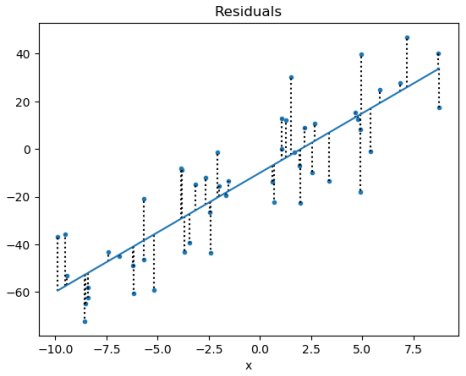
Clase 10: Clustering

Metodología supervisada

La metodología supervisada se diferencia en que se conoce la variable respuesta y. En otras palabras, un modelo supervisado se puede entender estadística mente como aprender de nuestra variable respuesta, sabiendo que ocurren sucesos.

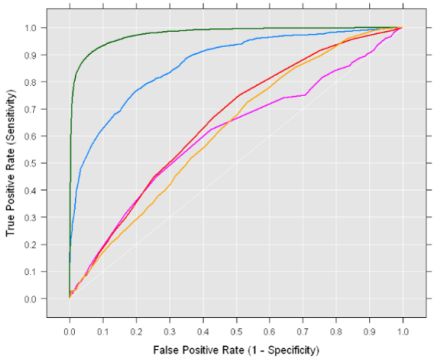
$$M = \mathrm{P}(y|X)$$



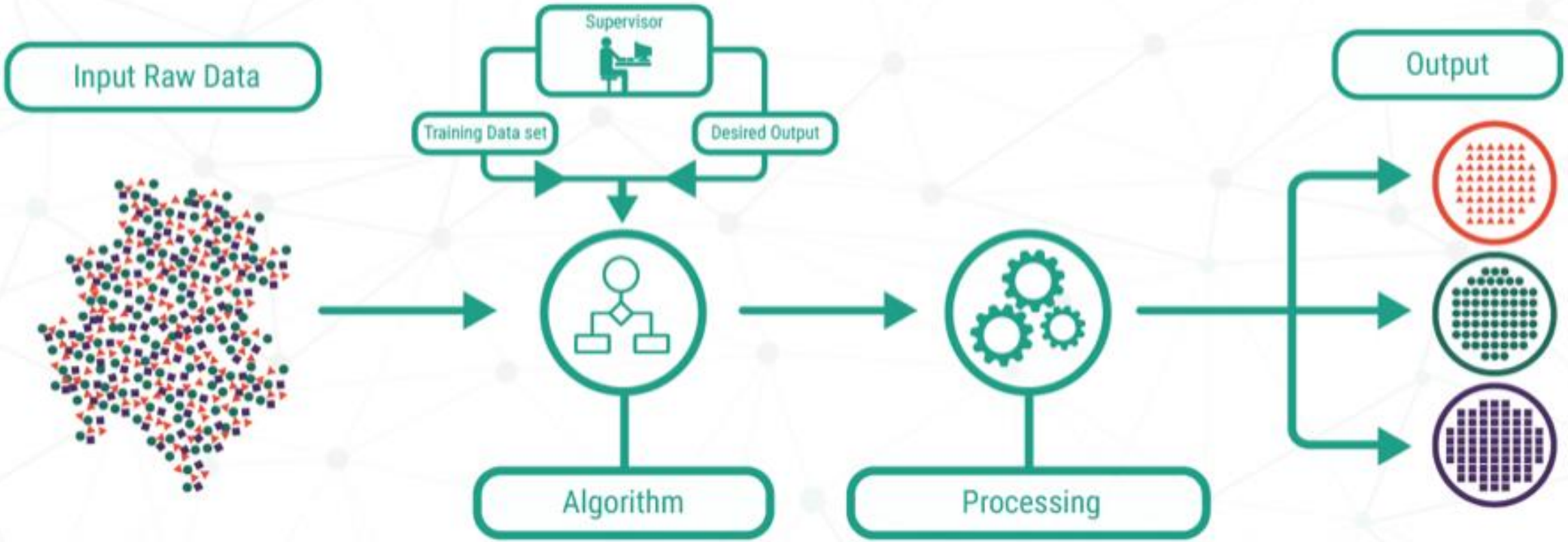
El ejemplo más básico es el modelo lineal, en el cual modelamos la media de la distribución de y dado X, es decir

$$E(y|X) = X\beta + \epsilon$$

Notar que esta formulación la variable de interés siempre es y



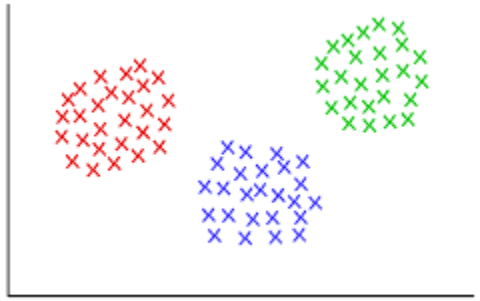
Nuestra evaluación siempre la podemos hacer con alguna validación (Hold-out o cross-validation) pues ya sabemos de antemano la respuesta correcta.


Input Raw Data
Supervisor
Training Data set
Desired Output
Algorithm
Processing
Output

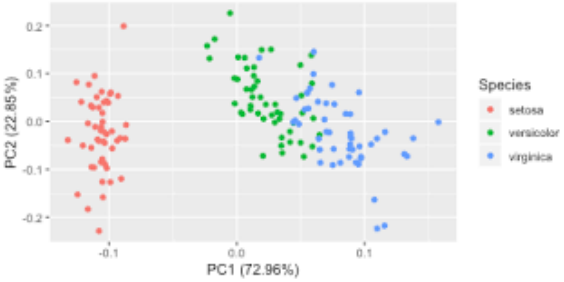
Metodología no supervisada

La metodología no supervisada no existe el rol de un supervisor en el modelo, es decir no se sabe a priori el verdadero valor de la etiqueta y. Estadísticamente se puede entender como aprender de la conjunta de los datos

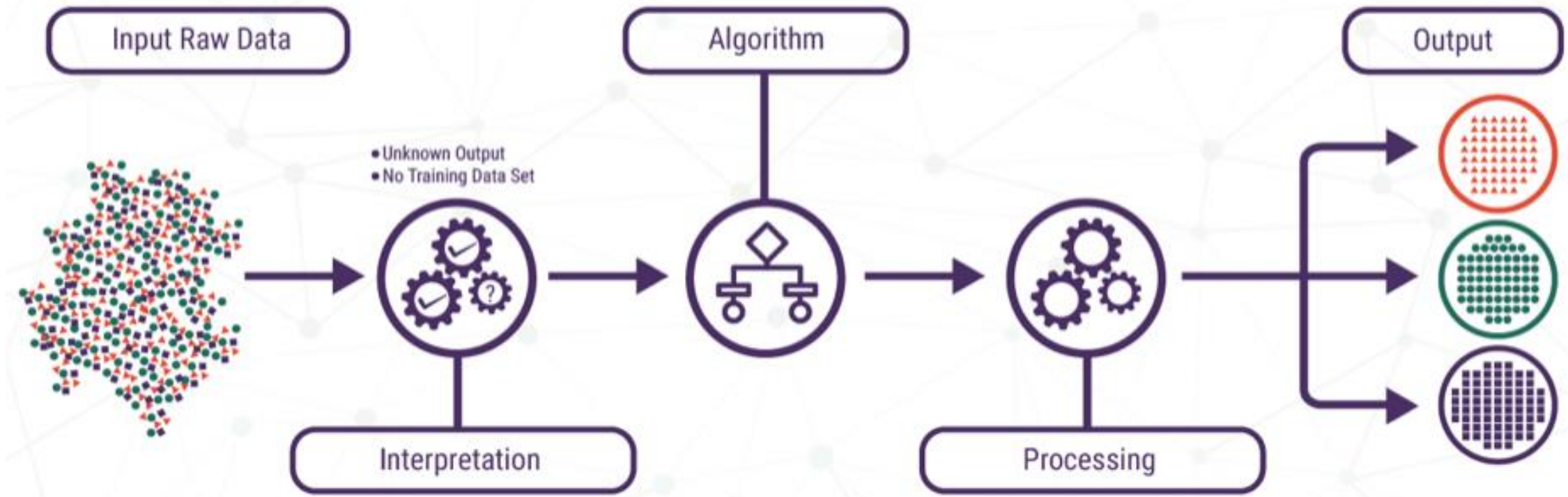
$$M = \mathrm{P}(y, X)$$



Un ejemplo en este caso podría ser el que veremos más adelante, el cual son los modelos de clusterización, sin embargo esta familia de modelos no son los únicos que caen en la categoría no supervisada; otros son: PCA, ICA, A priori, Filtros colaborativos.



La evaluación es distinta, al no saber la respuesta el compararse contra un conjunto de test no tiene mucho sentido, la evaluación va a depender de la metodología utilizada y de la interpretación de los mismos.


Input Raw Data
Unknown Output
No Training Data Set
Interpretation
Algorithm
Processing
Output

Classical Machine Learning

Task Driven

Data Driven

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

Unsupervised Learning

(Unlabelled Data)

Clustering

(Divide by Similarity)

Eg. Targeted Marketing

Association

(Identify Sequences)

Eg. Customer Recommendation

Dimensionality Reduction

(Wider Dependencies)

Eg. Big Data Visualization

Obj: Predications & Predictive Models

Pattern/ Structure Recognition

K-means

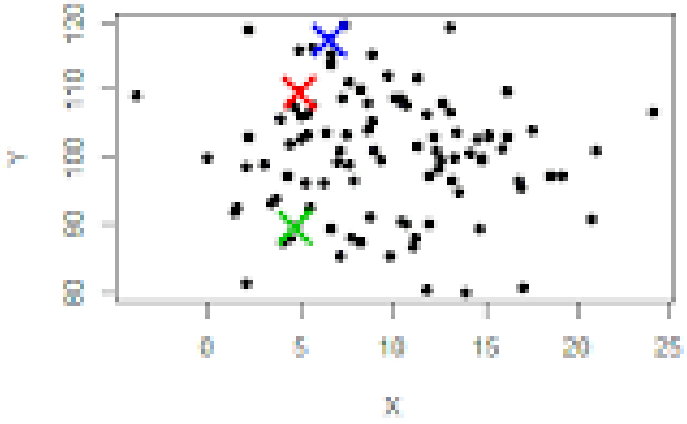
Probablemente el algoritmo de clustering más popular, consiste en dados unos centroides iniciales, vamos desplazando estos de tal forma que se le asignen los puntos más cercanos a ese centroide. Este centro de masa se irá movidendo de acuerdo a los puntos que se le irán asignando.

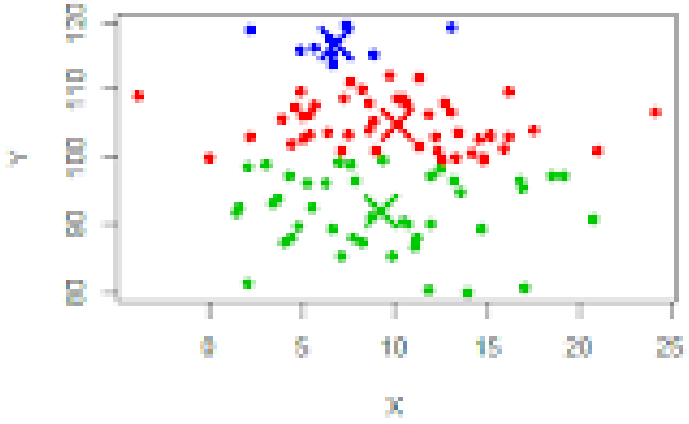
En este enlace se muestra una animación de su funcionamiento
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

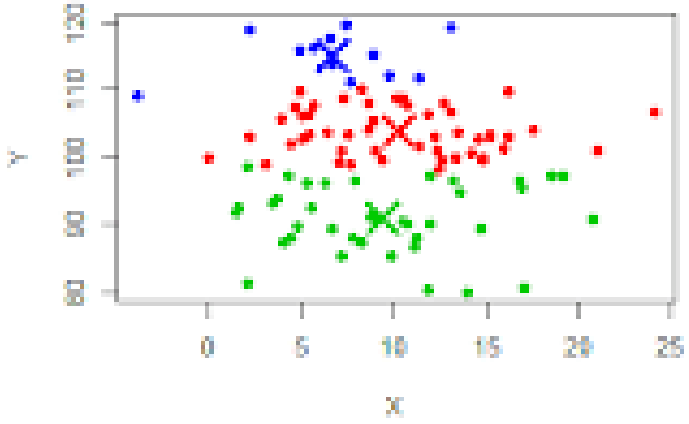
Algoritmo (algoritmo de Lloyd)
Sea k1, k2, …, km centroides, se realizan estos dos pasos de forma iterativa:

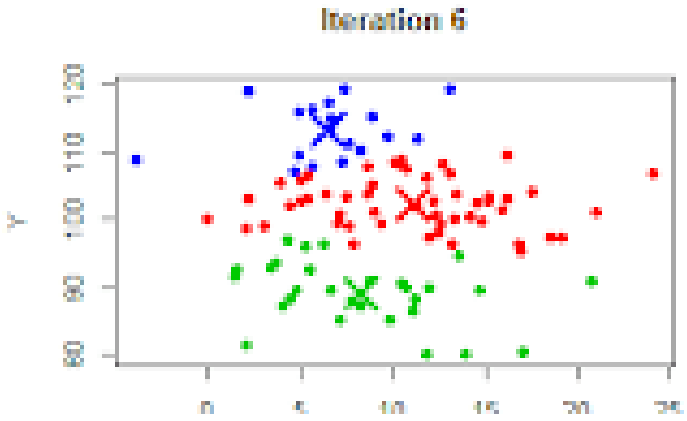
1. **Asignación**: Asignar cada punto a su centroide más cercano, de acuerdo a la métrica definida (normalmente distancia euclideana).
2. **Actualización**: Actualizar cada centroide se actualiza como la media del nuevo grupo.
3. **Criterio de detención**: Si no varían los centroides en relación al paso anterior, se detiene el algoritmo, caso contrario volver a 1.

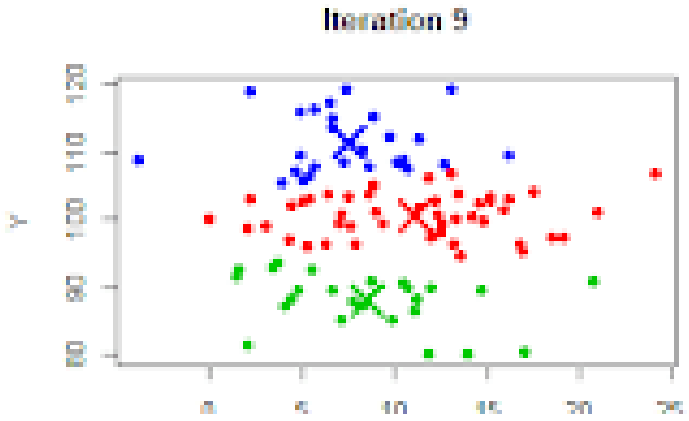
K-means

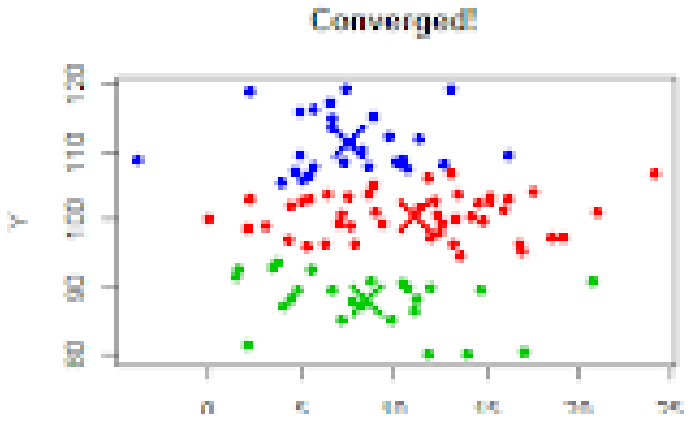

Y
X
0
5
10
15
20
25
80
90
100
110
120


Y
X
0
5
10
15
20
25
80
90
100
110
120


Y
X
0
5
10
15
20
25
80
90
100
110
120


Iteration 6
Y
80
90
100
110
120


Iteration 9
Y
80
90
100
110
120


Converged!
Y
80
90
100
110
120

Medición

En general para medir la efectividad de los clusters, podremos distinguir dos tipos de criterios, los "duros" y los "blandos". Los duros se refieren a métricas y números (que tampoco es algo tan indicativo como en las metodologías supervisadas). Por otra parte los "blandos" se hace referencia a la interpretación de los mismos y que tengan algún sentido de negocio.

La idea de clusterizar es ganar información dada la nueva partición, una forma de estudiar esto es tener un indicador de cuantos ganamos en varianza.
Suma de cuadrados dentro La suma cuadrática de los puntos dada la separación

$$\boldsymbol{SCE} = \sum_{k=1}^{K} \sum_{i=1}^{N} \boldsymbol{I(C_i = k)} \sum_{i=1}^{N_K} \left|\left| \boldsymbol{x_i} - \overline{\boldsymbol{x_k}} \right|\right|^2$$

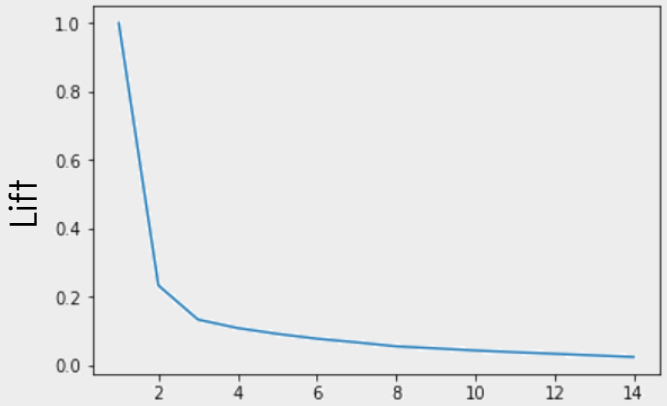
Suma de cuadrados totales La suma de todas las distancias al centro

$$\boldsymbol{SCT} = \sum_{i=1}^{n} \left|\left| \boldsymbol{x_i} - \overline{\boldsymbol{x}} \right|\right|^2$$

Con estos dos números se calcula el Lift, que es un indicador de que tanto se reduce la varianza.

$$Lift = \frac{SCE}{SCT}$$

Notarán que este número siempre va a ir bajando a medida que se incremente el número de clusters, en efecto si el número de clusters es igual a la cantidad de registros este número es cero. Se puede hacer un gráfico de sedimentación de cuanto se gana al clusterizar para escoger el número de clusters.

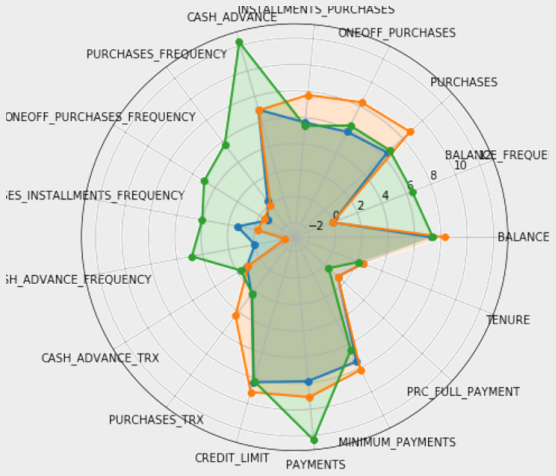


Medición

Al obtener los clusters, es importante la caracterización de los mismos, por ejemplo obtener métricas dentro de cada cluster y buscar definir que es lo que tienen en común todos los individuos dentro de un mismo cluster. Esto es a alto nivel sin embargo existen formas de ayudar a obtener una mejor interpretación

Caracterización por medias Obtener el promedio por cada partición y variable, permite tener algunos insight de los clusters. Debieran de ser distintas en más de una columna.

Visualización de medias Se sugiere complementar con una visualización de datos multivariados, puede ser útil el grafico de red, de araña o estrella que son los mismos.



Diferencia de medias. Todos conocemos el test t de diferencia de medias. puede ser utilizado para ir discriminando que variables para los mismos sujetos tienen diferencias fuertes, esto más que nada para validar con qué caracterizar un cluster.

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

En caso de muestras múltiples, no olviden aplicar la corrección de Bonferroni, es decir dividir su alfa por el número de pruebas de hipótesis simultaneas.

Variantes

K-medoides. Similar al anterior pero con una cuantas diferencias, la más notoria es que los centroides, un punto de los datos, el cual se llama medoide. Además optimiza en base a una función de costo la configuración actual.

El algorimo utilizado es normalmente el PAM (Partición Mlrededor de Medoides), el cual, dado unos puntos iniciales

1. **Asignación**: Asignar cada punto al mediode más cercano
2. **Actualización**: Para cada mediode m y cada punto p, intercambiar p y m
    1. Si el costo aumenta, por ejemplo la suma total de las distancias al medoide, aumenta, entonces deshacer el cambio

K-modes. Sigue un concepto similar al k-means pero diseñado para agrupar datos categóricos. En escencia es lo mismo, sin embargo tiene algunas sutilezas en la implementación.

Dado k modas,

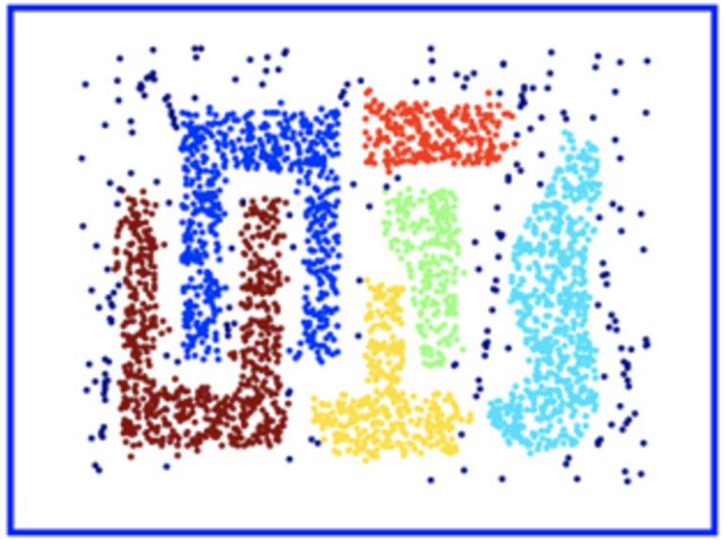
1. **Disimilitud**: Se calcula la disimilitud entre cada punto y la moda, esto es sumar uno si es distinto, 0 si no.
2. **Asignación**: Se asigna cada punto a la moda con menos disimilitud.
3. **Actualización**: Dentro de cada grupo se calula la moda y se vuelve al paso 1

Consideraciones

- Los métodos anteriores requieren que a priori se especifique un número de clusters, privilegiar en lo posible números pequeños.
- En caso de no poder reducir su número, evaluar la posibilidad de reducir dimensionalidad de los datos, por ejemplo con PCA (o ACP en español)
- K-medoids a diferencia del k-means muestra ser más robusto ante presencia de datos anómalos.
- K-modes es una alternativa para cuando el conjunto de datos no son continuos, si no más bien son categorías.
- El escalamiento de la data es super importante en estos métodos, pues se basan en la distancia de un punto a otro.
- Alternativamente es posible usar otras métricas de distancia, sin embargo el k-means puede ser sensible en la convergencia según la métrica que se elija, k-medoids es algo más robusto.

DBSCAN

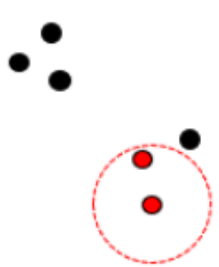
DBSCAN: Density Based Spatial Clustering with Applications with Noise, efectivamente una sigla para que diga DBSCAN, consiste en dado una vecindad con distancia especificada, ir iterando sobre los puntos en esa vecindad e ir incorporando secuencialmente más puntos.
Popular por su capacidad de reconocer "caminos" a diferencia de los anteriores que solo ven la distancia al punto.
No se requiere de especificar el número inicial de clusters y tiene excelente capacidad de reconocer puntos de alta densidad, sin embargo puede presentar problemas con clusters de varianza heterogénea.



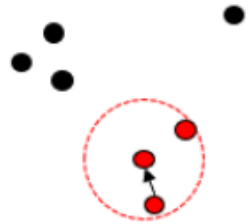
DBSCAN

Se tienen un conjunto de puntos (datos)

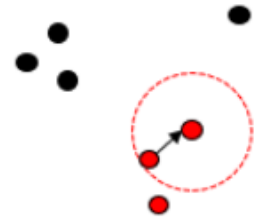
Se selecciona uno al azar



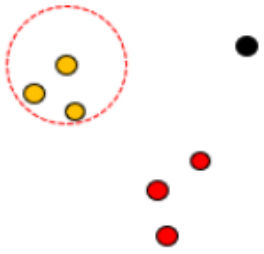
Se buscan cuantos puntos están en la región, esto es a distancia ε del escogido. Si hay al menos de N (en este ejemplo 2) puntos, se marcan como integrantes del clúster, caso contrario el punto se marca como ruido y se pasa a otro punto.



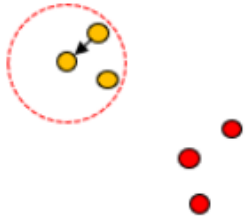
Se repite la operación con uno de los puntos incorporados a la región.



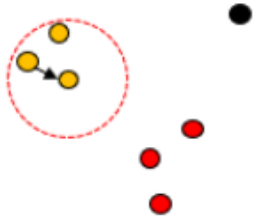
Cuando ya se recorran todos los puntos de la región, se cierra el clúster.



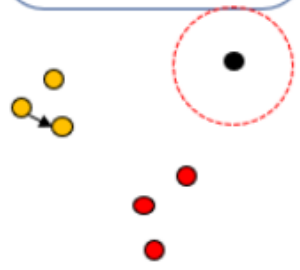
Se toma otro punto al azar no visitado.



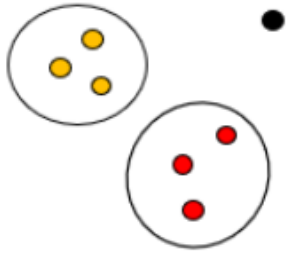
Se repite el proceso anterior hasta completar un clúster.



Cuando ya no se encuentren más puntos, se cierra el clúster y se pasa a otro punto no visitado.



En este caso se toma un último punto pero no tiene al menos dos integrantes en su región (sólo uno). Por lo cual se considera ruido

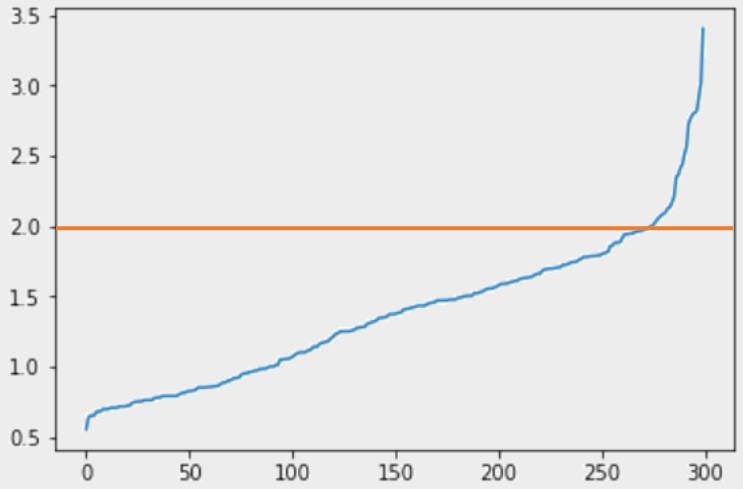


Felicitaciones!, ya tenemos dos clústeres y un punto que no formaría agrupación y sería ruido

DBSCAN

Hiperparámetros Es de los puntos más problemáticos de esta metodología, no es de elección trivial, sobre todo en data mutidimensional. Para obtener una idea de como ajustarlo, se puede usar lógica de k vecinos más cercanos de la siguiente forma

1. Fijar los mínimos puntos para armar vecindad, llamar a este punto k
2. Obtener los el k-ésimo vecino más cercano de cada punto y su distancia al punto.
3. Ordenar todas las distancias de menor a mayor
4. Graficar las distancias y observar el punto de quiebre, es decir cuando pasada esa distancia ocurre el ruido.



Notar en el ejemplo que a distancia 2 es el radio que mayoritariamente agrupa k vecinos por lo que sería un buen punto de partida para elegir el épsilon

DBSCAN

- Al igual que el K-means va a tener complicaciones con dimensiones muy elevadas, si es de muchas dimensiones la data, evaluar reducirla antes de aplicarlo.
- Puede fallar con dispersiones heterogéneas entre los clusters.
- En caso de no tener clusters muy definidos en la data, suele entregar solo uno.
- Los clusters obtenidos es más difícil de interpretar, evitar usar cuando se requiera análisis a profundidad de los clusters.
- Puede servir como un detector de anomalías en datos multivariados.
- Preferible su utilización en contextos como análisis de imágenes, de audio, etc.