodos de la forma en que las máquinas "aprenden": supervisado, no supervisado y optimización.

Aprendizaje no supervisado

- Por el contrario, el aprendizaje no supervisado es un tipo de Machine Learning que se utiliza para identificar nuevos patrones y detectar anomalías. Los datos que se introducen en los algoritmos de aprendizaje no supervisados no están etiquetados. El algoritmo (o modelos) intentan dar sentido a los datos por sí mismos mediante la búsqueda de características y patrones.

Componentes basicos de los algoritmos de machine learning

- Dataset (informacion estructurada o no estructurada según sea el caso).
- Eleccion del modelo.
- Modelo programado.
- Funcion de error
- Modelo entrenado.

The base of K-means: optimizing the sum squared error

La idea detrás del algoritmo k-means es clasificar cada observacion de un dataset en un numero k de grupos que llamaremos clusters. El algoritmo se denomina k-means. Como el algoritmo decide a que cluster pertenece cada observacion?

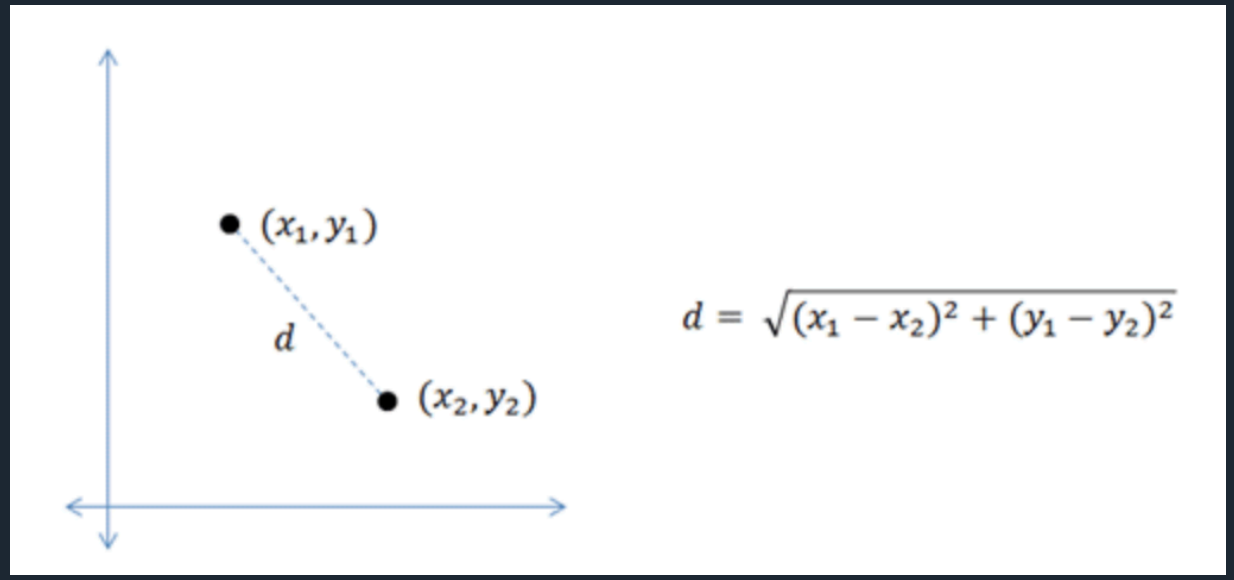
El algoritmo sigue estos pasos:

1. Inicializa los k centroides de manera aleatoria.
2. Para cada observacion, se calcula el error cuadrado (sum squared error) hacia cada centroide.
3. Para cada observacion, se asigna el centroide que minimice el error que calculamos hacia su cluster.

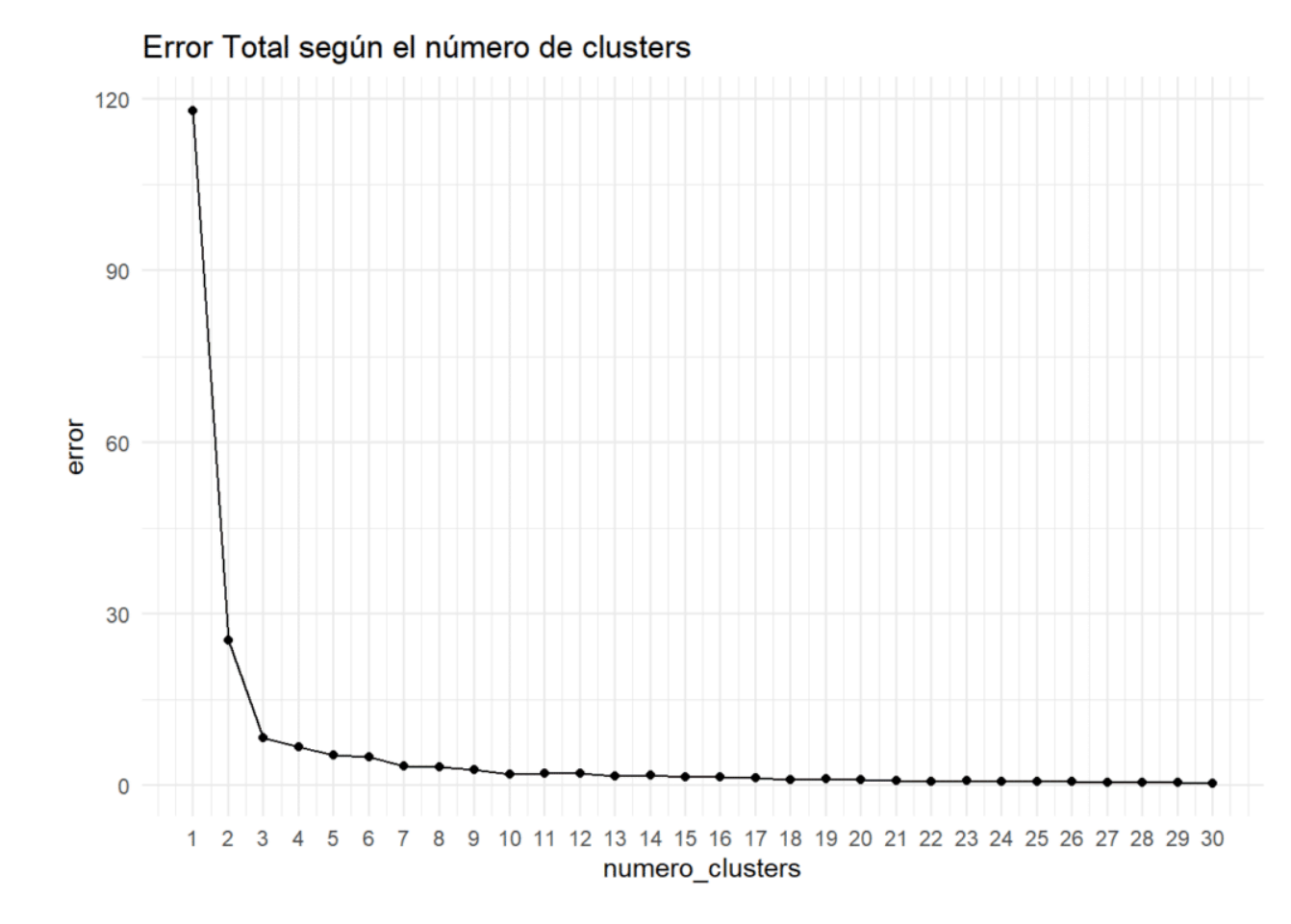
Sum of squared errors, Euclidean distance and the Pythagoras theorem

La suma de los errores al cuadrado es igual a la distancia euclediana entre cada punto y el centroide. Esta regla se aplica sin importar el numero de variables (y dimensiones) que tengamos. Asi que indirectamente, el algoritmo k-means encuentra el centroide que es linealmente mas cercano a cada observacion.

La distancia euclediana esta basada en el teorema de pitagoras, que tiene la siguiente formula:



K-means: how to choose the number of clusters (elbow method)



En nuestro ejemplo ya conociamos exactamente el numero de clusters ideal para el algoritmo, sin embargo en la vida real no lo conocemos. Es por esto, que para determinar el numero de los clusters es clave en el aprendizaje no supervisado.

Una de las maneras mas comunes de determinar este numero es el elbow method. Este metodo esta bastado en correr varias veces el algoritmo de kmeans con diferentes numeros de cluster. Se grafica el error para los diferentes clusters: