Métodos de Clasificación y Agrupamiento (clustering)

Métodos de clasificación

- Regresión Logística

Métodos de clusterización

- Clustering
- K-Means

¿Cómo saber cuántos conjuntos elegir?

- Regla del codo (Elbow Method)
- Método Silhouette
- Estadística de Hopkins

Clasificación

Cuando resolvemos un problema de clasificación, buscamos, para ciertos datos de entrada, una categoría c de un conjunto C de categorías posibles. Estas categorías no solo son finitas, sino que además son conocidas de antemano.

Cuando tenemos varias categorías posibles y quiero saber a cuál de ellas pertenece una observacion dada.

Regresión Logística

(no es un problema de regresión, es de clasificación)

Es un algoritmo de aprendizaje supervisado.

En la regresión logística lo que quiero es **categorizar, clasificar**. Dado una serie de puntos quiero encontrar una función (no es una recta en este caso) que separe los puntos en dos conjuntos. Y una vez que la encontré puedo determinar para cualquier valor X futuro, el conjunto al cual pertenecerá.

Está asociado a problemas de probabilidad.

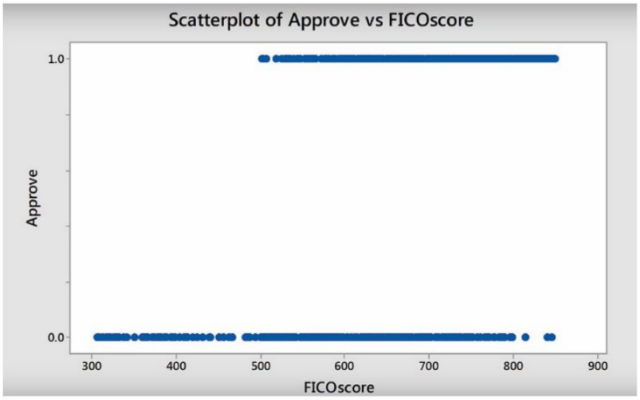
Le tenemos que decir las categorias de cada una de las observaciones que le damos para que entrene.

Ejemplo: Una persona quiere comprar una casa y para ello necesita pedir un préstamo hipotecario. Esta persona quiere saber si se lo van a otorgar o no. Pero el único dato fehaciente que tiene es su puntaje crediticio, el cual es de 720.

La única información que se tiene es una lista de puntajes crediticios de otras 1000 personas con el resultado del otorgamiento, es decir si el crédito fue otorgado: 1 o bien si no fue otorgado: 0.

creditScore	approved
655	0
692	0
681	0
663	1
688	1
693	1
699	0
699	1
683	1
698	0
655	1
703	0
704	1
745	1
702	1

Puntaje crediticio vs si esta aprobado o no. Como esto no es variable continua, hay dos valores (1 y 0) ⇒ solo hay puntos arriba o abajo



De 500 para atras → no se aprobaron

De 500 para adelante → se aprueban, pero no a todos

De 800 para adelante → hay mas arriba que abajo

Funcion sigma o sigmoidea

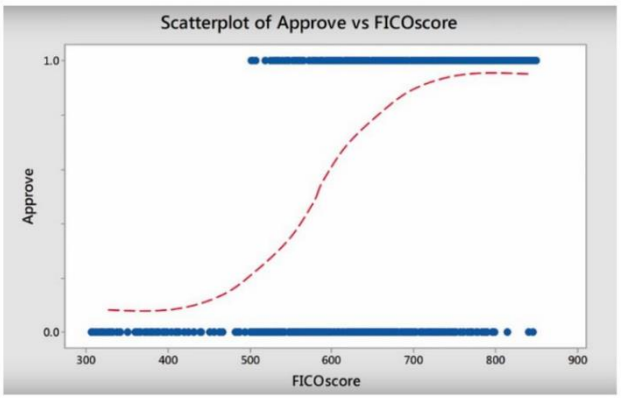
Probabiblity of a class

$$\theta(y) = \frac{1}{1 + e^{-x}}$$

Forma suave, derivable

La puedo hacer valer entre 0 y 1, para un valor dado.

Esta funcion me puede determinar mas o menos la probabilidad.



Acá podemos ver que la curva, para cada valor de X, asigna un valor entre 0 y 1 que indica la probabilidad de que el préstamo sea otorgado. Como la curva es creciente, la probabilidad será más alta cuanto el score esté más cerca de 850 y será más baja cuando el score esté cerca de 300. **Encontrar los parametros óptimos para esta curva consiste en construir un estimador de regresión logística**.

Métodos de clusterización

Clustering

En este tipo de problemas se trata de **agrupar los datos**.
Agruparlos de tal forma que queden definidos N conjuntos distinguibles, aunque no necesariamente se sepa que signifiquen esos conjuntos. El agrupamiento siempre será por características similares.

No tenemos variables de salida.

K-Means

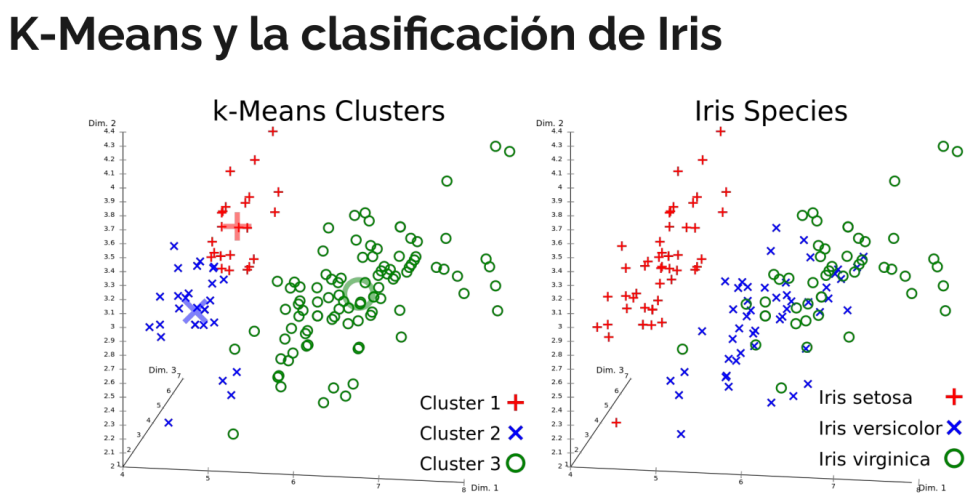
Es un algoritmo de aprendizaje no supervisado. El mismo determina como se van a agrupar las observaciones que tenemos.

Pasos:

1. El usuario decide la cantidad de grupos
2. K-Means elige al azar K centroides
3. Decide qué grupos están más cerca de cada centroide. Esos puntos forman un grupo
4. K-Means recalcula los centroides al centro de cada grupo
5. K-Means vuelve a reasignar los puntos usando los nuevos centroides. Calcula nuevos grupos
6. K-means repite punto 4 y 5 hasta que los puntos no cambian de grupo.

Va a calcular la distancia euclidea de esos puntos al centroide. Por defecto es euclidea, pero se puede cambiar → es un hiperparametro

Ejemplo: Conjunto de datos Iris: setosa, versicolor, virginica
Datos para cuantificar la variación morfológica de la flor Iris de tres especies relacionadas.
Cada punto es en R4
La variable dependiente es la clase: setosa, versicolor, virginica



Clusters de K-Means vs Las especies reales

¿Cómo saber cuántos conjuntos elegir?

A veces es obvio cuantos clusters elegir de antemano.

Si estamos trabajando con el conjunto MNIST claramente son 10 conjuntos los que tengo. Si estoy trabajando con el conjunto IRIS serán 3.

Cuando no es tan obvio, hay varias tecnicas para elegirlo.

Hiperparametro y parametros

Hiperparametro: son los parámetros que le vamos a pasar al algoritmo de entrenamiento. No son del modelo

En Kmeans los parámetros son los centroides → los parametros es lo que hay que ajustar

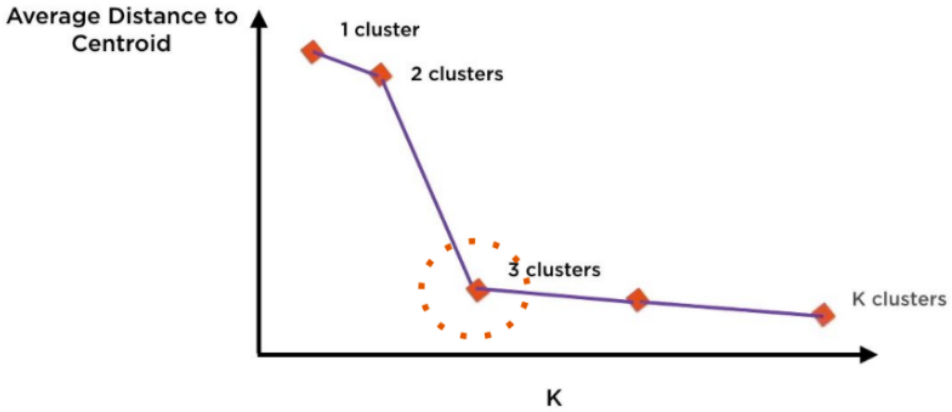
En Kmeans la distancia es un hiperparametro y también lo es el número de clusters K.

Regla del codo (Elbow Method)

Elegimos un rango, ejemplo 1 a 10, y para cada valor:

- Para cada centroide calculamos la distancia promedio

El gráfico tiene un "codo"



Método Silhouette

Elegimos un rango, ejemplo 1 a 10, y para cada valor:

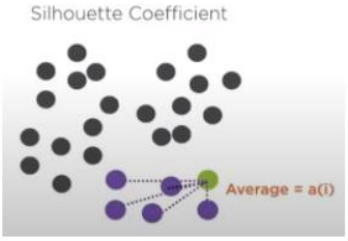
- Para cada valor de K graficamos la **silhouette**
  i. El mejor valor posible es silhouette = 1
  ii. El peor valor posible es silhouette = -1

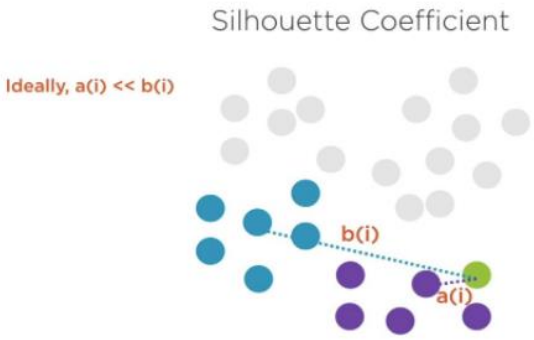
Primero, tendremos que calcular el coeficiente de Silhouette

Coeficiente de Silhouette

Cada punto en el conjunto de datos tiene un coeficiente de Silhouette.
Para calcular este coeficiente necesitamos calcular a(i) y b(i)

a(i) = distancia promedio del punto i a cada uno de los puntos de su cluster
b(i) = distancia promedio del punto i a cada uno de los puntos del cluster más cercano a su propio cluster





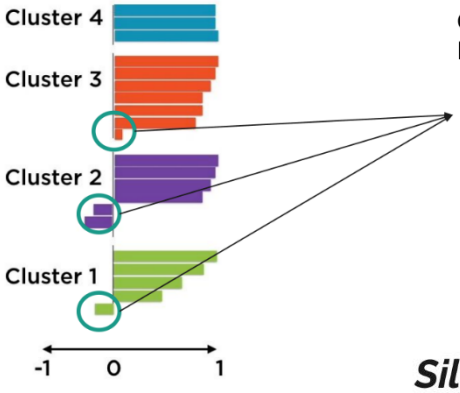
Si a(i) > b(i) i está posiblemente mal clasificado.

¿Tiene sentido que la distancia promedio a los demás puntos de su cluster sea mayor que la distancia promedio a los puntos de otro cluster? → NO

$$s(i) = \frac{b(i) - a(i)}{\text{El mayor de } (b(i) \text{ o } a(i))}$$

En el peor de los casos s(i) es -1

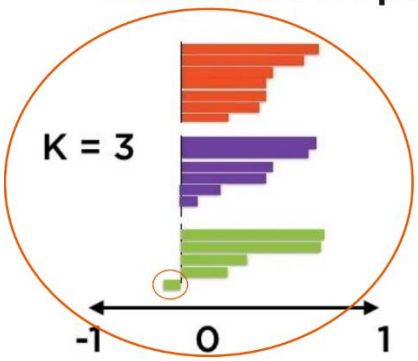
Silhouette Plot

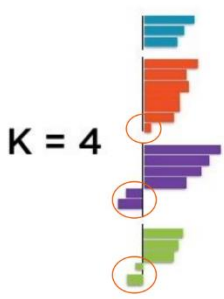


Calculamos s(i) para cada punto
Lo graficamos para identificar *outliers*

Outliers

Silhouette Plot para buscar el mejor K





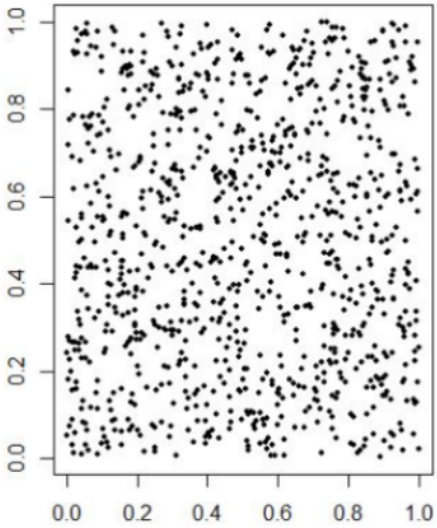
K=3, es decir con 3 clusters se ajustan mejor los valores del conjunto.

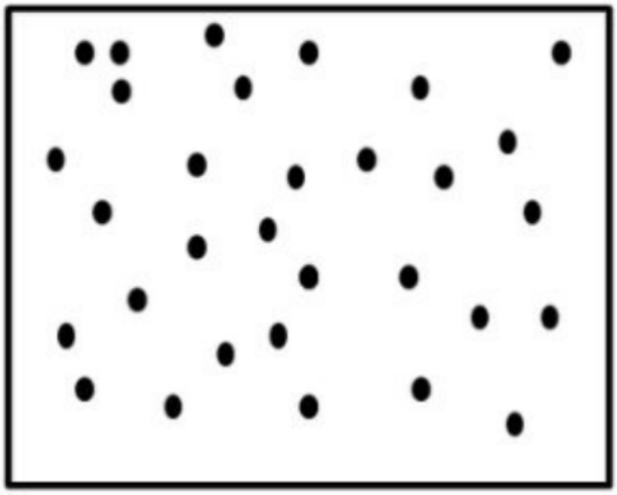
Estadística de Hopkins

Se utiliza para evaluar la tendencia de agrupación de un conjunto de datos midiendo la probabilidad de que un conjunto de datos dado sea generado por una distribución de datos uniforme.

En otras palabras, prueba la **aleatoriedad** espacial de los datos.

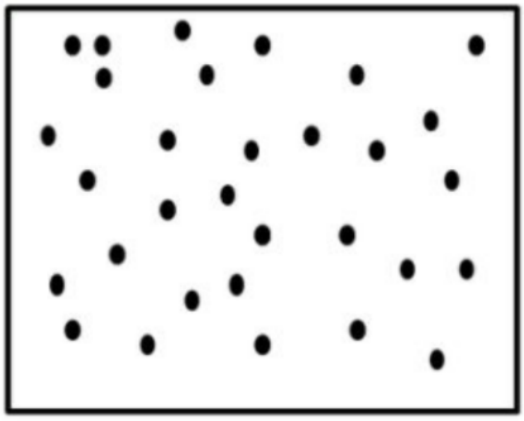
Veamos una muestra uniforme aleatoria, para ver que se puede deducir a simple vista:

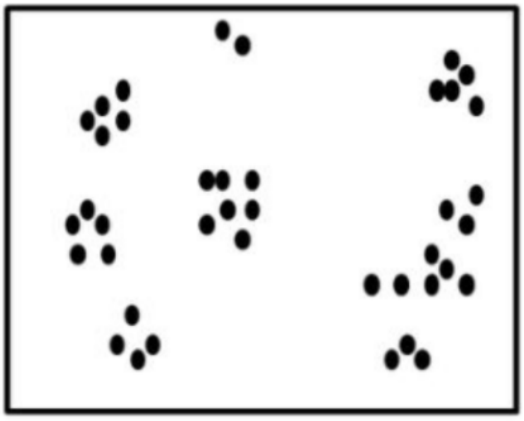




¿Vemos "tendencia" al agrupamiento aquí? Vemos clusters? NO. No son evidentes. No hay tendencia al agrupamiento.

¿Qué observamos entre estas dos muestras? ¿Cuál tiene tendencia al agrupamiento?





En el de la derecha si podemos ver unos clusters

La idea es comparar una muestra cualquiera con una muestra uniforme (creada de forma aleatoria) y ver cómo se distribuyen los ejemplos (los puntos) en dicho espacio.

Sea D un conjunto de datos reales:

1. Tomar una muestra uniformemente de n puntos (p1 ,..., pn) de D.
2. Calcular la distancia, xi , de cada punto real a cada vecino más cercano.
   a. Para cada punto pi ∈ D, encuentre su vecino más cercano pj
   b. Calcular la distancia entre pi y pj y llámela xi = dist(pi , pj)
3. Generar un conjunto de datos simulados (randomD) extraído de una distribución uniforme aleatoria con n puntos (q1, ... , qn) y la misma variación que el conjunto de datos reales original D.
4. Calcular la distancia yi desde cada punto artificial hasta el punto de datos real más cercano.
   a. Para cada punto qi ∈ randomD , encuentre su vecino más cercano qj en D
   b. Ccalcular la distancia entre qi y qj y llámela yi = dist(qi , qj)
5. Calcule la estadística de Hopkins (H) como: la distancia media del vecino más cercano en el conjunto de datos aleatorios dividida por la suma de las distancias medias del vecino más cercano en el conjunto de datos real y simulado.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

¿Cómo interpretar la estadística de Hopkins?

Si D está distribuida de forma uniforme, entonces ∑xi y ∑yi serían muy parecidos, entonces H sería aproximadamente ½ (0.5) ⇒ no hay esperanza de encontrar clusters

Pero si hay clústeres en D, las distancias de los puntos artificiales ∑yi serían mucho más grandes que las distancias de los puntos reales ∑xi ⇒ por lo tanto H sería mayor que 0.5.

Valor de H superior a 0,75 indica una tendencia a la agrupación en un nivel de confianza del 90%

Hipótesis que maneja Hopkins:

● **Hipótesis nula:** el conjunto de datos D se distribuye uniformemente
→ no hay clusters significativos

● **Hipótesis alternativa:** el conjunto de datos D no está uniformemente distribuido
→ contiene clusters significativos

Podemos realizar la prueba de la estadística de Hopkins de forma iterativa, utilizando 0,5 como umbral para rechazar la hipótesis alternativa.
● Si H < 0,5, es poco probable que D tenga conglomerados estadísticamente significativos.
● Si el valor de la estadística de Hopkins es cercano a 1, entonces podemos rechazar la hipótesis nula y concluir que el conjunto de datos D es significativamente un dato agrupable.

Indica si tiene sentido utilizar algun metodo de agrupamiento (ej K-Means), o no.