Introducción a la ciencia de datos

Ing. Juan M. Rodríguez

Model


Ciencia de datos: Modelos
Datamining
Machine Learning
Regresión Lineal
Analisis discriminante
Analisis de componentes principales
Id3
KMeans
Bayes Naive
Redes Neuronales
Support Vector Machines (SVM)
Estadistica (metodos del 1700)
Formulas
Inteligencia Artificial
Entrenamiento

Variables

- Variables Independientes (entradas)
- Variables dependientes (salidas, categorías)

Variables

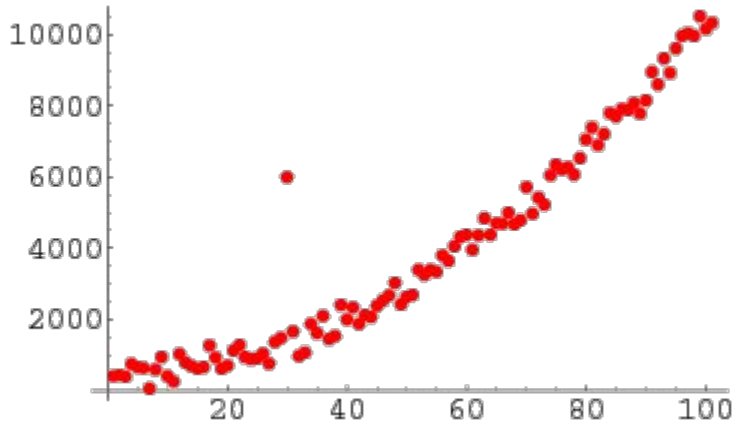
Variables Independientes:

- Cualitativas
  - Texto
    - Nominales (categorías, ejemplo: países)
    - Ordinales (poco, mucho, muchísimo)
  - Númericas
    - Nominales
    - Ordinales
- Cuantitativa
  - Discreta
  - Continua

Variables y tipos de problemas

1. Si la variable dependiente es **cualitativa**, el tipo de problema es de **clasificación**
2. Si la variable dependiente es **cuantitativa**, el problema es de **regresión**
3. Si **NO hay variable** dependientes, el problema es de **agrupamiento**

Outliers (valor atípico)



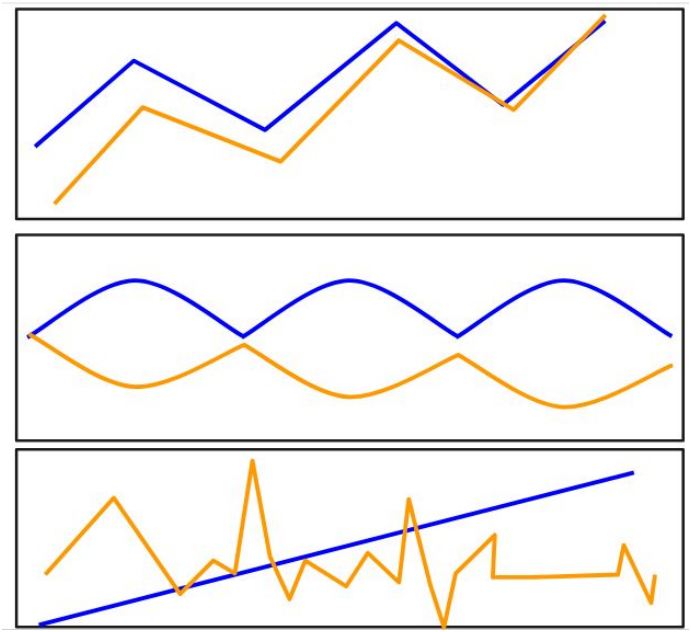
Correlación de variables

Correlación de variables

Dos variables están correlacionadas cuando varían de igual forma sistemáticamente.

Correlación de variables

- Positiva
- Negativa
- Sin correlación

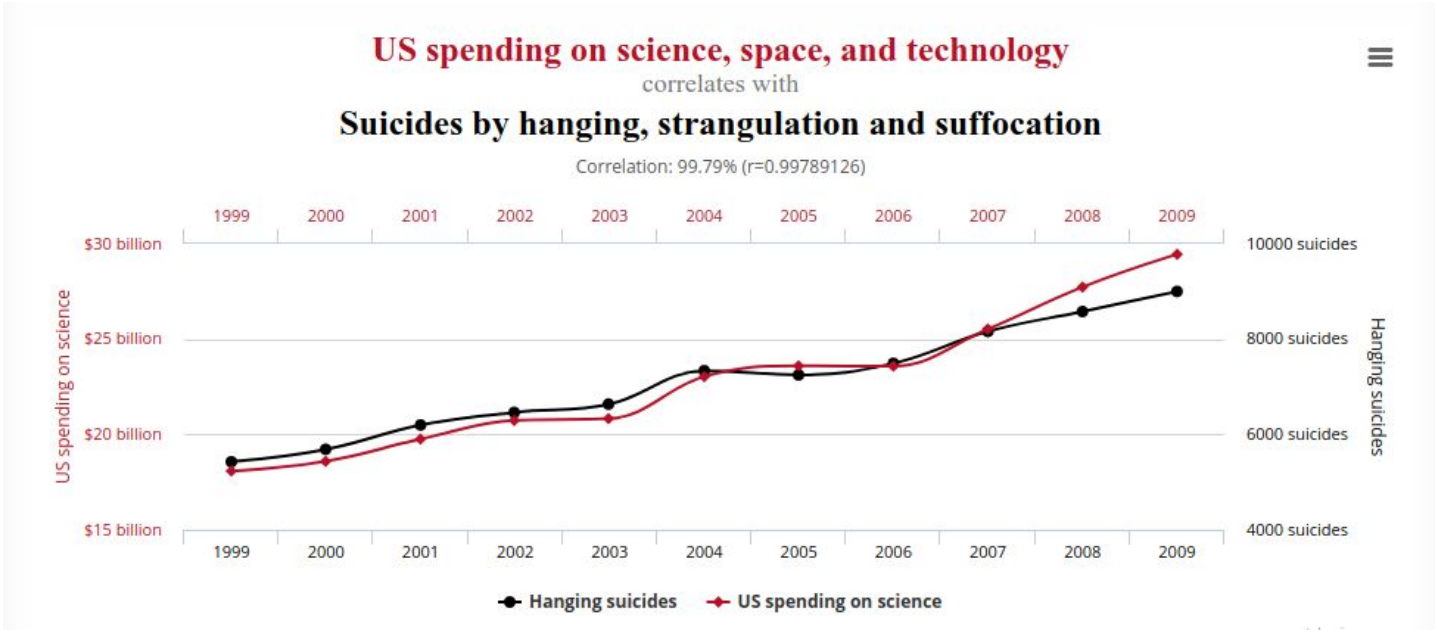


Correlación de variables

Correlación NO IMPLICA Causalidad

- Que dos variables tengan alto índice de correlación no significa que una cause la otra
- Las relaciones de causalidad son mucho más difíciles de encontrar y demostrar
- Las correlaciones pueden suceder por otros motivos como: Una tercer variable que “empuja” a ambas o simplemente azar

Ejemplos de correlaciones sin sentido

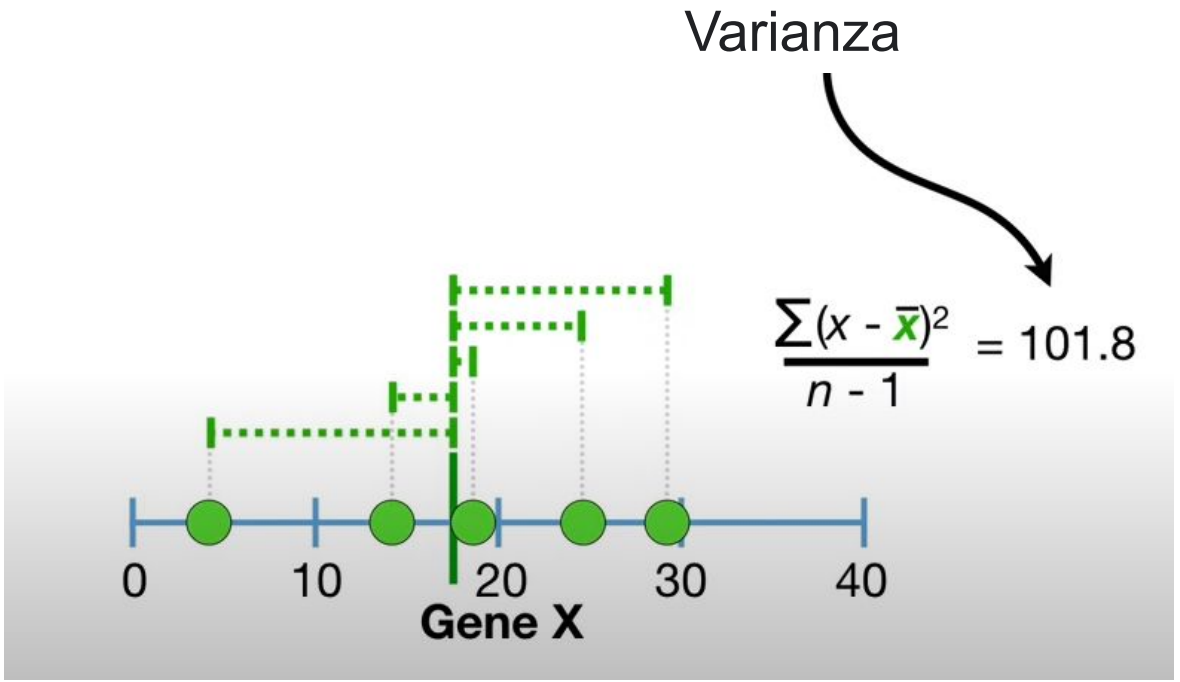


Varianza

Varianza

Promedio de la diferencia, entre todas las observaciones, respecto de su **media.**

La media es: **17.6**



Covarianza

Covarianza

En probabilidad y estadística, la **covarianza** es un valor que indica el **grado de variación conjunta de dos variables aleatorias respecto a sus medias**.

Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal o la recta de regresión.

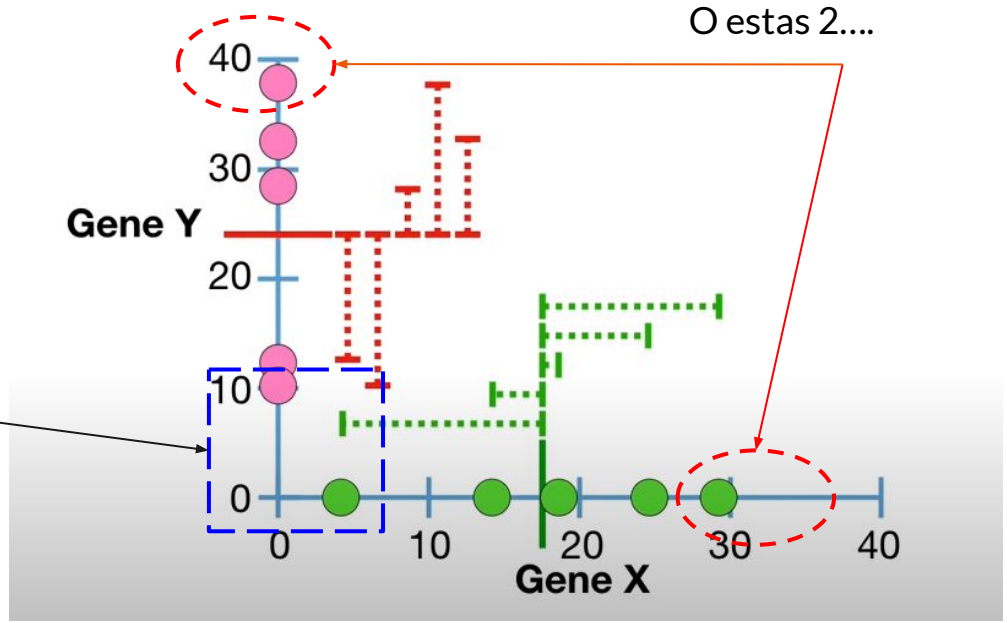
Covarianza

Dos variables.

Calculamos la varianza de cada una….

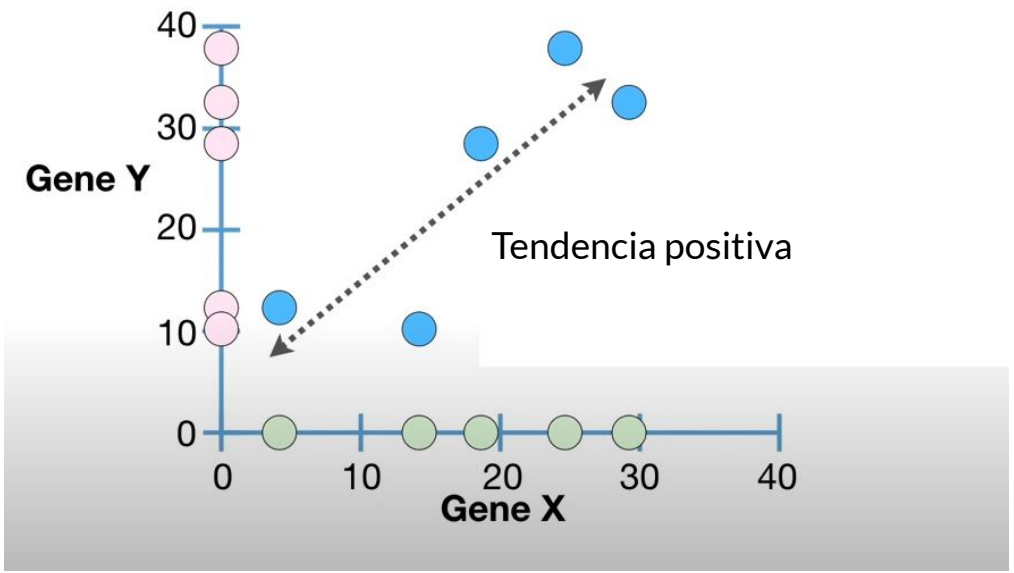
Estas dos medidas corresponden a una misma medición. A la misma observación.

¿Son sus variaciones respecto de la media simiares?

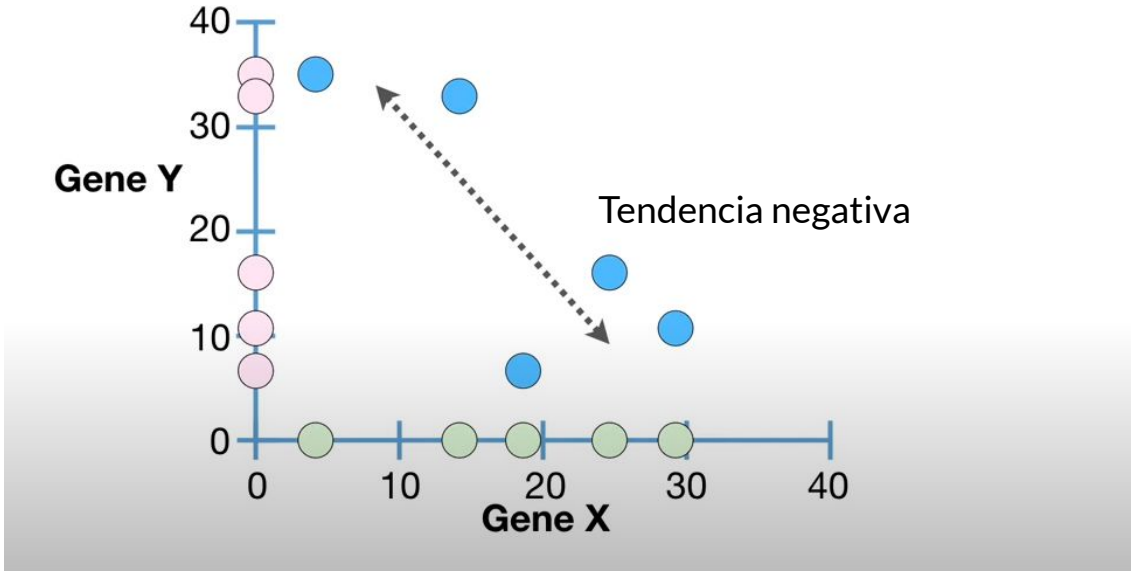


Covarianza

Cómo los datos en Y y X, pertenecen a una misma medición podemos graficarlos en 2D y ver si hay alguna tendencia…

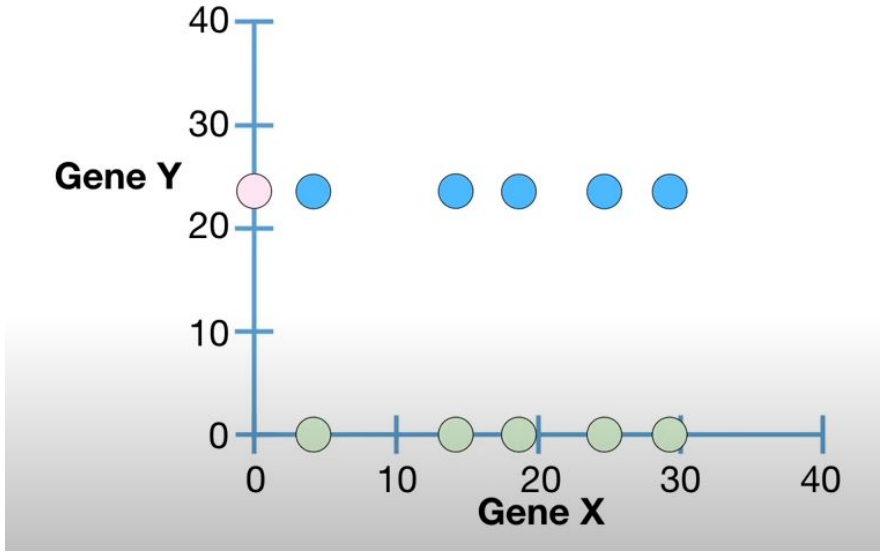


Covarianza



Covarianza

No hay tendencia



Correlación de Pearson

Correlación de Pearson

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

donde

- σ_{XY} es la covarianza de (X, Y)
- σ_X es la desviación estándar de la variable X
- σ_Y es la desviación estándar de la variable Y

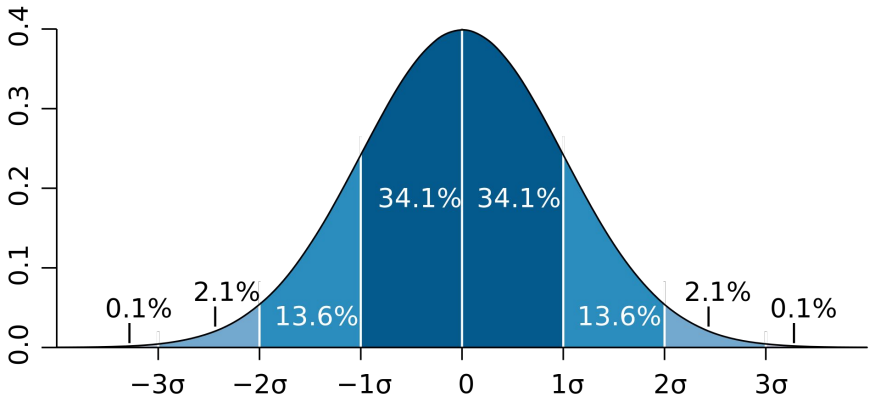
Correlación de Pearson

Para 2 variables podemos medir su correlación lineal con el coeficiente de correlación r (Pearson). Este coeficiente, es una función que mide cuán relacionada estan 2 variables de forma lineal.

- Si da 0 NO existe correlación
- Si da 1 Están relacionadas linealmente de forma perfecta (todos los puntos están en una línea)
-
- Si da -1 Existe una correlación negativa perfecta.

Correlación de Pearson: desvio estandar

Es una medida que se utiliza para **cuantificar la variación o la dispersión de un conjunto de datos numéricos**.
Una **desviación estándar baja** indica que la mayor parte de los datos de una muestra tienden a estar agrupados **cerca de su media** (también denominada el valor esperado), mientras que una desviación estándar alta indica que los datos se extienden sobre un rango de valores más amplio.



Métodos de regresión

Regresión

La primera forma de regresión lineal documentada fue el método de los **mínimos cuadrados** que fue publicada por Legendre en 1805, Gauss publicó un trabajo en donde desarrollaba de manera más profunda el método de los mínimos cuadrados,

Regresión

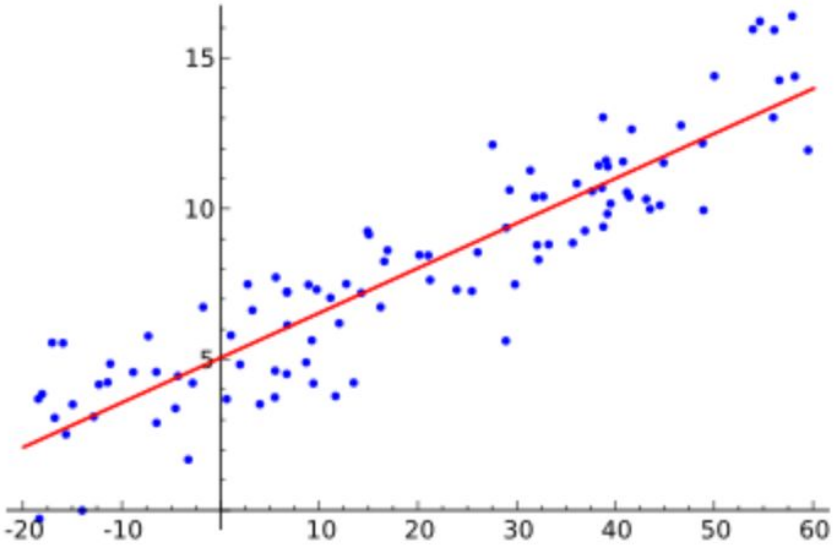
El concepto de regresión proviene de la genética y fue popularizado por Sir Francis Galton a finales del siglo XIX con la publicación de Regression towards mediocrity in hereditary stature.7 Galton observó que las c**aracterísticas extremas** (por ejemplo, la altura) de los padres no se transmiten por completo a su descendencia. Más bien, las características de la descendencia retroceden hacia un punto mediocre (un punto que desde entonces ha sido identificado como la media)

Regresión

Buscamos predecir un valor en un rango continuo, para ciertos valores de entrada. Ejemplos:

- Temperatura
- Valor de una propiedad

Regresión lineal o ajuste lineal



Pseudo-code

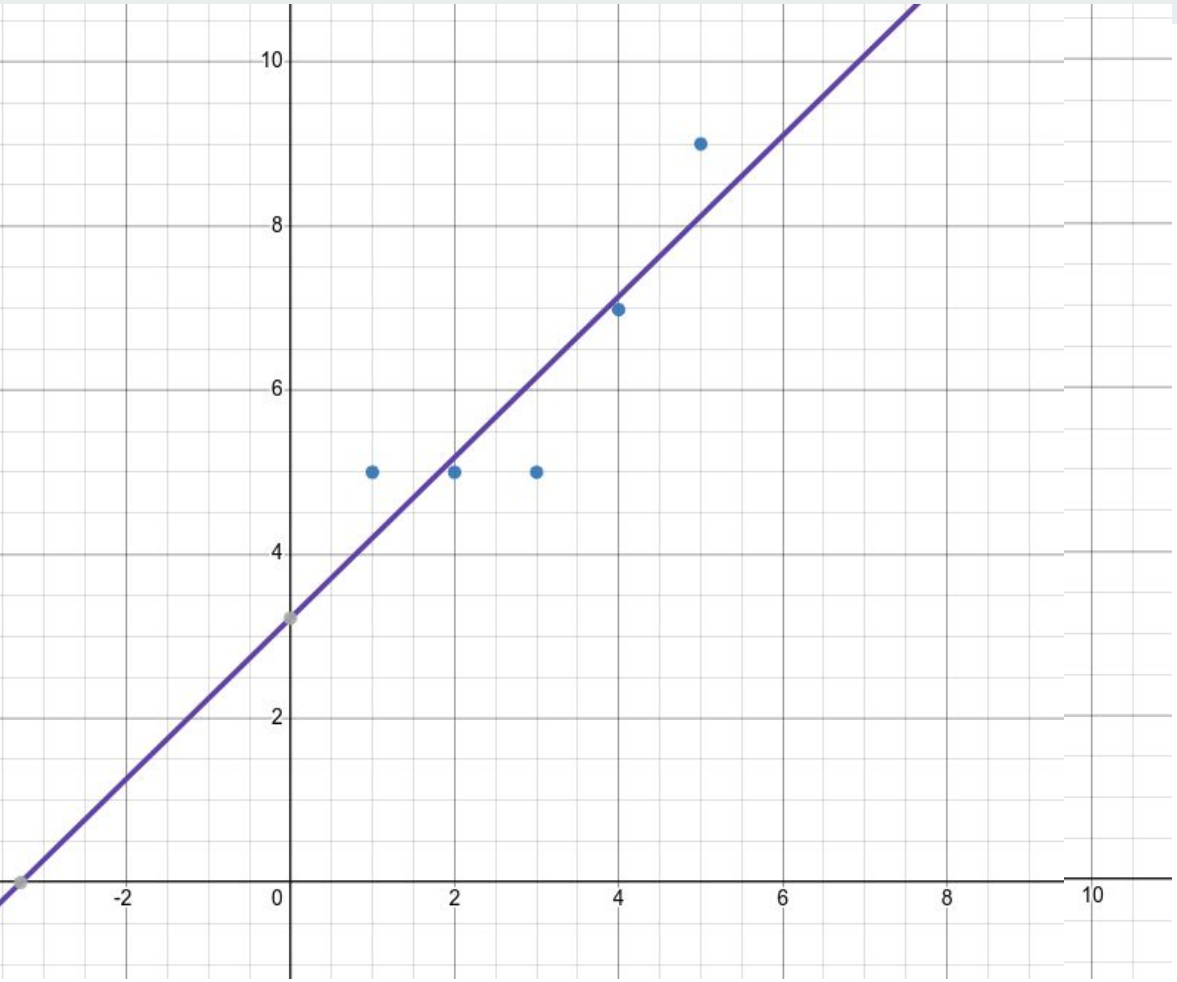
```
vars
xarray = [            1, 2, 3, 4, 5      ],
yarray = [            5, 5, 5, 6.8, 9    ],
x = y = xy = xx = a = b = resultado = 0,
cantidad = xarray.length,
```

```
for (i = 0; i < cantidad; i++) {
      x += xarray[i];
      y += yarray[i];
      xy += xarray[i]*yarray[i];
      xx += xarray[i]*xarray[i];
}
b = ((cantidad * xy) - (x * y)) /
     ((cantidad * xx) - (x * x));
a = (y - (b * x)) / cantidad;
```

$y = x^*b + a$

b = 0.98

a =3.22


10
8
6
4
2
-2
0
2
4
6
8
10

Generalización

Ecuación Normal:

$$\widehat{\theta} = (X^T X)^{-1} \; X^T \; y$$

```
import numpy as np
xarray = [            1, 2, 3, 4, 5      ],
yarray = [            5, 5, 5, 6.8, 9    ]
X_b = np.c_[np.ones((5, 1)), xarray] # add x0 = 1 to each instance
theta_best = np.linalg.inv(X_b.T.dot(X_b)).dot(X_b.T).dot(yarray)
```

array([3.22, 0.98])

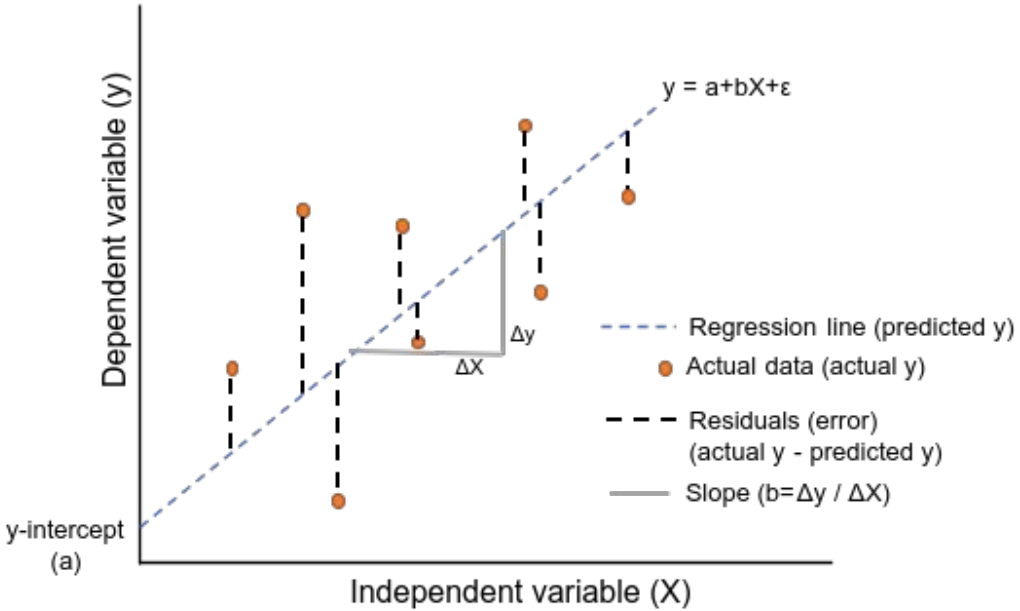
Problemas con la ecuación normal

La ecuación normal calcula el inverso de $X^T X$, que es una matriz de $(n + 1) \times (n + 1)$ (donde ***n*** es el número de características).
La complejidad computacional de invertir tal matriz es típicamente alrededor de $O(n^{2.4})$ a $O(n^3)$, dependiendo de la implementación.

En otras palabras, si se duplica el número de características, el tiempo se multiplica por aproximadamente $2^{2.4} = 5.3$ a $2^3 = 8$.

Existen otros mecanismos que buscan de forma iterativa, por aproximación y son computacionalmente menos costosos como el **descenso por gradiente**

Error en regresión



Métrica para regresión

Root Mean Square Error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Raíz del error cuadrático medio

m: número de instancias

h: función de hipótesis, es el modelo entrenado. En este caso regresión lineal

x: todos los valores de entradas, todas las columnas

Métrica para regresión

Mean absolute error (MAE)

$$\mathrm{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left| h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right|$$

Error medio absoluto

m: número de instancias

h: función de hipótesis, es el modelo entrenado. En este caso regresión lineal

x: todos los valores de entradas, todas las columnas

Métrica para regresión

$$\text{Mean Square Error}$$

$$\text{MSE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right)^2$$

Error cuadrático medio.

m: número de instancias

h: función de hipótesis, es el modelo entrenado. En este caso regresión lineal

x: todos los valores de entradas, todas las columnas

Métodos de clasificación

Clasificación

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

Regresión Logística

En la regresión logística lo que quiero es categorizar, clasificar.
Es decir que dado una serie de puntos quiero encontrar una función (no es una recta en este caso) que separe los puntos en dos, 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.

Regresión Logística

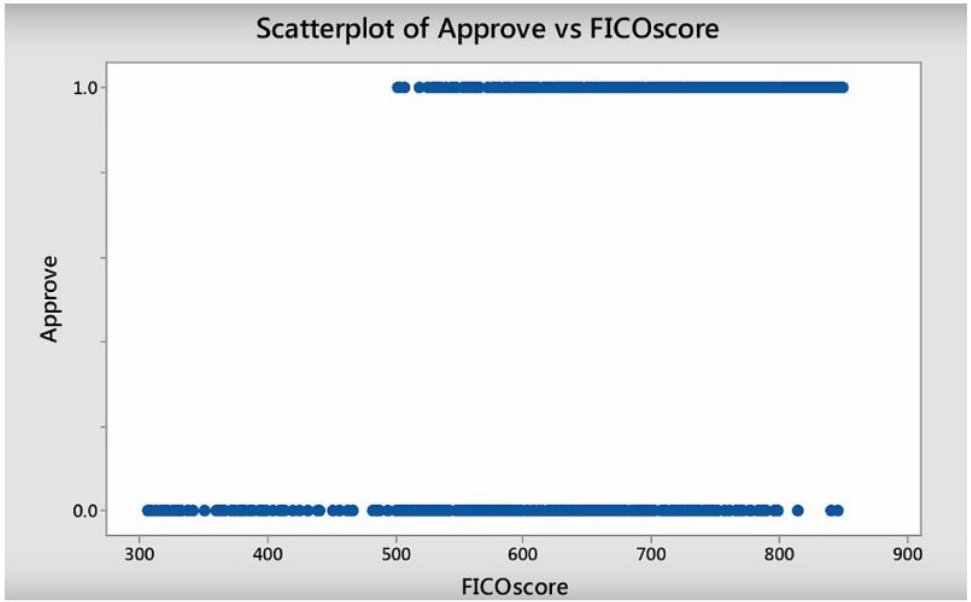
Una persona quiere comprar una casa y para ello necesita pedir un préstamo hipotecario. Esta persona quiere saber si se lo van a otorgar o no. Pero el único dato fehaciente que tiene es su puntaje crediticio, el cual es de 720

Regresión Logística

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

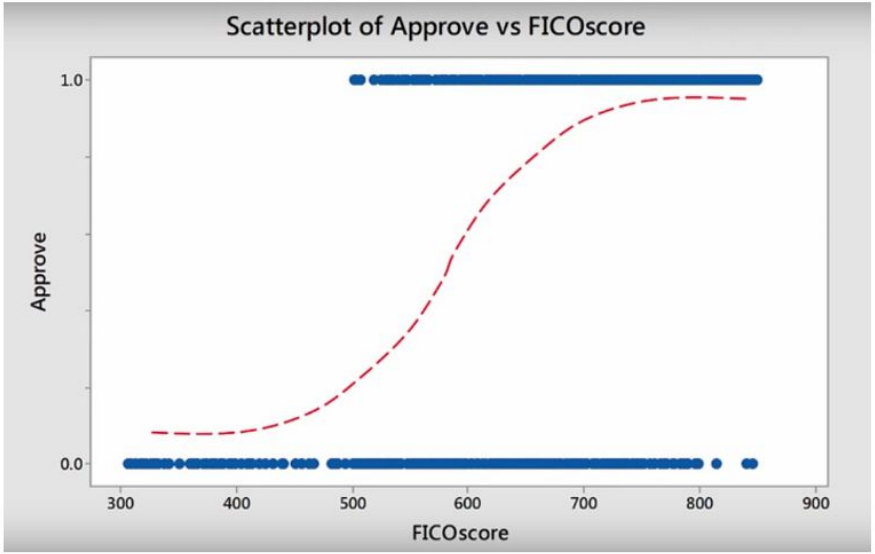
La única información que se tiene es una lista de puntajes crediticios de otras 1000 personas con el resultado del otorgamiento, es decir si el crédito fue otorgado: 1 o bien si no fue otorgado: 0.

Regresión Logística



Regresión Logística

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



Regresión Logística

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 optimos 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.

K-Means

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.

Conjunto de datos Iris: setosa, versicolor, virginica

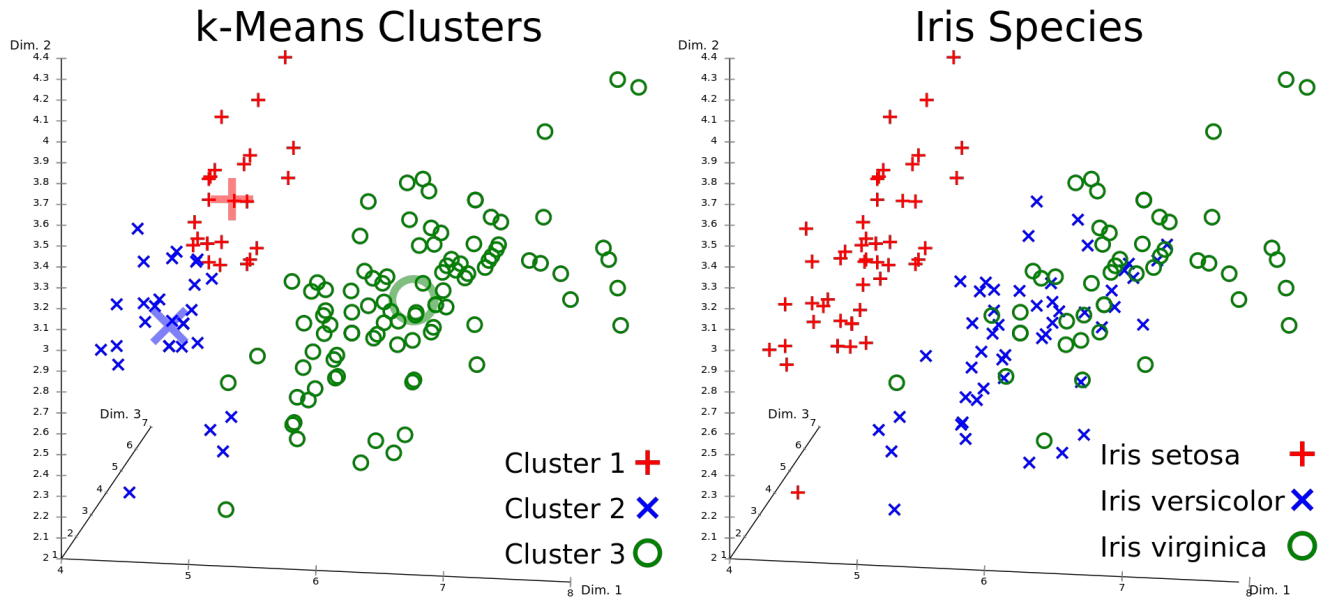
Fisher's *Iris* Data

Largo de sépalo	Ancho de sépalo	Largo de pétalo	Ancho de pétalo	Especies
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*

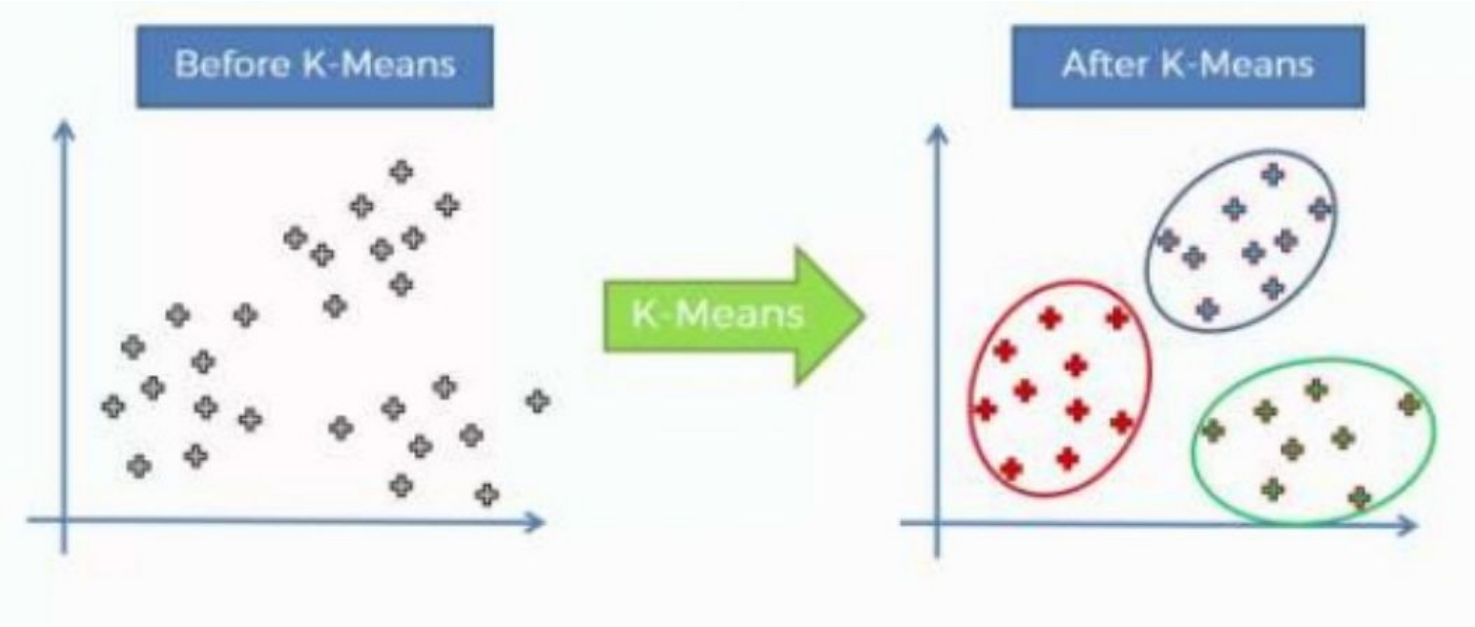
Conjunto de datos Iris: setosa, versicolor, virginica

El conjunto de datos flor Iris o conjunto de datos iris de Fisher es un conjunto de datos introducido por Ronald Fisher en un artículo de 1936.
A veces, se llama Iris conjunto de datos de Anderson porque Edgar Anderson coleccionó los datos para cuantificar la variación morfológica de la flor **Iris** de **tres especies relacionadas**.

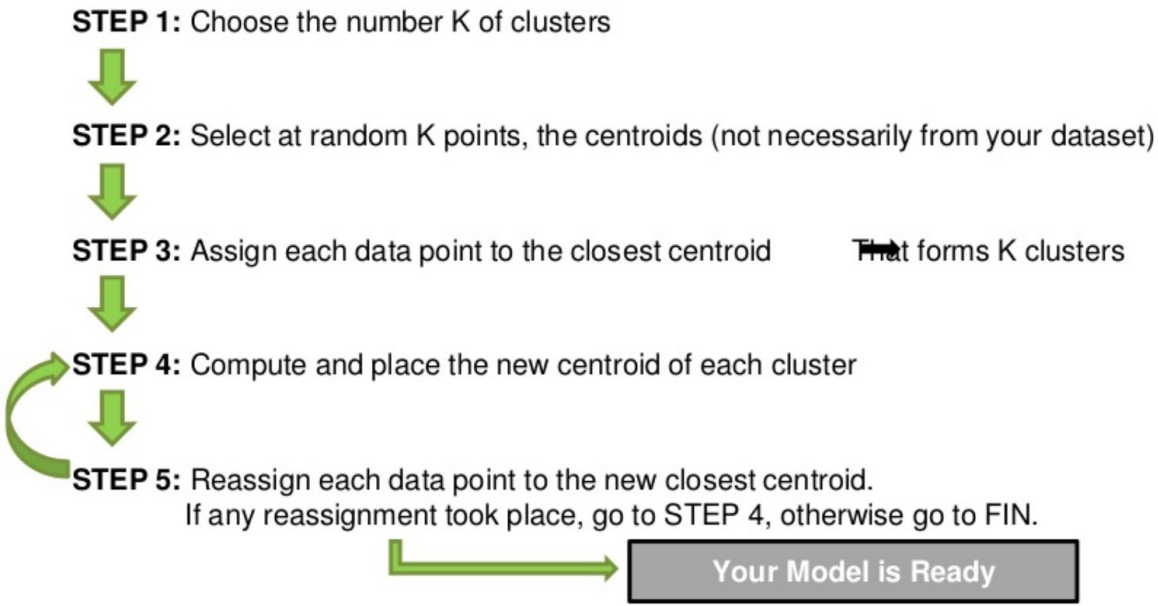
K-Means y la clasificación de Iris



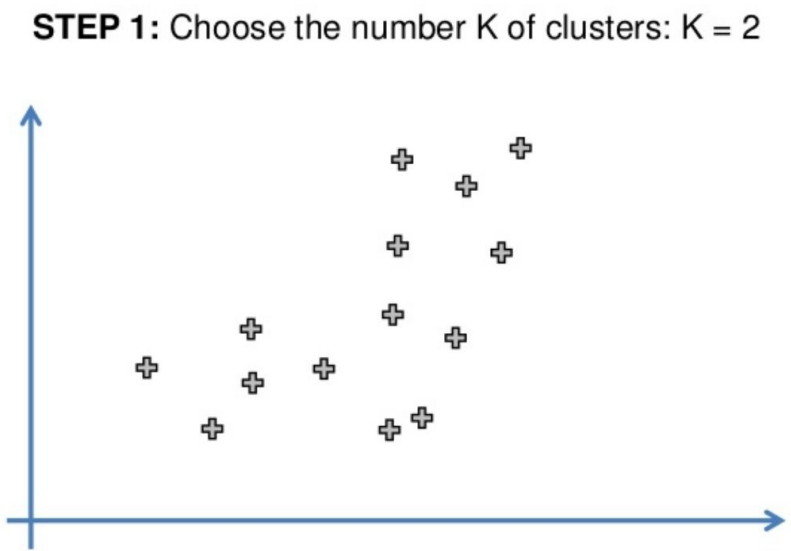
K-Means



K-Means

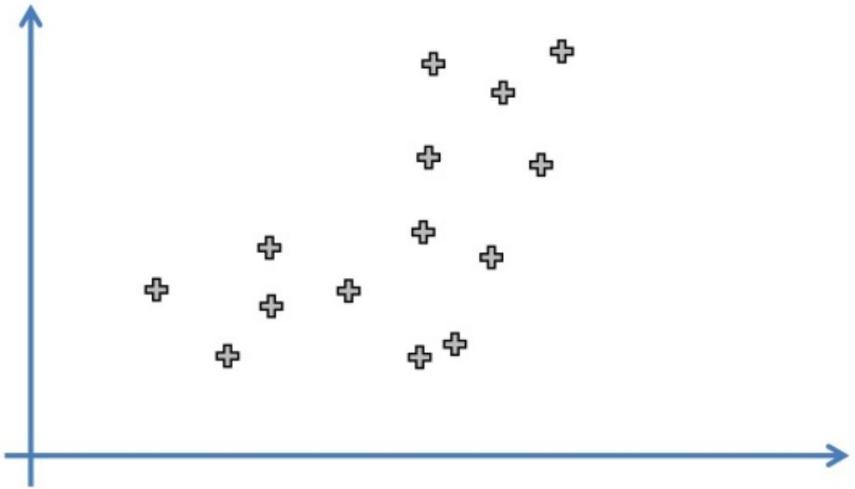

STEP 1: Choose the number K of clusters
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)
STEP 3: Assign each data point to the closest centroid
That forms K clusters
STEP 4: Compute and place the new centroid of each cluster
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.
Your Model is Ready

K-Means



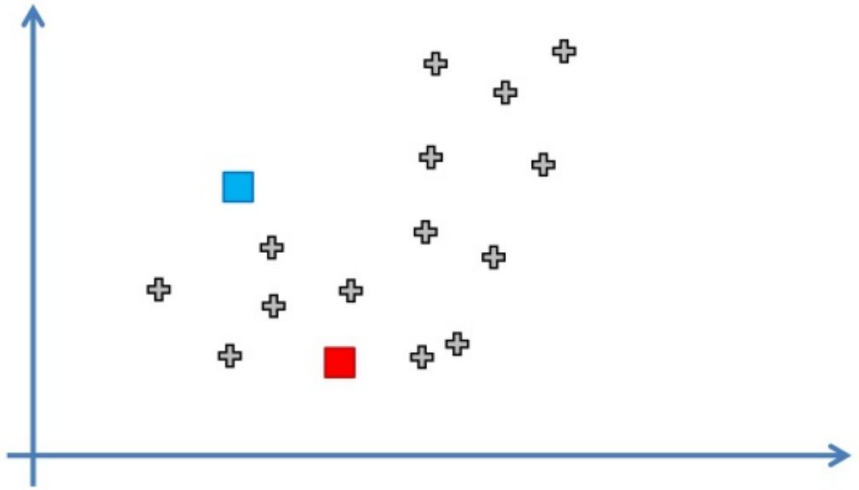
K-Means

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



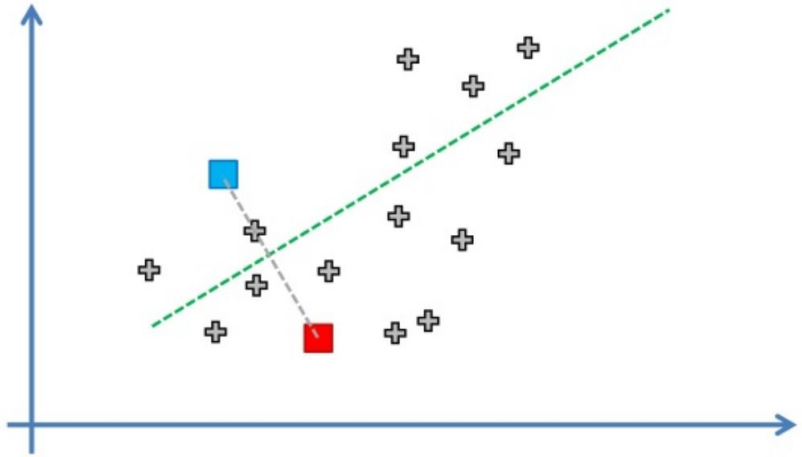
K-Means

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



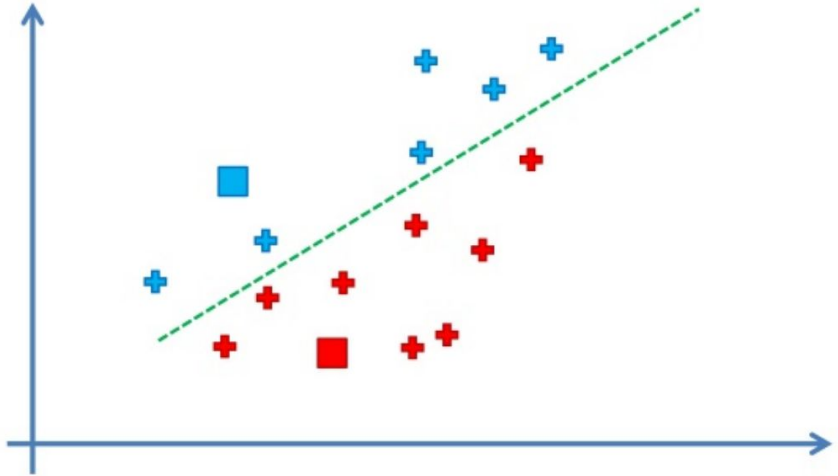
K-Means

STEP 3: Assign each data point to the closest centroid ➡That forms K clusters



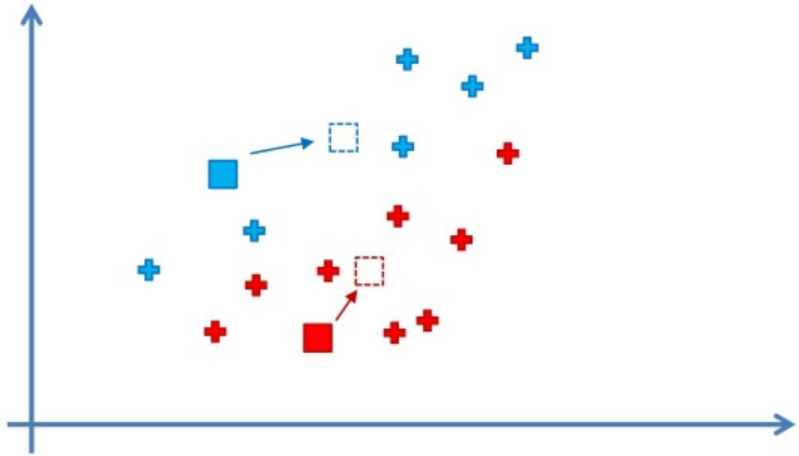
K-Means

STEP 3: Assign each data point to the closest centroid ➡That forms K clusters



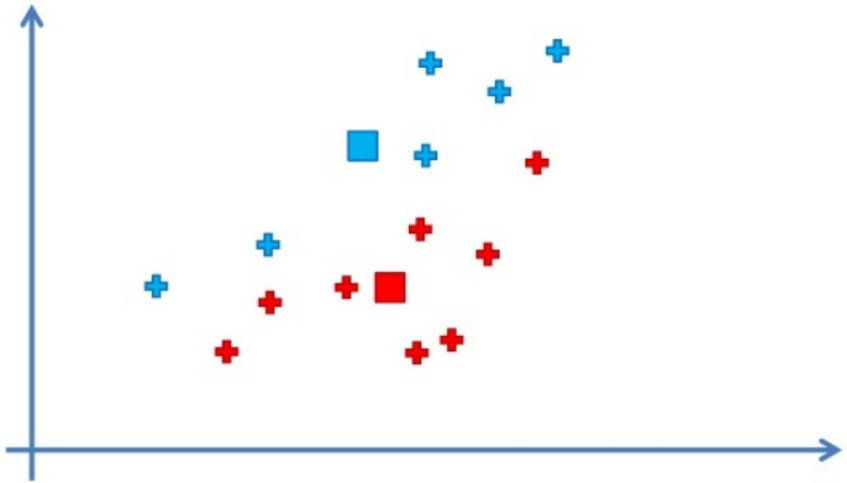
K-Means

STEP 4: Compute and place the new centroid of each cluster



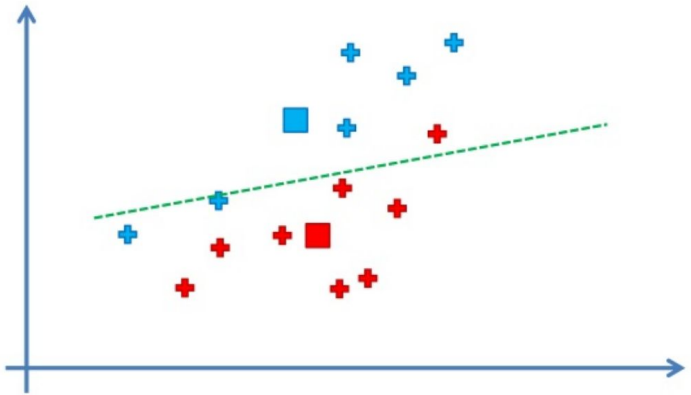
K-Means

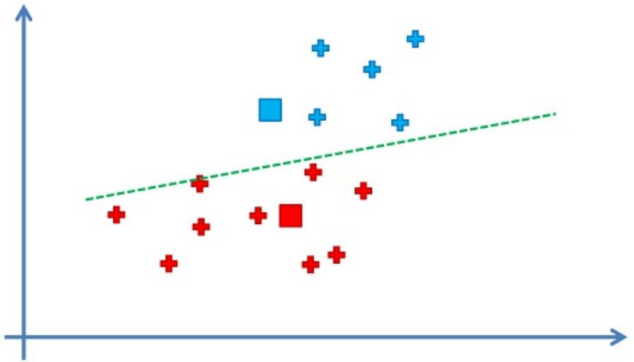
STEP 4: Compute and place the new centroid of each cluster



K-Means

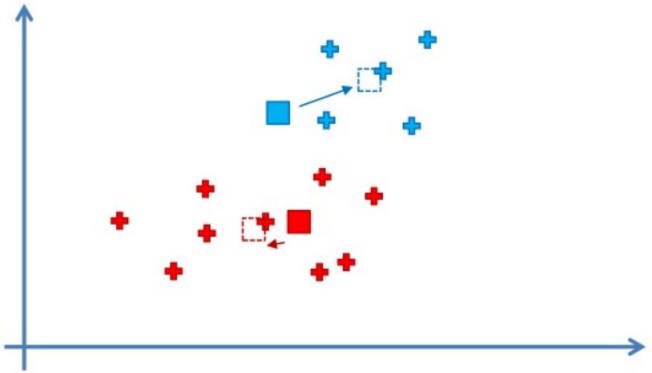
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

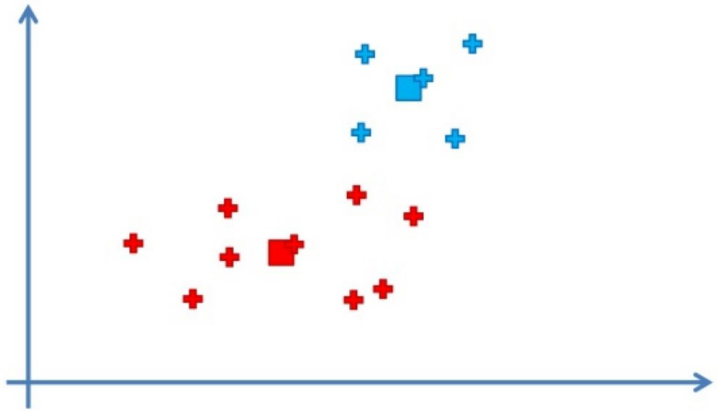




K-Means

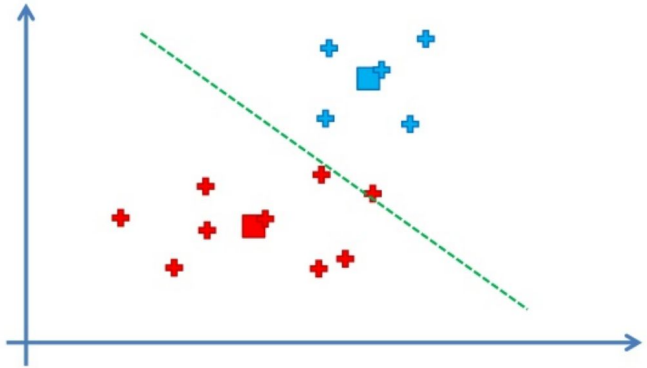
STEP 4: Compute and place the new centroid of each cluster

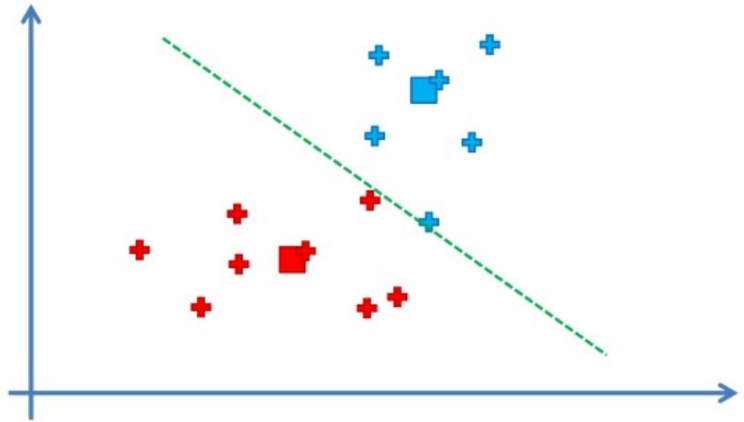




K-Means

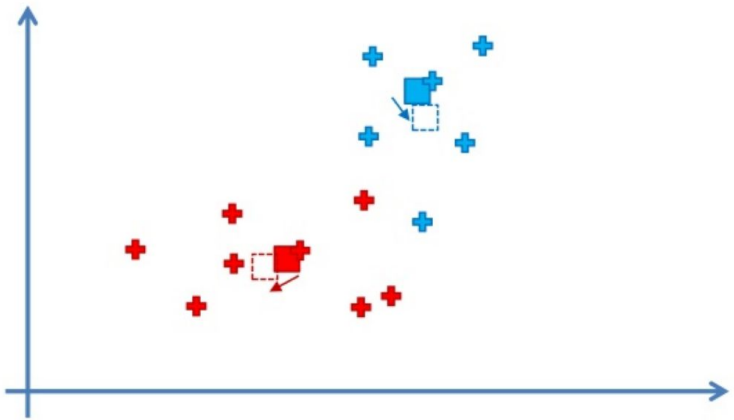
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

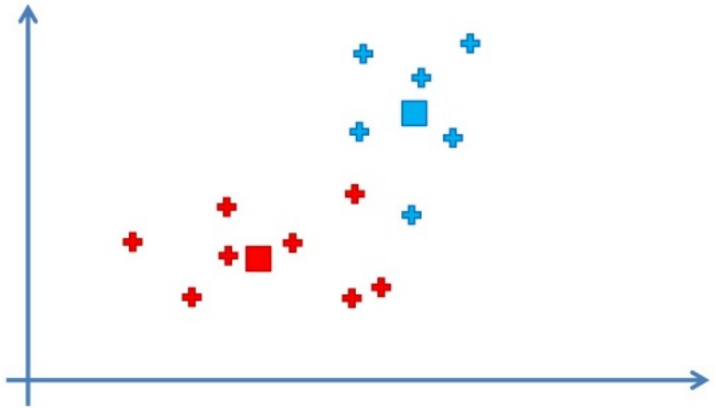




K-Means

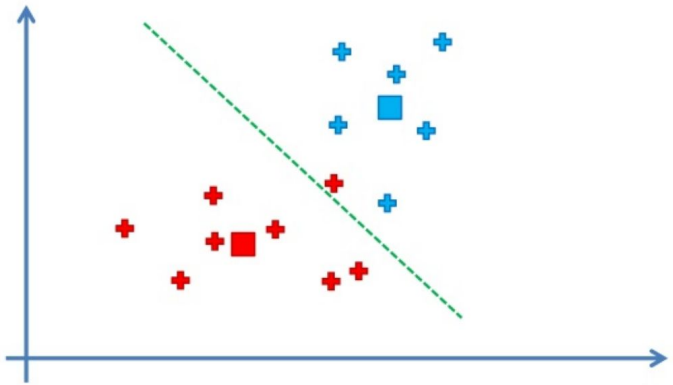
STEP 4: Compute and place the new centroid of each cluster

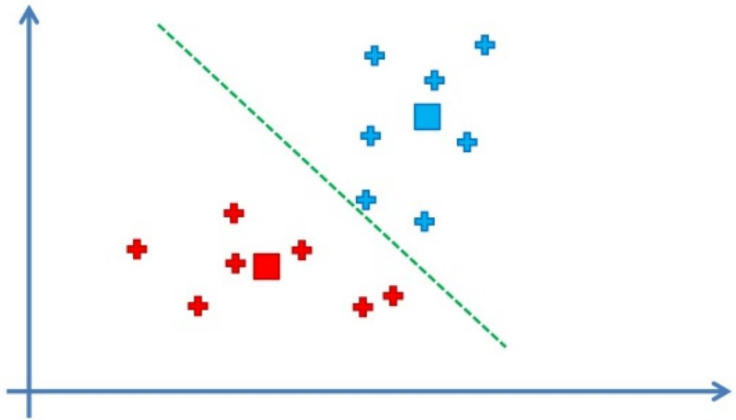




K-Means

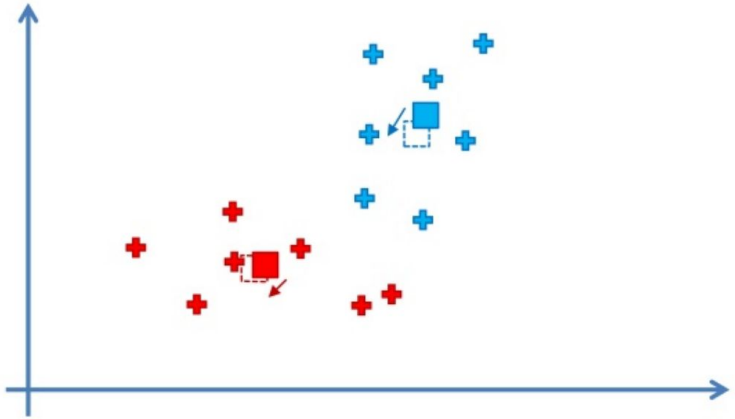
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

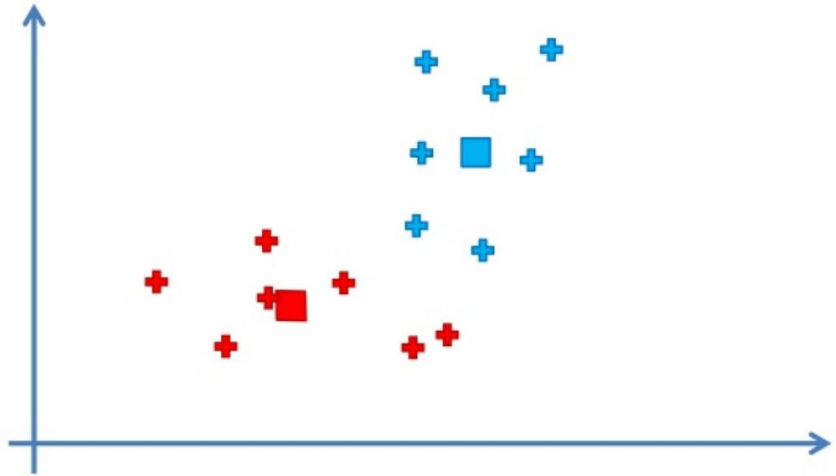




K-Means

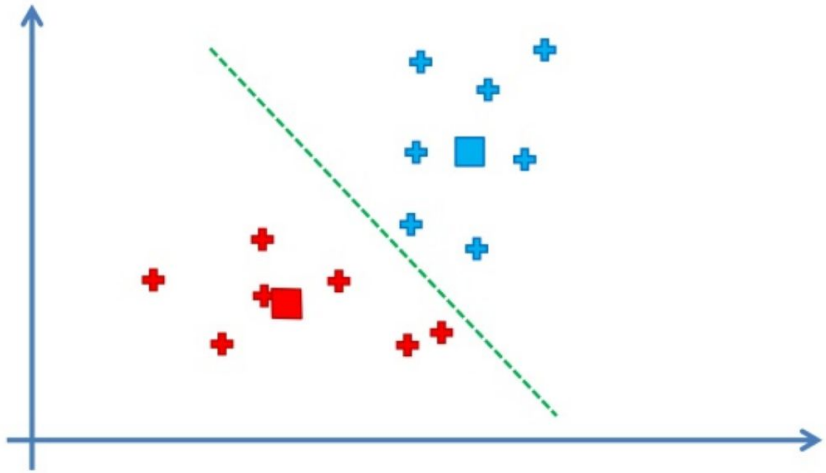
STEP 4: Compute and place the new centroid of each cluster



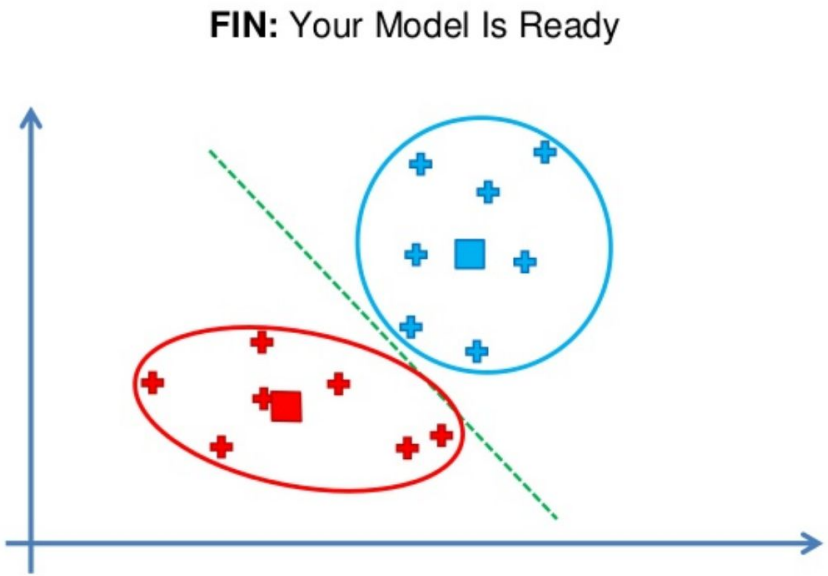


K-Means

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



K-Means



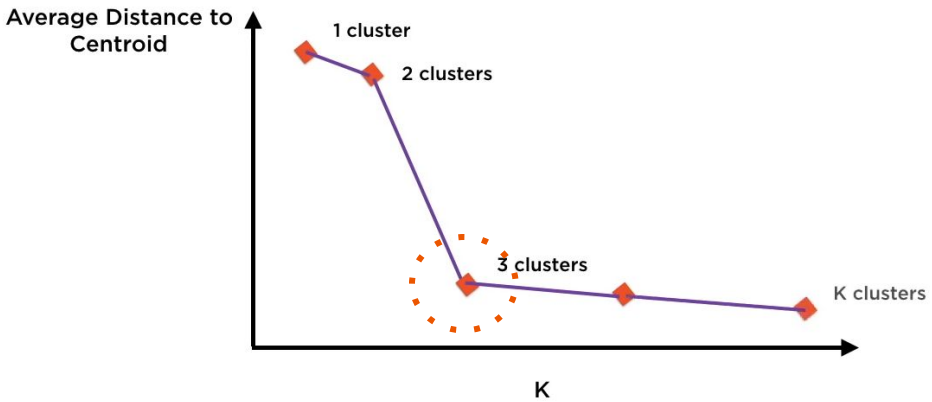
¿Cómo saber cuántos conjuntos elegir?

¿Cuántos conjuntos elegir?

A veces es obvio, si estamos trabajando con el conjunto MNIST claramente son 10 conjuntos los que tengo. Si estoy trabajando con el conjunto IRIS serán 3

Regla del codo (Elbow Method)

1. Elegimos un rango, ejemplo 1 a 10, y para cada valor:
    a. Para cada centroide calculamos la distancia promedio



El gráfico tiene un “codo”

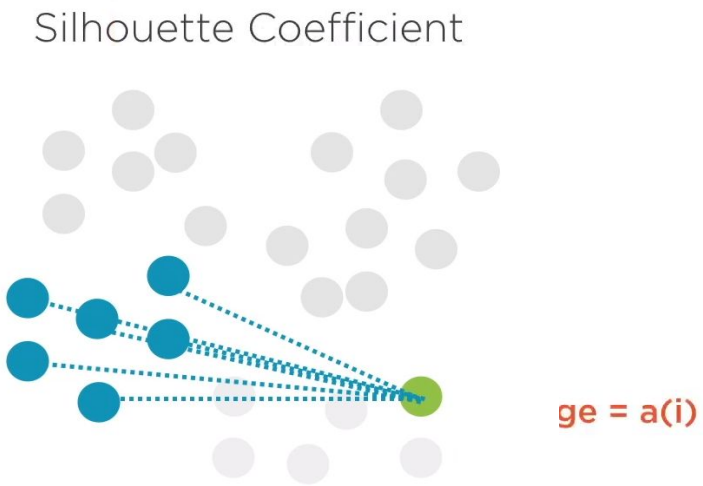
Método *Silhouette*

1. Elegimos un rango, ejemplo 1 a 10, y para cada valor:
    a. Para cada valor de **K** graficamos la ***silhouette***
        i. El mejor valor posible es *silhouette* = 1
        ii. El peor valor posible es *silhouette* = -1

Primeramente tendremos que calcular el coeficiente de Silhouette

Coeficiente de *Silhouette*

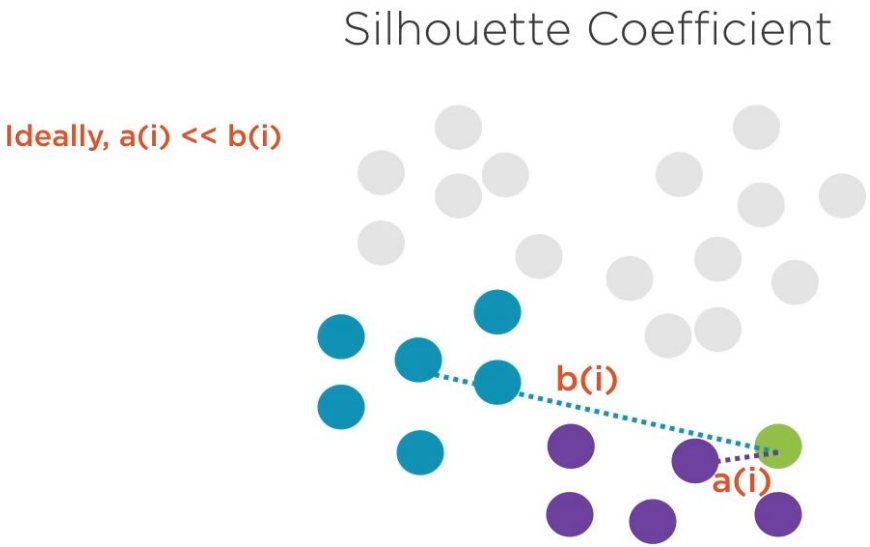
Cada punto en el conjunto de datos tiene un coeficiente de *Silhouette*.
Para calcular este coeficiente necesitamos calcular a(i) y b(i)



a(i) = distancia promedio del punto ***i*** a cada uno de los puntos de su cluster

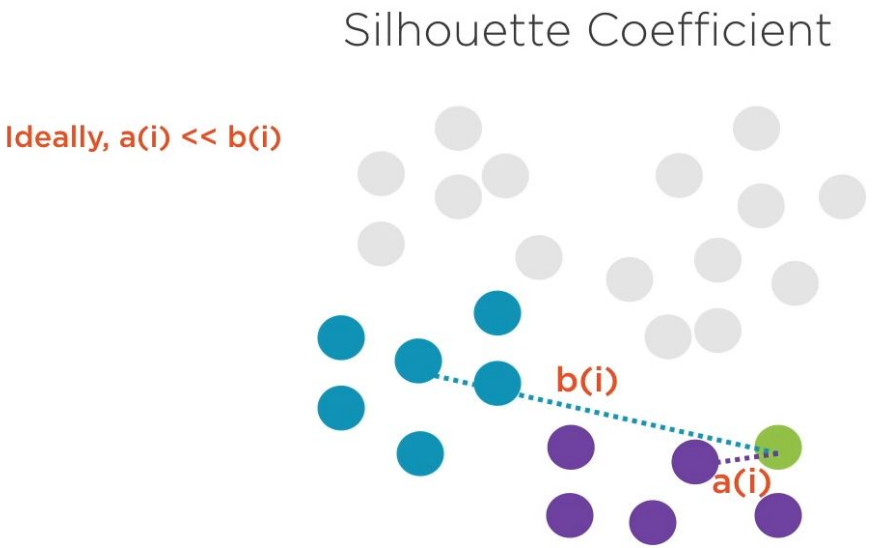
b(i) = distancia promedio del punto ***i*** a cada uno de los puntos del cluster más cercano a su propio cluster

Coeficiente de *Silhouette*



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?

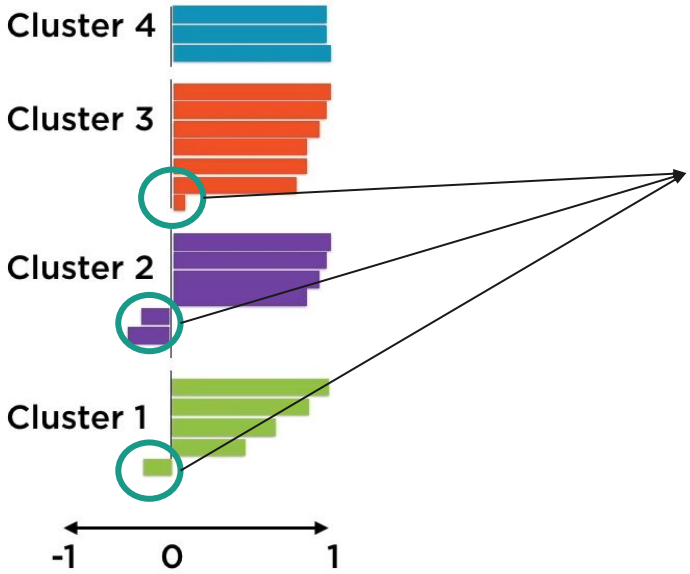
Coeficiente de *Silhouette*



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

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

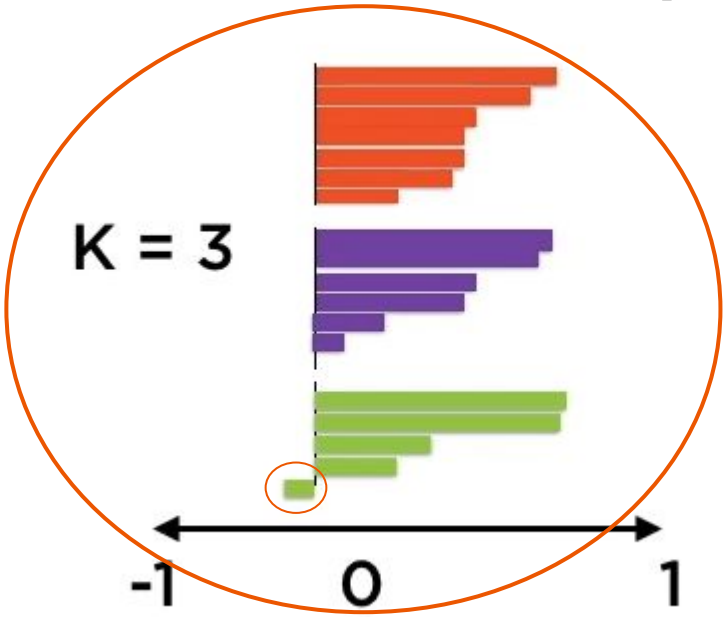
Silhouette Plot

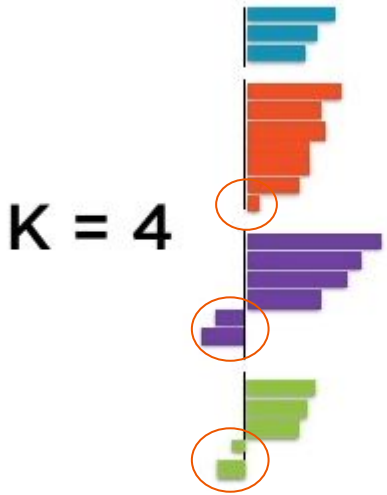


Calculamos s(i) para cada punto
Lo graficamos para identificar *outliers*

Outliers

Silhouette Plot para buscar el mejor K





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

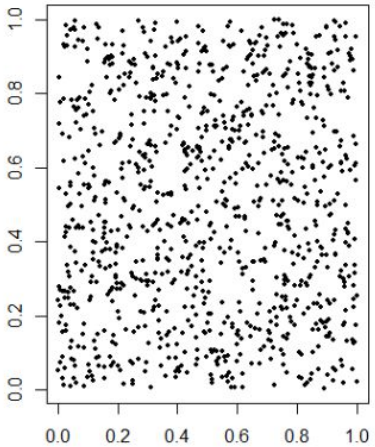
Estadística de Hopkins

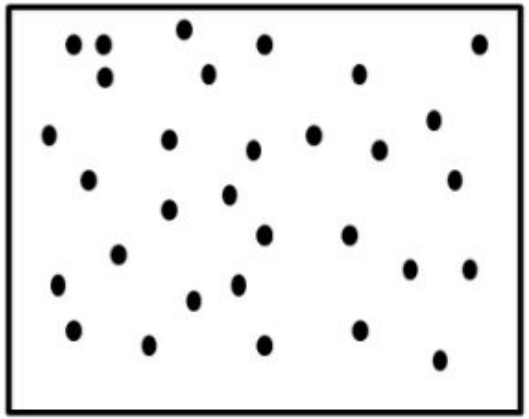
La estadística de *Hopkins* (Lawson y Jurs 1990) se utiliza para evaluar la tendencia de agrupación de un conjunto de datos midiendo la probabilidad de que un conjunto de datos dado sea generado por una distribución de datos uniforme.

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

Estadística de Hopkins

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

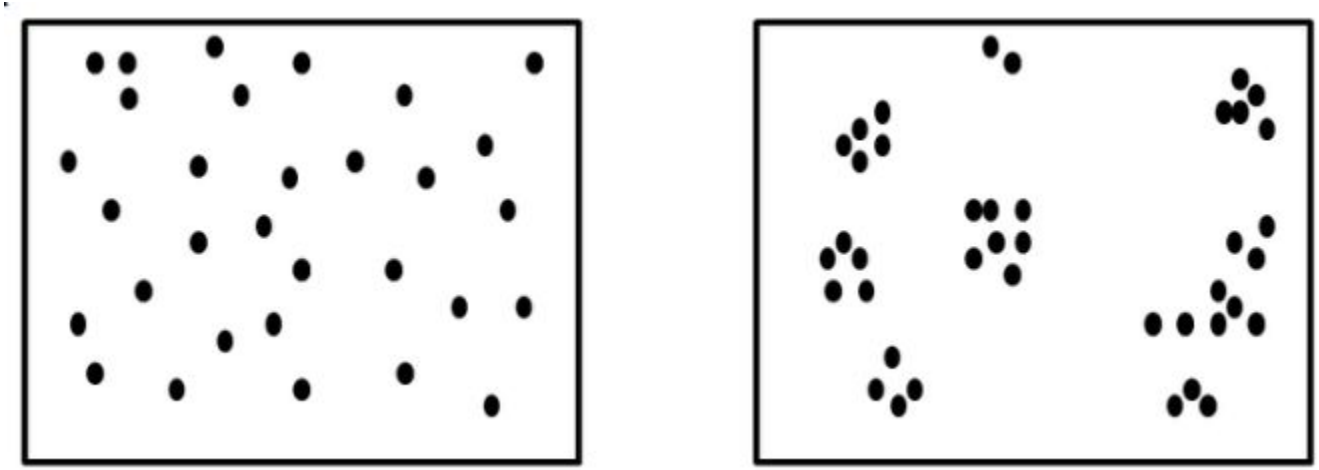




¿Vemos “tendencia” al agrupamiento aquí?

Estadística de Hopkins

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



Estadística de Hopkins

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 (p_1,…, p_n) de D.
2. Calcular la distancia, $\boldsymbol{x_i}$, de cada punto **real** a cada vecino más cercano.
   a. Para cada punto $\boldsymbol{p_i} \in$ D, encuentre su vecino más cercano $\boldsymbol{p_j}$; l
   b. calcular la distancia entre $\boldsymbol{p_i}$ y $\boldsymbol{p_j}$ y llámela $\boldsymbol{x_i = dist(p_i , p_j)}$
3. Generar un conjunto de datos simulados ($\boldsymbol{random_D}$) extraído de una distribución uniforme aleatoria con ***n*** puntos (q_1,…, q_n) y la misma variación que el conjunto de datos reales original **D**.
4. Calcular la distancia, $\boldsymbol{y_i}$ desde cada punto artificial hasta el punto de datos ***real*** más cercano.
   a. Para cada punto $\boldsymbol{q_i} \in \boldsymbol{random_D}$, encuentre su vecino más cercano $\boldsymbol{q_j}$ en **D**
   b. calcular la distancia entre $\boldsymbol{q_i}$ y $\boldsymbol{q_j}$ y llámela $\boldsymbol{y_i = dist(q_i , q_j)}$
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.

Estadística de Hopkins

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

¿Cómo interpretar la estadística de Hopkins?

Si D está distribuida de forma uniforme, entonces $\sum x_i$ y $\sum y_i$ serían muy parecidos, entonces H sería aproximadamente ½ (0.5).

Pero si hay clústeres en D, las distancias de los puntos artificiales $\sum y_i$ serían mucho más grandes que las distancias de los puntos reales: $\sum x_i$ y por lo tanto H sería mayor que 0.5.

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

Estadística de Hopkins: hipótesis

Hipótesis que maneja Hopkins:

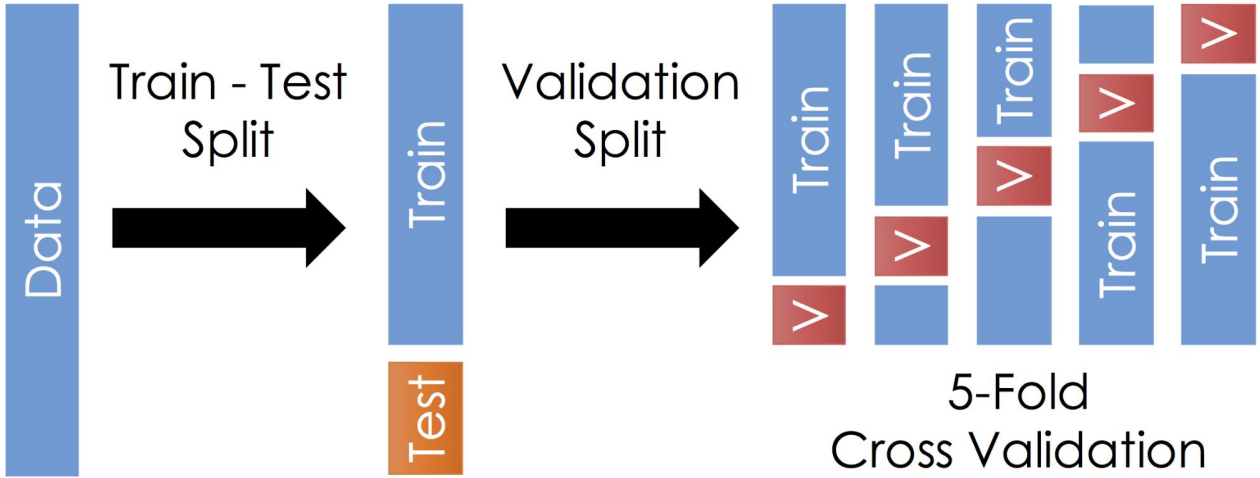
- **Hipótesis nula**: el conjunto de datos D se distribuye uniformemente (es decir, no hay clusters significativos)
- **Hipótesis alternativa**: el conjunto de datos D no está uniformemente distribuido (es decir, contiene clusters significativos)

Podemos realizar la prueba de la estadística de Hopkins de forma iterativa, utilizando 0,5 como umbral para rechazar la hipótesis alternativa.

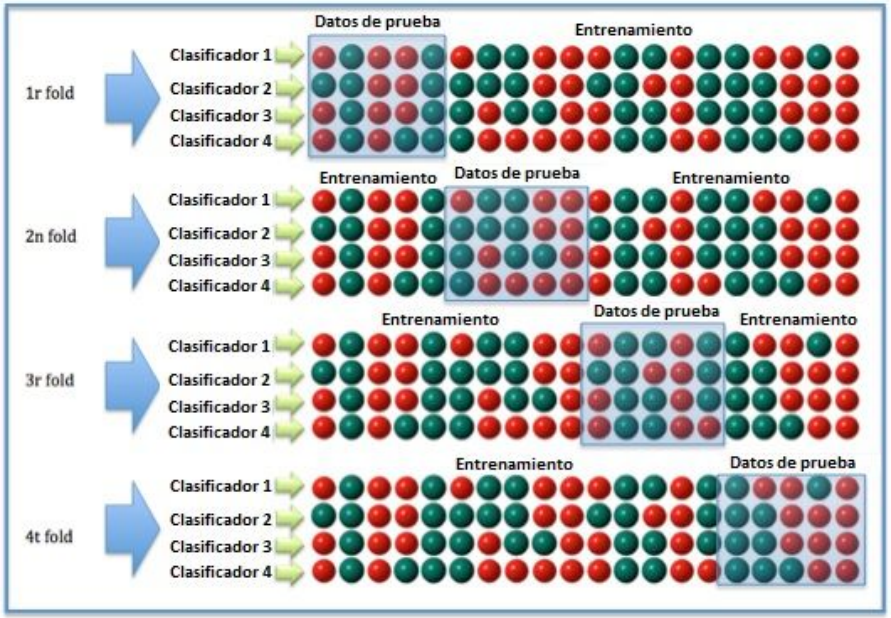
- Es decir, si **H < 0,5**, es poco probable que D tenga conglomerados estadísticamente significativos.
- O si el valor de la estadística de Hopkins es cercano a 1, entonces podemos rechazar la hipótesis nula y concluir que el conjunto de datos D es significativamente un dato agrupable.

Entrenamiento

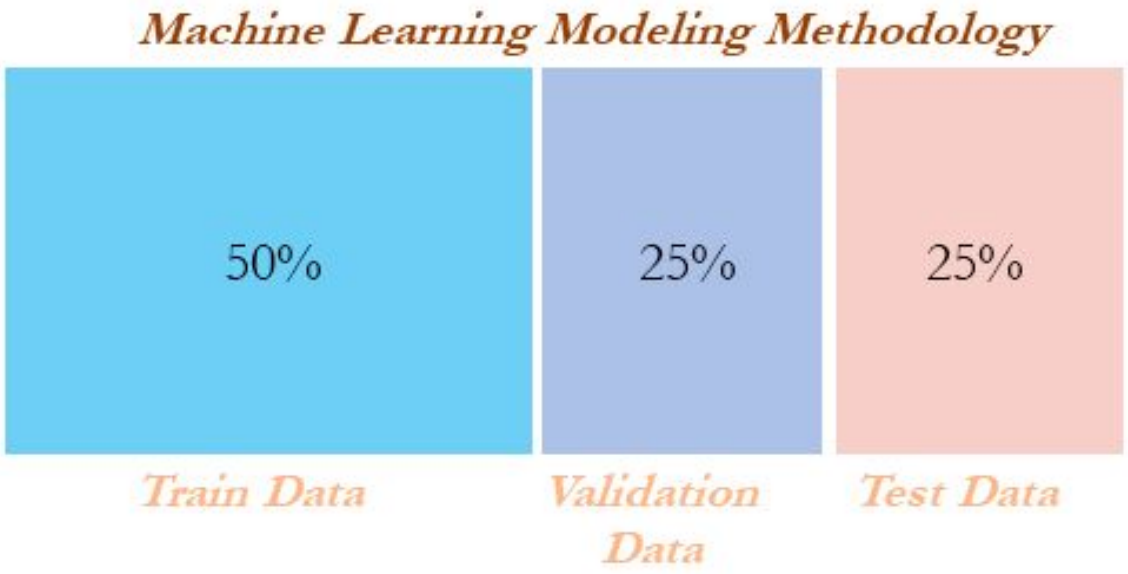
Conjuntos de entrenamiento y prueba



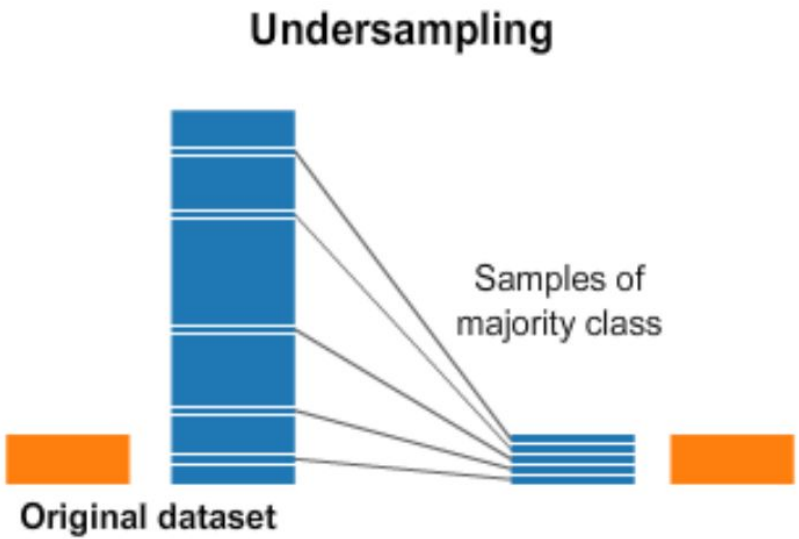
Conjuntos de entrenamiento y prueba

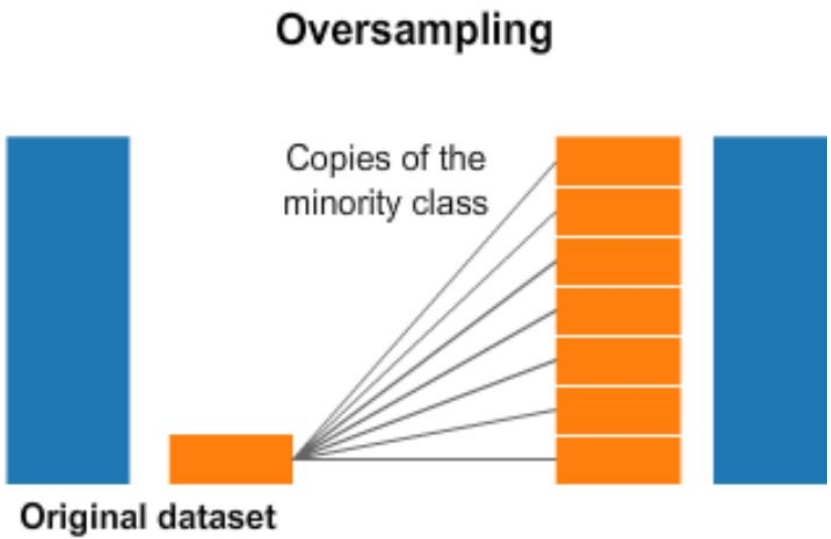


Conjuntos de entrenamiento y prueba

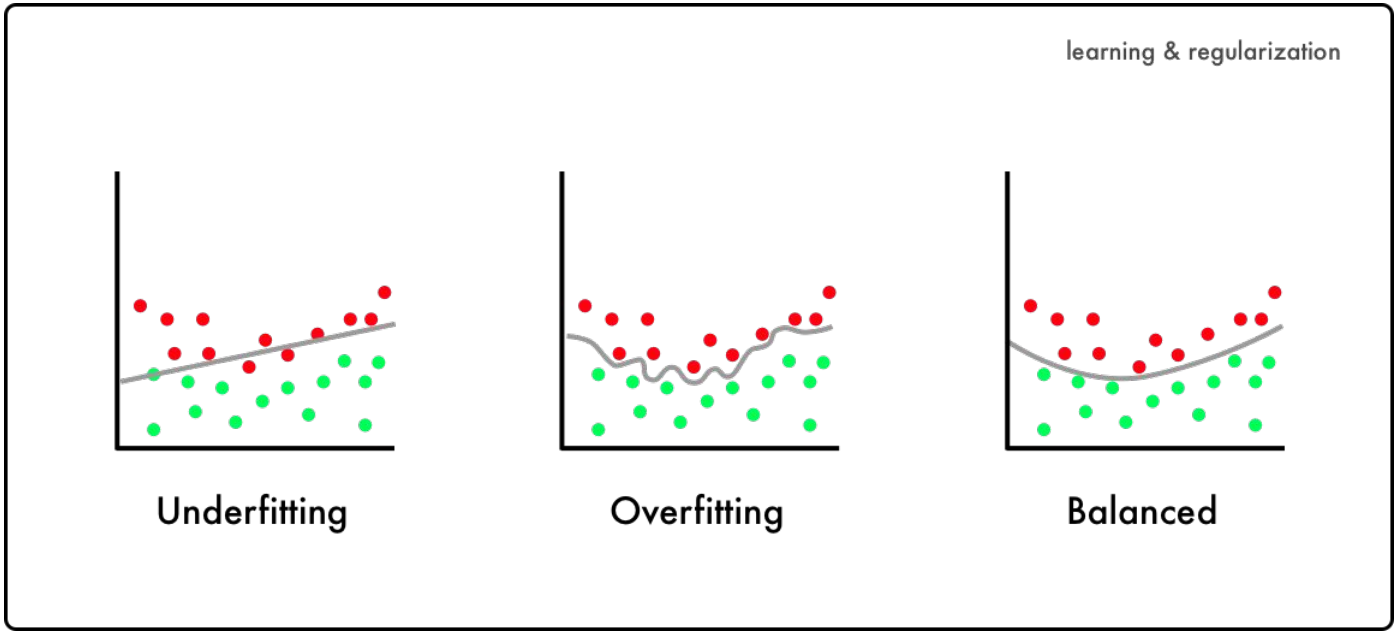


Conjuntos balanceados

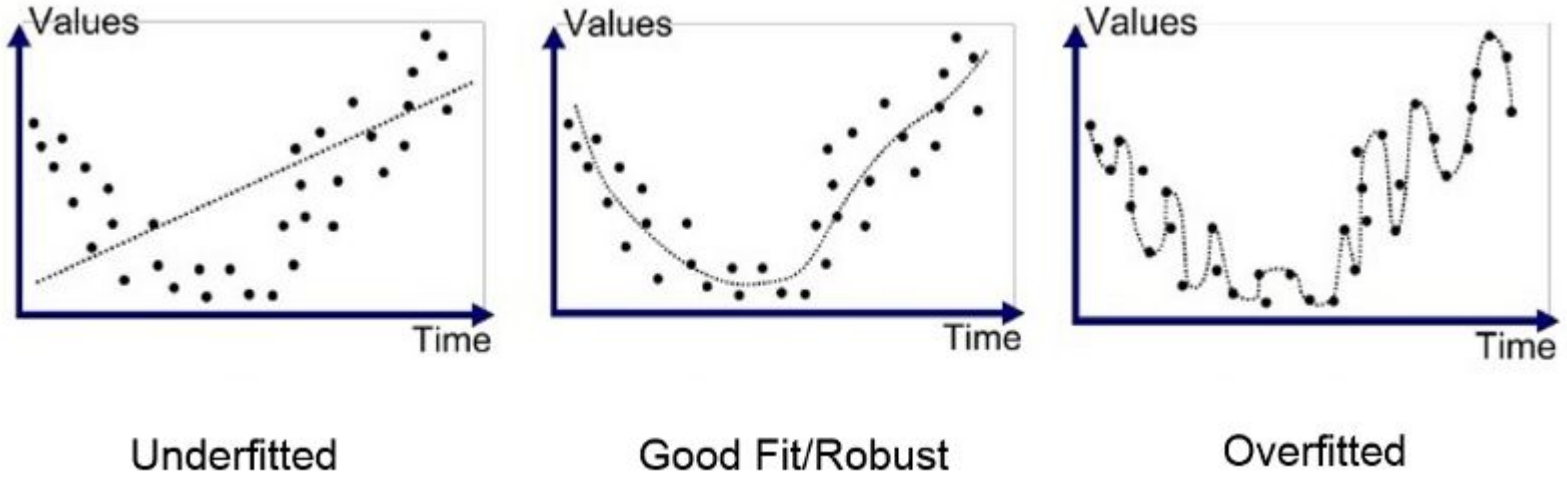




Overfitting



Overfitting



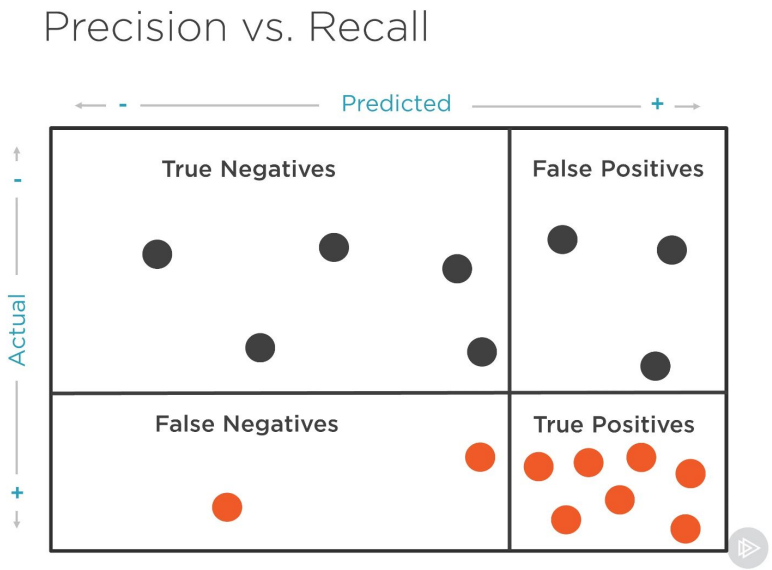
Métricas

Precision

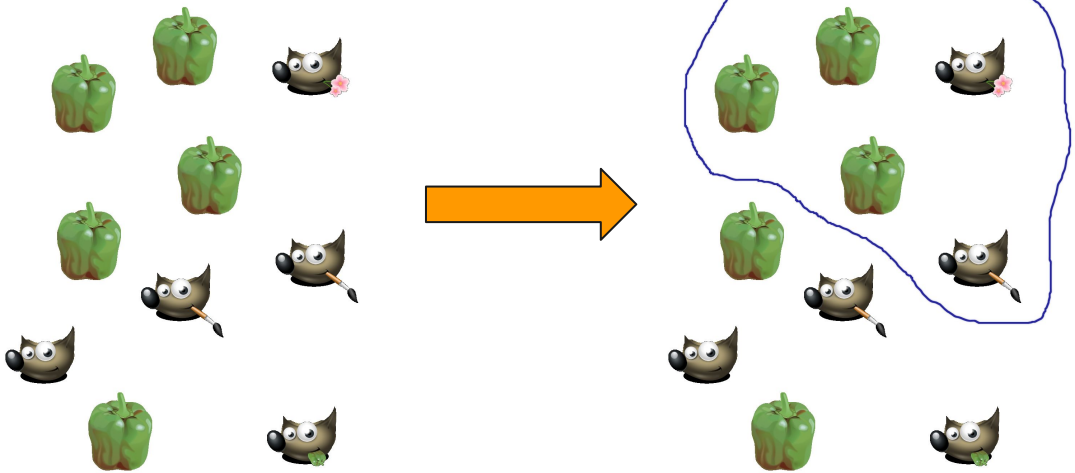
Precision

$$\frac{TP}{TP + FP} = \frac{7}{10}$$

.70



Precisión



Clasificación de la red en azul

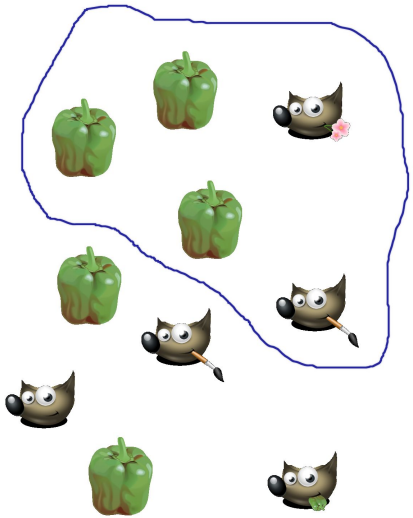
Verdaderos Positivos = 3
Falsos Positivos = 2

Falsos Negativos = 2
Verdaderos Negativos =2

Precisión = VP / VP + FP = 3 / 5

detección de morrones

Recall - exhaustividad



Verdaderos Positivos = 3
Falsos Positivos = 2

Falsos Negativos = 2
Verdaderos Negativos =3

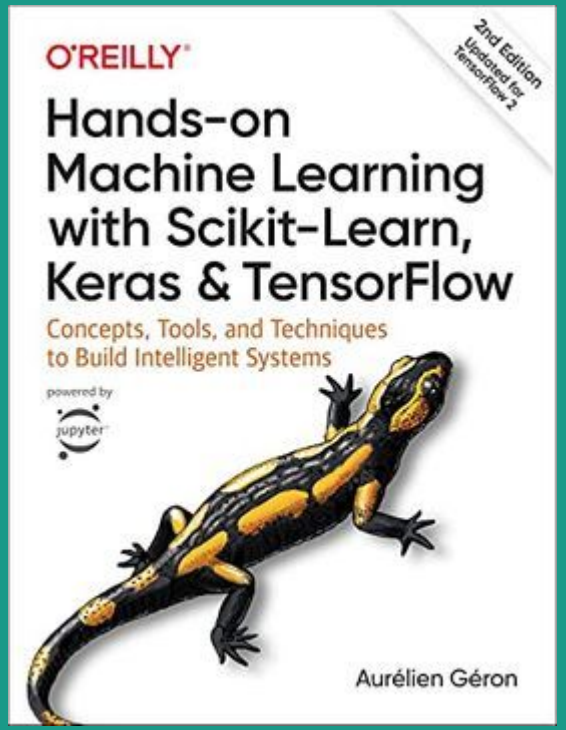
Recall = VP / VP + FN = 3 / 5

Formulas

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Bibliografía



Aurélien Géron

También en español