Aprendizaje Automático

Modelos de clasificación y clustering

Master Executive Big Data y Business Analytics

Edición 2015

Temario

- Sesión 1: Introducción a aprendizaje automático
- **Sesión 2: Modelos de clasificación y clustering**
- Sesión 3: Aprendizaje profundo, PCA y aplicaciones

Tabla de contenidos

Tabla de contenidos

Selección de variables

En los conjuntos de datos utilizados para entrenar los modelos

- Eliminar variables con poca capacidad predictiva:
  - Variables con baja varianza: constantes.
  - Identificadores: nombres o ID.
- Análisis univariante
- Eliminación recursiva de variables

Factor de inflación de la varianza (VIF)

El factor de inflación de la varianza (variance inflation factor, VIF) cuantifica la multicolinealidad en un análisis de regresión.

Proporciona un índice que mide cuánto se incrementa la varianza de un coeficiente de regresión estimado debido a la colinealidad.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Se considera que su valor es elevado cuando supera un valor de 5 a 10.

Stepwise

Los principales enfoques para la regresión Stepwise son:

- **Forward selection**, lo que implica comenzar sin variables en el modelo, poniendo a prueba la adición de cada variable utilizando un criterio de comparación modelo, añadiendo la variable que mejora el modelo, y repetir este proceso hasta que no mejora el modelo.

- **Backward elimination**, lo que implica comenzar con todas las variables candidatas, probando la eliminación de cada variable utilizando un criterio de comparación modelo, eliminando en cada paso la variable que mejora el modelo más al ser eliminado, y repetir este proceso hasta que no mejora adicional es posible.

- **Bidirectional elimination**, una combinación de lo anterior, las pruebas en cada paso implica la eliminación y adicción de variables.

Tabla de contenidos

Modelos de clasificación

En los **problemas de clasificación** el objetivo es encontrar una función que **asigna** cada uno de los registros **a una clase o etiqueta** asociada.

En los algoritmos utilizados para la búsqueda de la función de clasificación se ha de utilizar un conjunto de vectores de características y las correspondientes etiquetas para estimar los valores que producen el mejor clasificador.

Definiciones

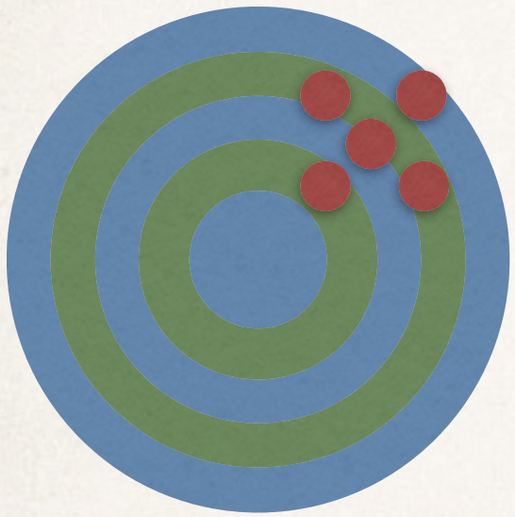
En los problemas de clasificación de clase binarios, en la que los resultados se etiquetan positivos (P) o negativos (N), hay cuatro posibles resultados que se pueden obtener:

- Verdaderos Positivos (TP)
- Verdaderos Negativos (TN)
- Falsos Positivos (FP) o Error tipo I
- Falsos Negativos (FN) o Error de tipo II

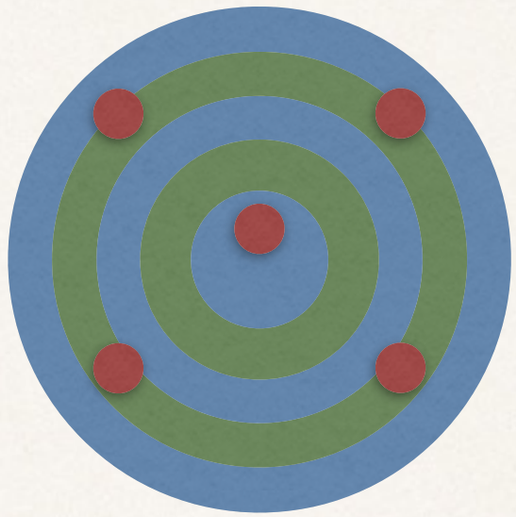
Con estos valores se pueden crear la matriz de confusión (o tabla de contingencia):

$$\begin{vmatrix} TP & FP \\ FN & TN \end{vmatrix}$$

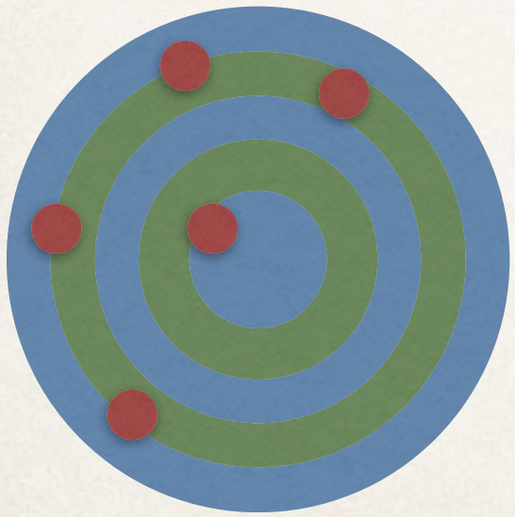
Clasificación: Sesgo y varianza



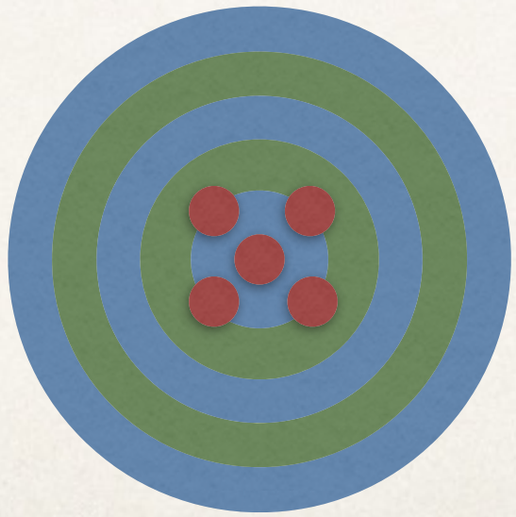
Gran sesgo
Baja varianza



Bajo sesgo
Alta varianza



Gran sesgo
Gran varianza



Bajo sesgo
Baja varianza

Métricas de rendimiento

- Precisión (Accuracy):

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

- Exactitud (Precision)

$$P = \frac{TP}{TP + FP}$$

- Exhaustividad (Recall):

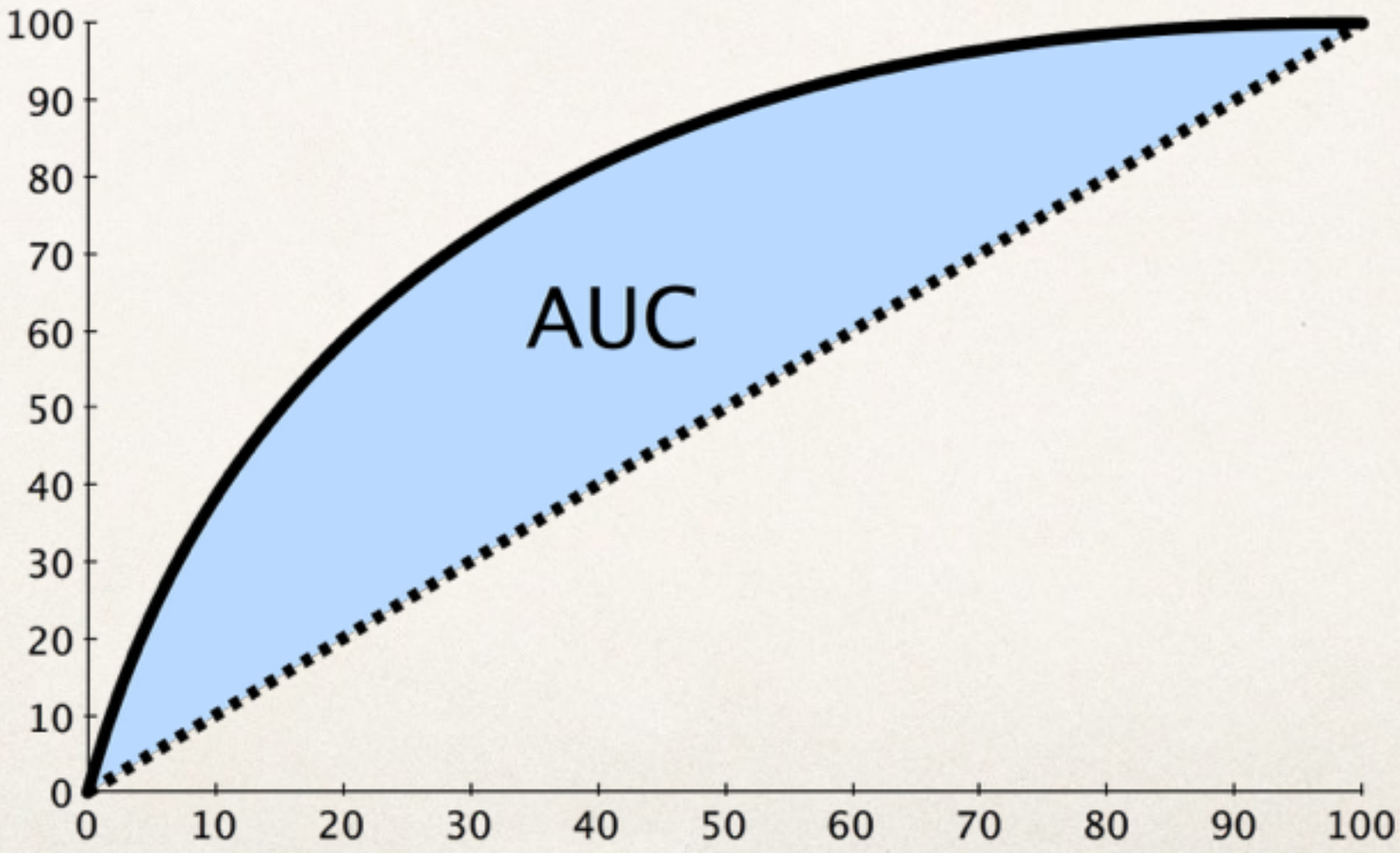
$$R = \frac{TP}{TP + FN}$$

- F1:

$$F1 = \frac{PR}{P + R}$$

Área bajo la curva ROC (ROC AUC)

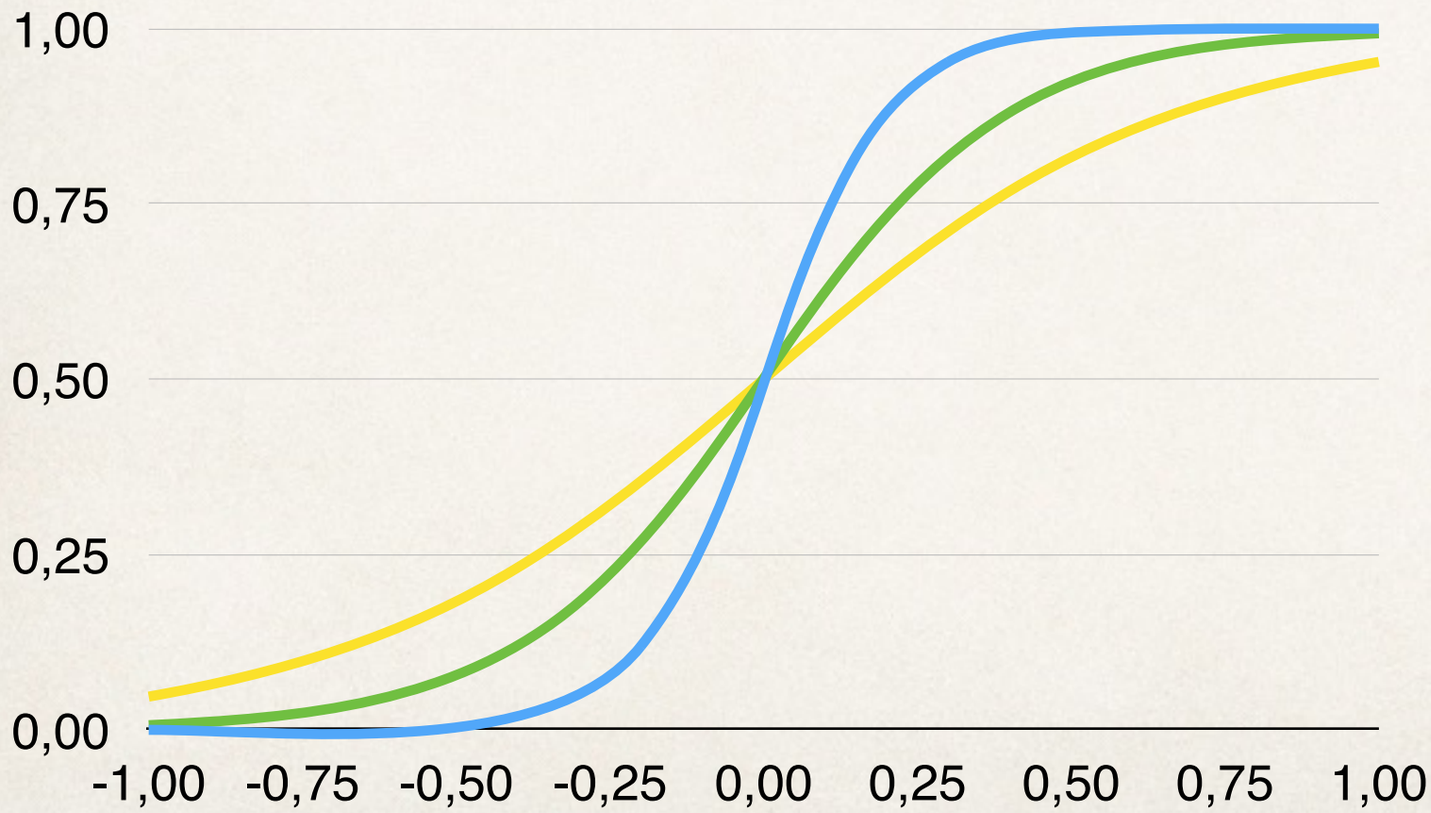
El Área Bajo la Curva permite medir la precisión de los modelo de clasificación. En el eje x se representa el Ratio de Falsos Positivos y el el eje y el Ratio de Verdaderos Positivos.



Regresión logística

La clasificación binaria de eventos se puede realizar a partir de un modelo regresión logística donde se utiliza la expresión:

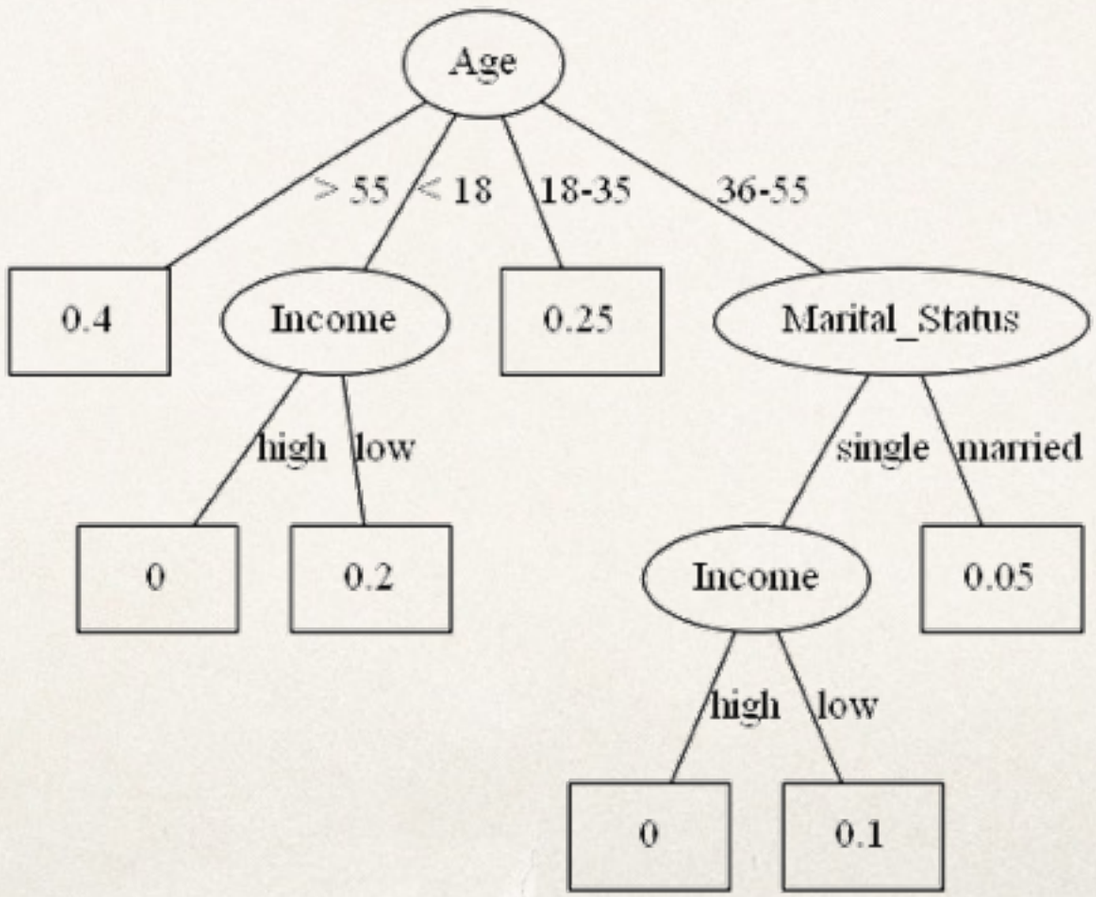
$$F(x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{\sum_i \beta_i x_i}}$$



Arboles de decisión

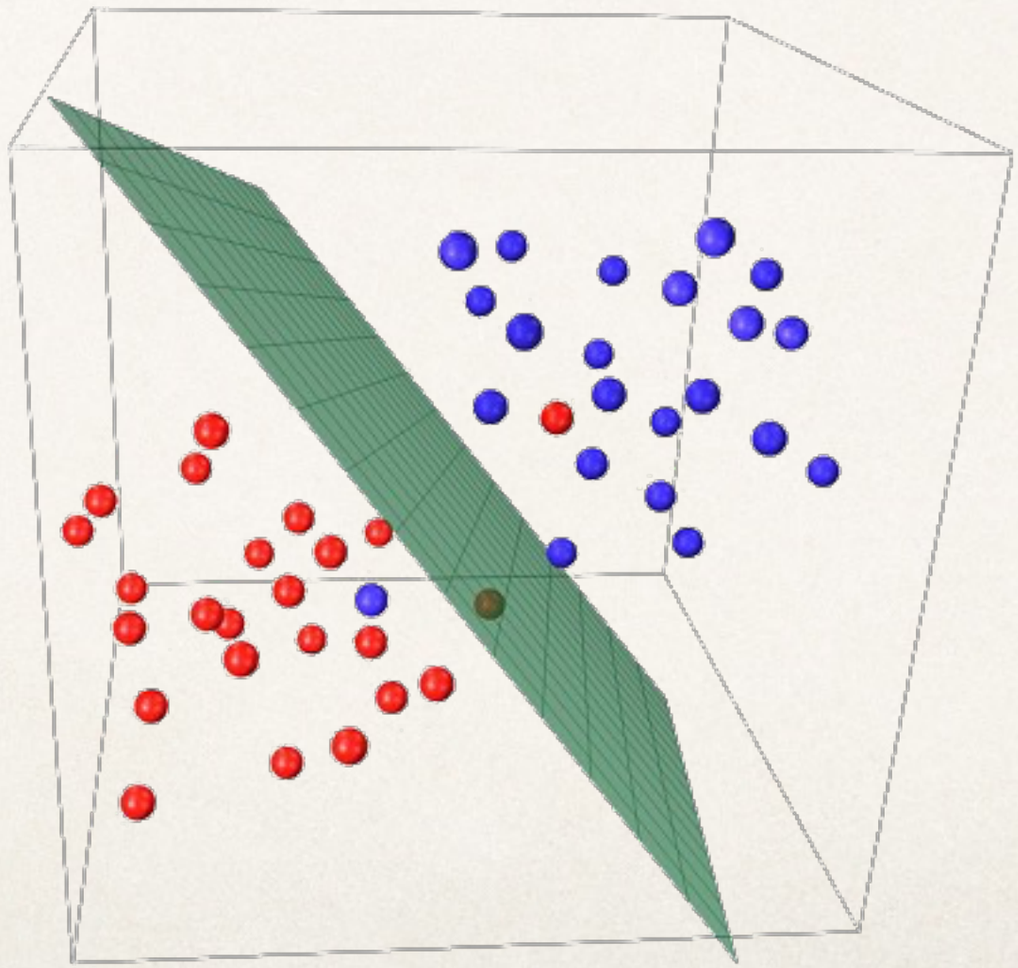
Los arboles de decisión crean construcciones similares a los sistemas de decisión basados en reglas.

Random forest: son una combinación de árboles de decision



Máquinas de vectores de soporte

Las Máquinas de vectores de soporte (SVM, Support vector machines) busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra.



Naïve Bayes

Un Clasificador Bayesiano Ingenuo (Naive Bayes) es un clasificador probabilístico que se basa en el teorema de Bayes y alguno hipótesis simplificadoras adicionales.

Teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naïve de Bayes:

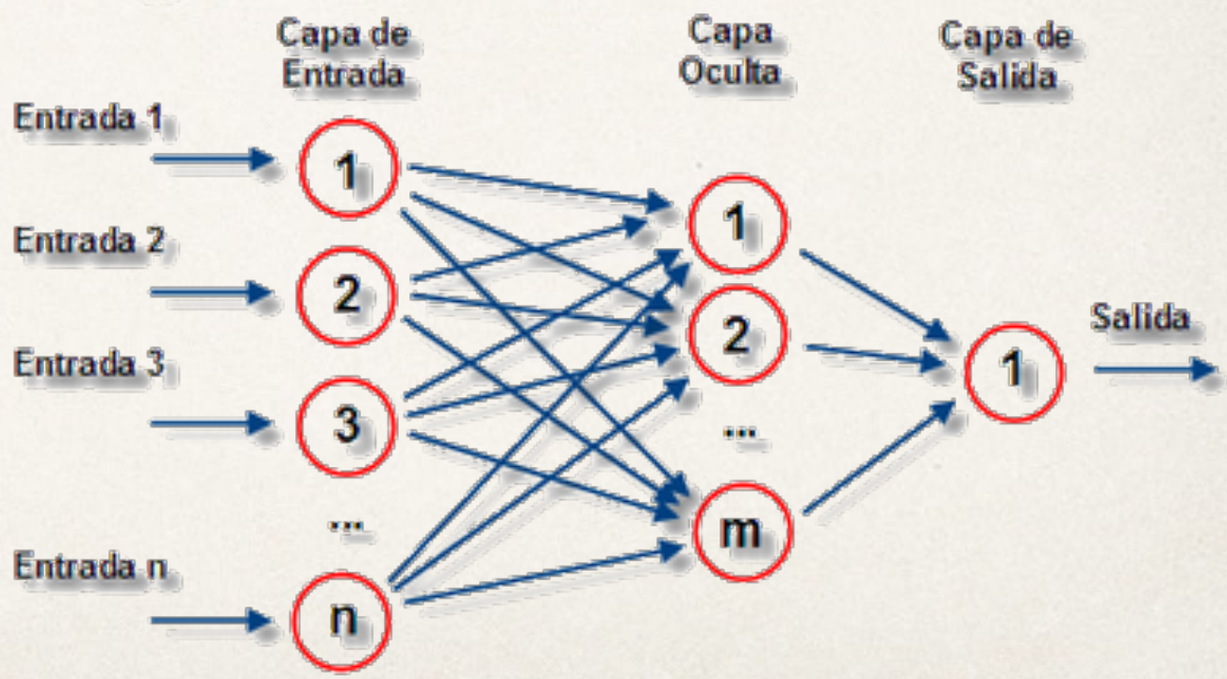
$$P(A|X_1, X_2, \ldots, X_n) = \frac{1}{Z} P(A) \prod_n P(F_i|A)$$

Redes Neuronales

Las redes neuronales es un paradigma de aprendizaje automático supervisado inspirado en el funcionamiento del sistema nervioso.

Las redes neuronales están basadas neurona, cada una de las cuales recibe una serien de entradas y emite una salida.

Las neuronas se agrupan en capas, pudiendo se las redes monocapa o multicapa. En las redes multicapa la capa donde recibe la entrada se llama capa de entrada, la salida se produce en la de salida y las intermedias se llama capas ocultas.



Selección de variables con Information Value (IV)

En los modelos de clasificación se puede utilizar el Information Value (IV) para analizar la capacidad de clasificación de cada variable:

$$IV = \sum_{i=1}^{N} (R_i(T) - R_i(F)) \ln \left| \frac{R_i(T)}{R_i(F)} \right|$$

$R_i(T)$ Porcentaje de T en la categoria i

Valor de IV	Capacidad de clasificación
< 0,02	muy débil
0,02 a 0,1	débil
0,1 a 0,3	promedio
0,3 a 0,5	fuerte
> 0,5	muy fuerte

Tabla de contenidos

Clustering

En el análisis de clustering se busca **agrupar** las observaciones de un conjunto de datos de tal manera que **los miembros de un mismo cluster son más similares entre sí de lo que son los miembros de los otros grupos**.

La similitud entre dos registros de datos se calcula utilizando una métrica.

Métricas (I)

- Euclídea

$$\sqrt{\sum_i (u_i - v_i)^2}$$

- Euclídea Normalizada

$$\sqrt{\sum (u_i - v_i)^2 / V[x_i]}$$

- Minkowski

$$\left(\sum |u_i - v_i|^p\right)^{1/p}$$

- Coseno

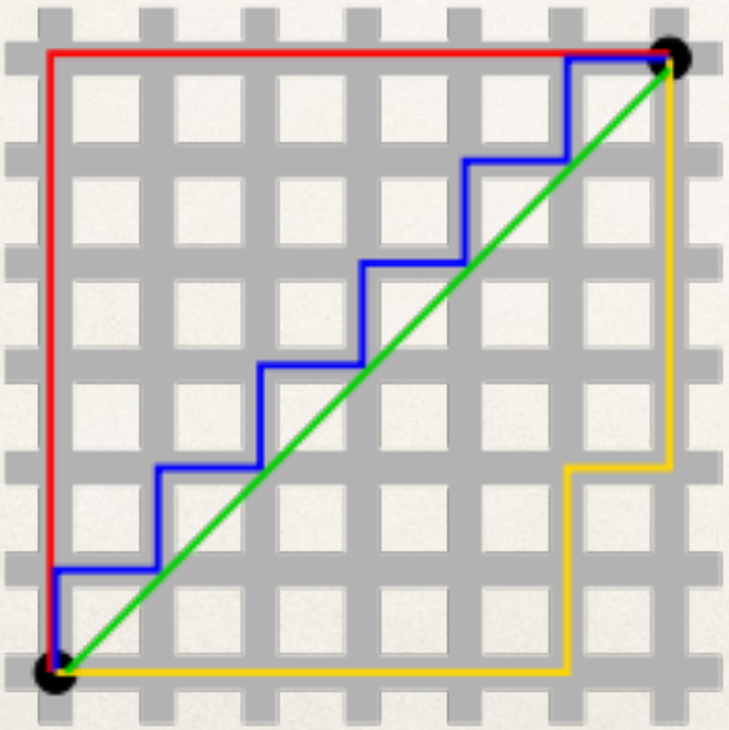
$$1 - \frac{u \cdot v}{|u||v|}$$

Métricas (y II)

- Correlación

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{|(u - \bar{u})|\,|(v - \bar{v})|}$$

- Geometría del taxista (City block o Manhattan)

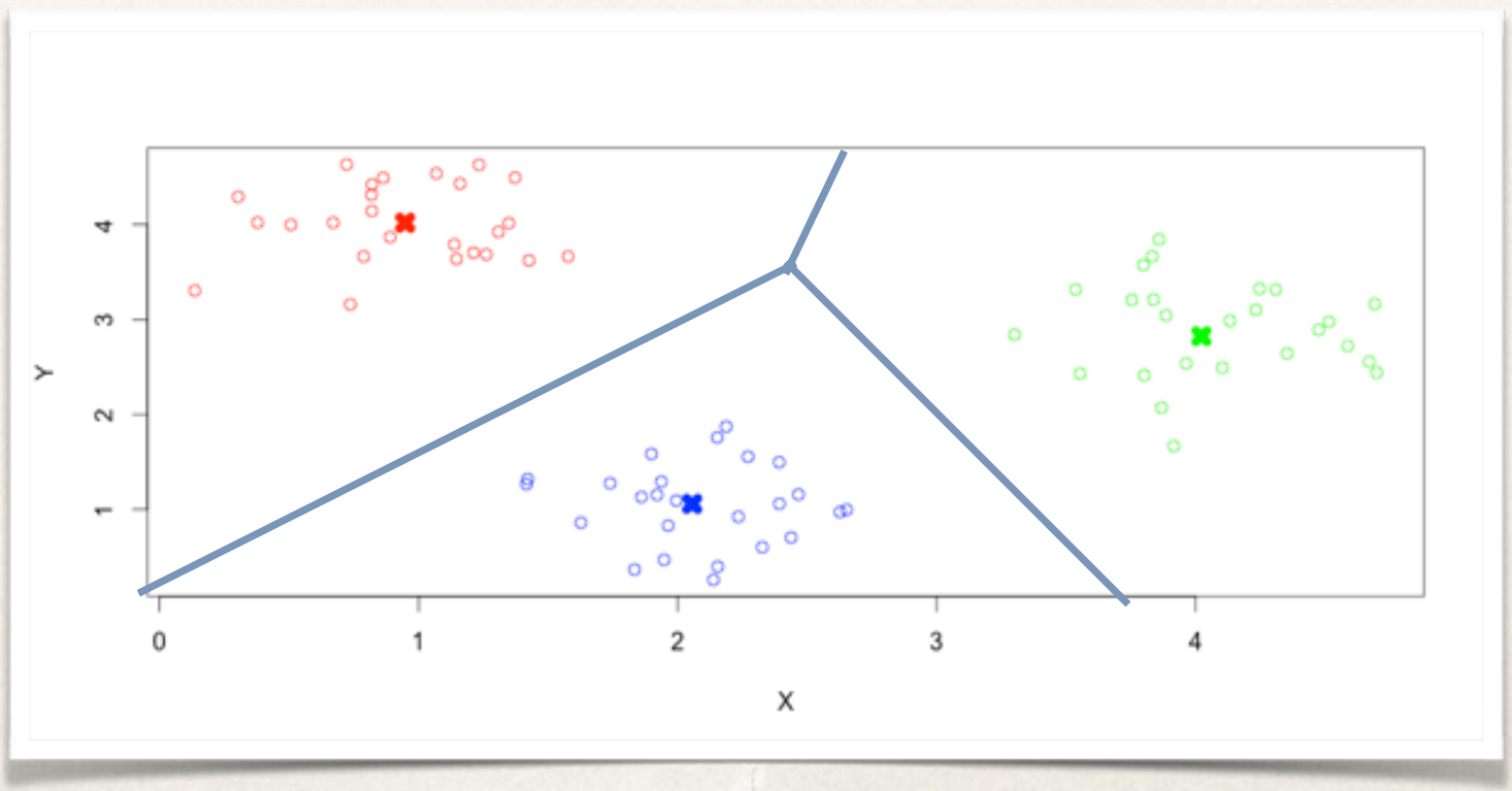


$$\sum_{i=1}^{n} |u_i - v_i|$$

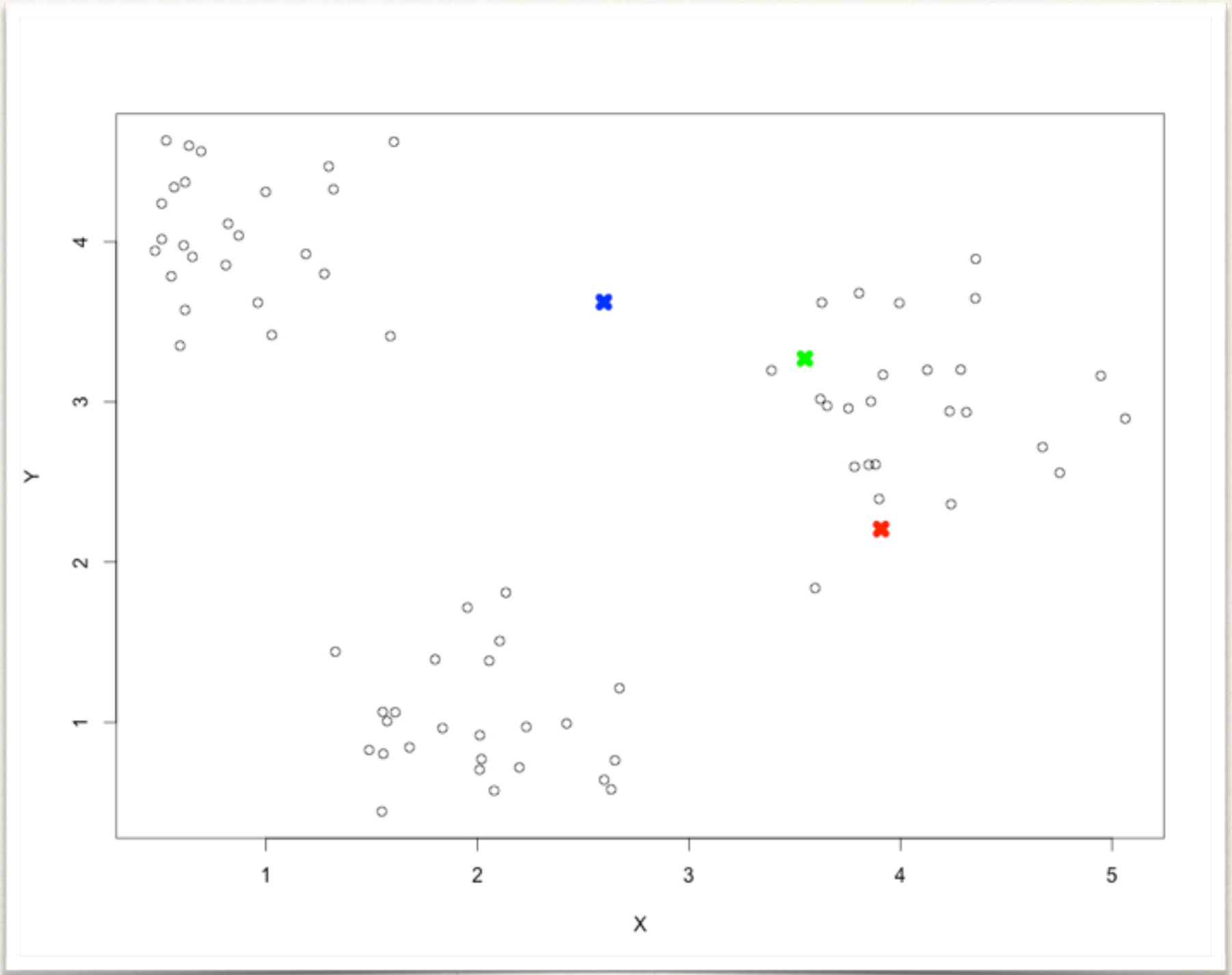
k-means

- k-means es uno de los métodos más usados para el análisis de clustering.

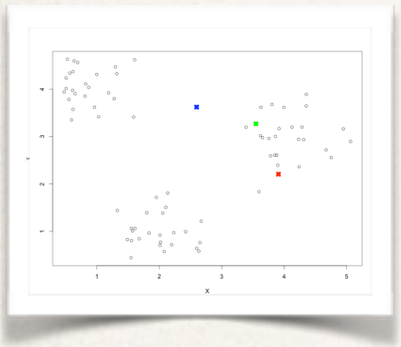
- Para funcionar el algoritmo requiere el número de clusters en los que se ha de separar los datos.

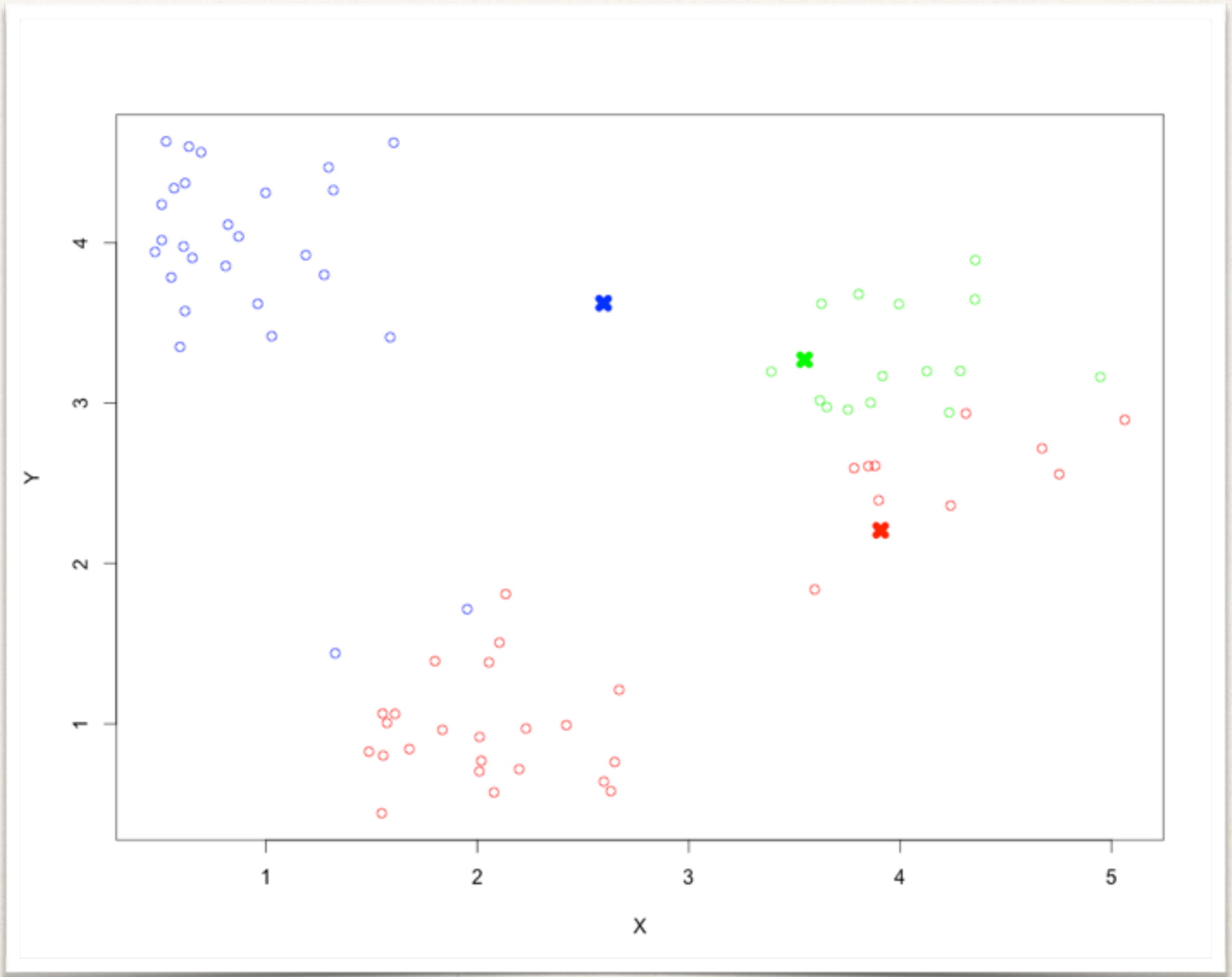


k-means funcionamiento

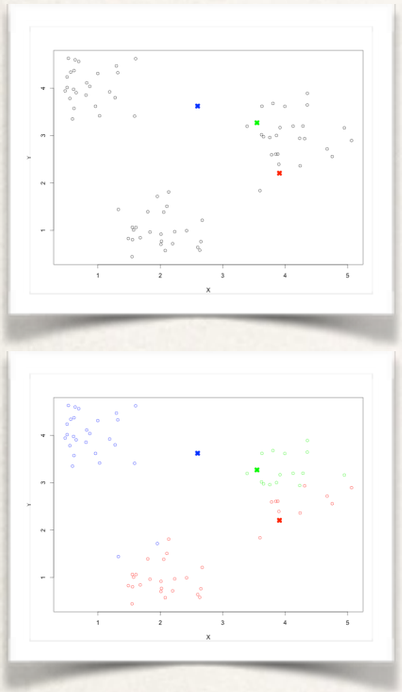


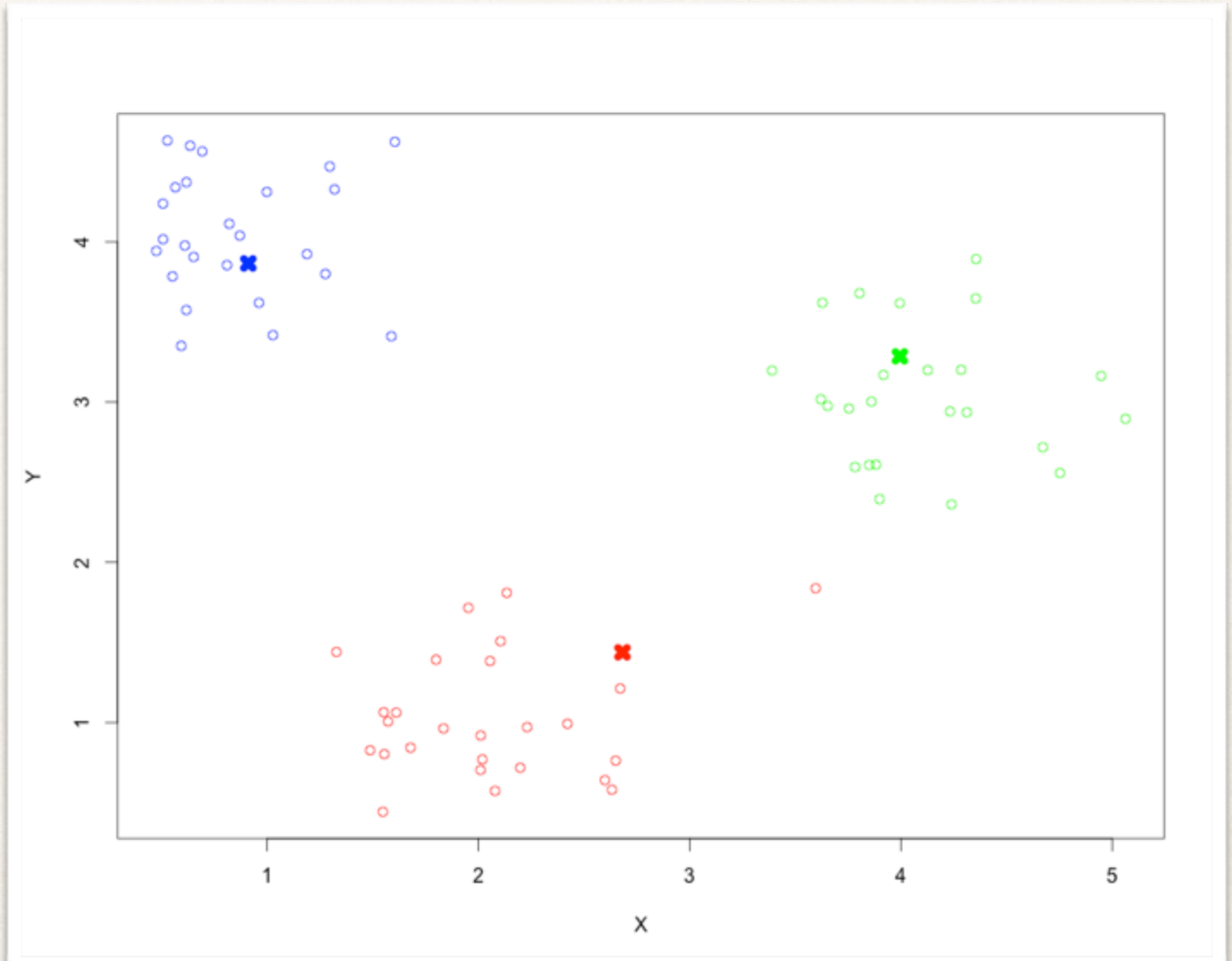
k-means funcionamiento



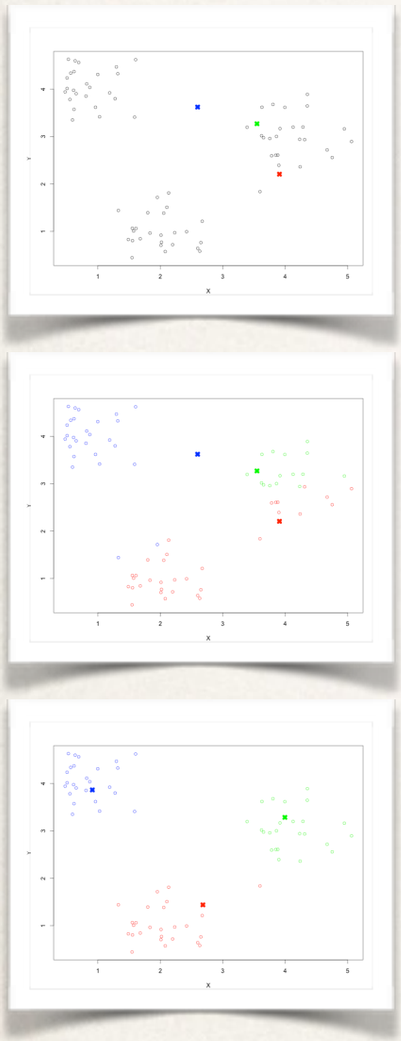


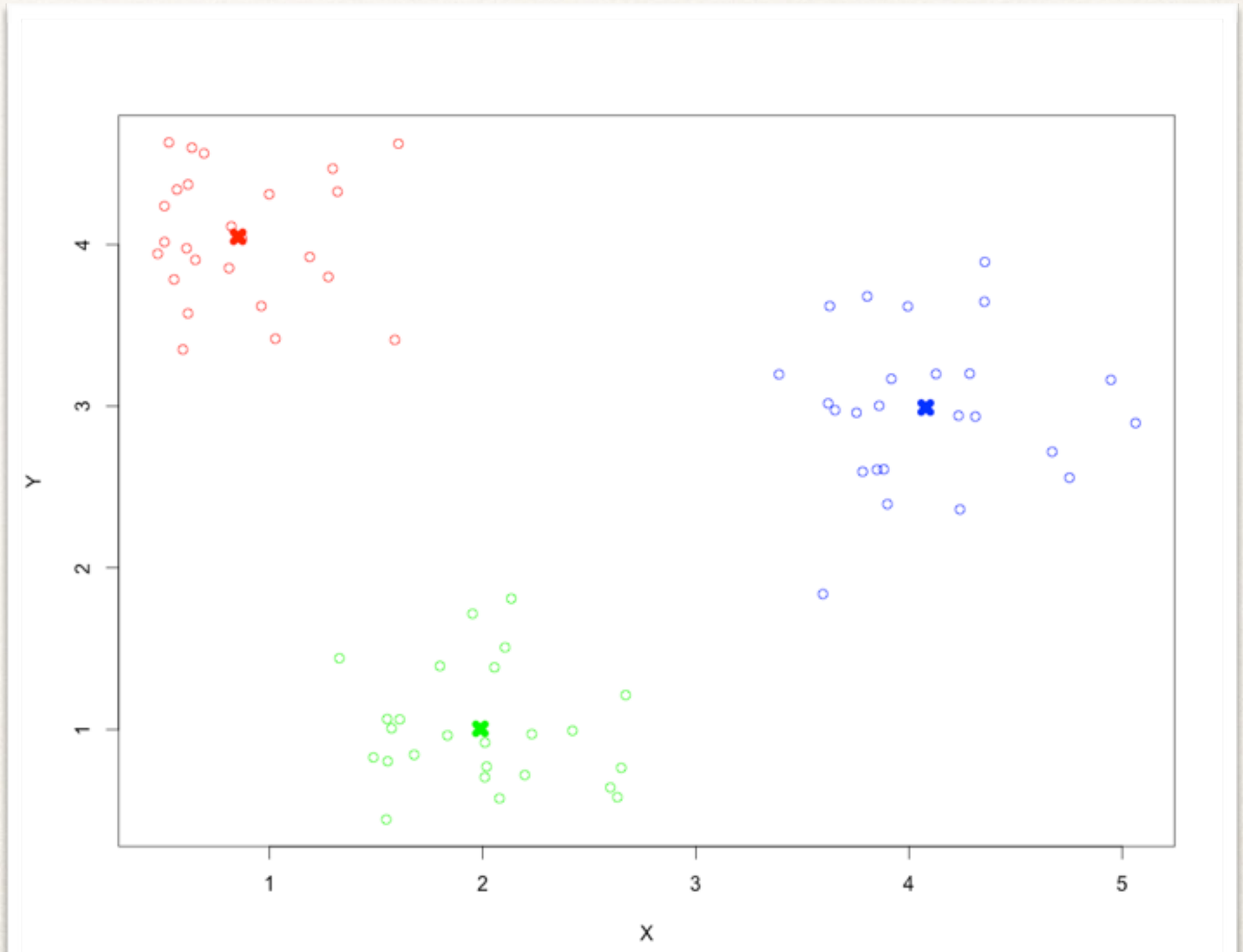
k-means funcionamiento





k-means funcionamiento





k-means en scikit-learn

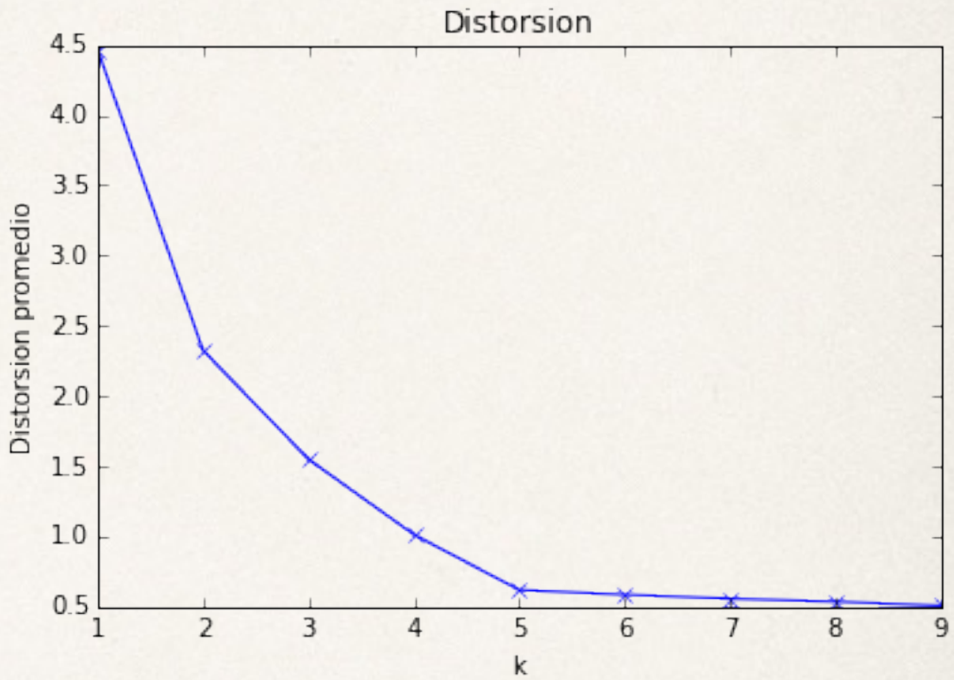
- En scikit-learn existen dos implementaciones
  - KMeans: es la implementación estándar
  - MiniBatchKMeans: utiliza subconjunto de datos para reducir el tiempo de computación, siendo adecuado para grandes conjuntos de datos.

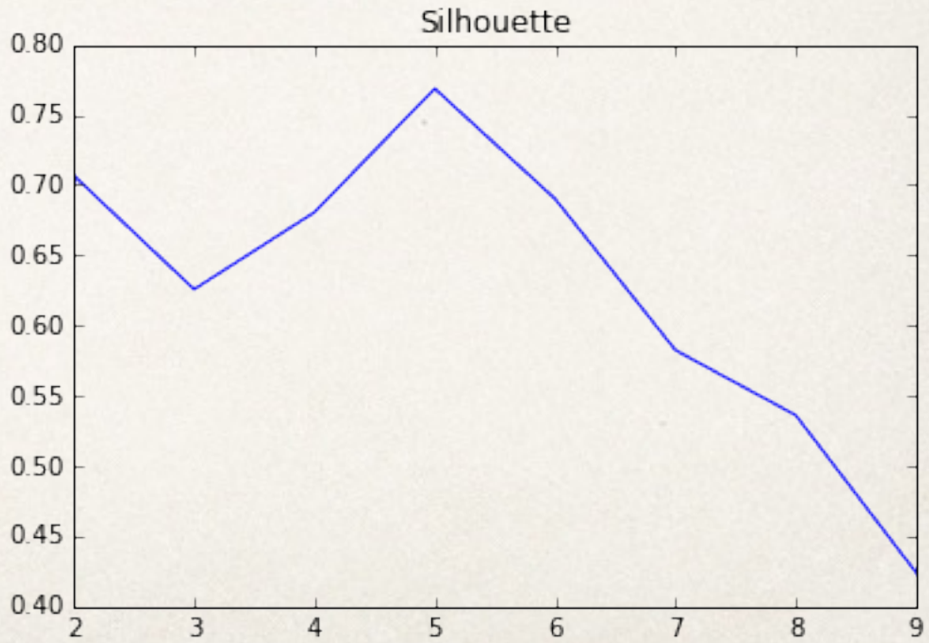
Selección del número de clusters (I)

El número de clusters es un parámetro de entrada en varios algoritmos.

Algunos de los métodos más utilizados para determinar el valor son:

- Evaluación de la distorsión.
- Silhouette.





Selección del número de clusters (y II)

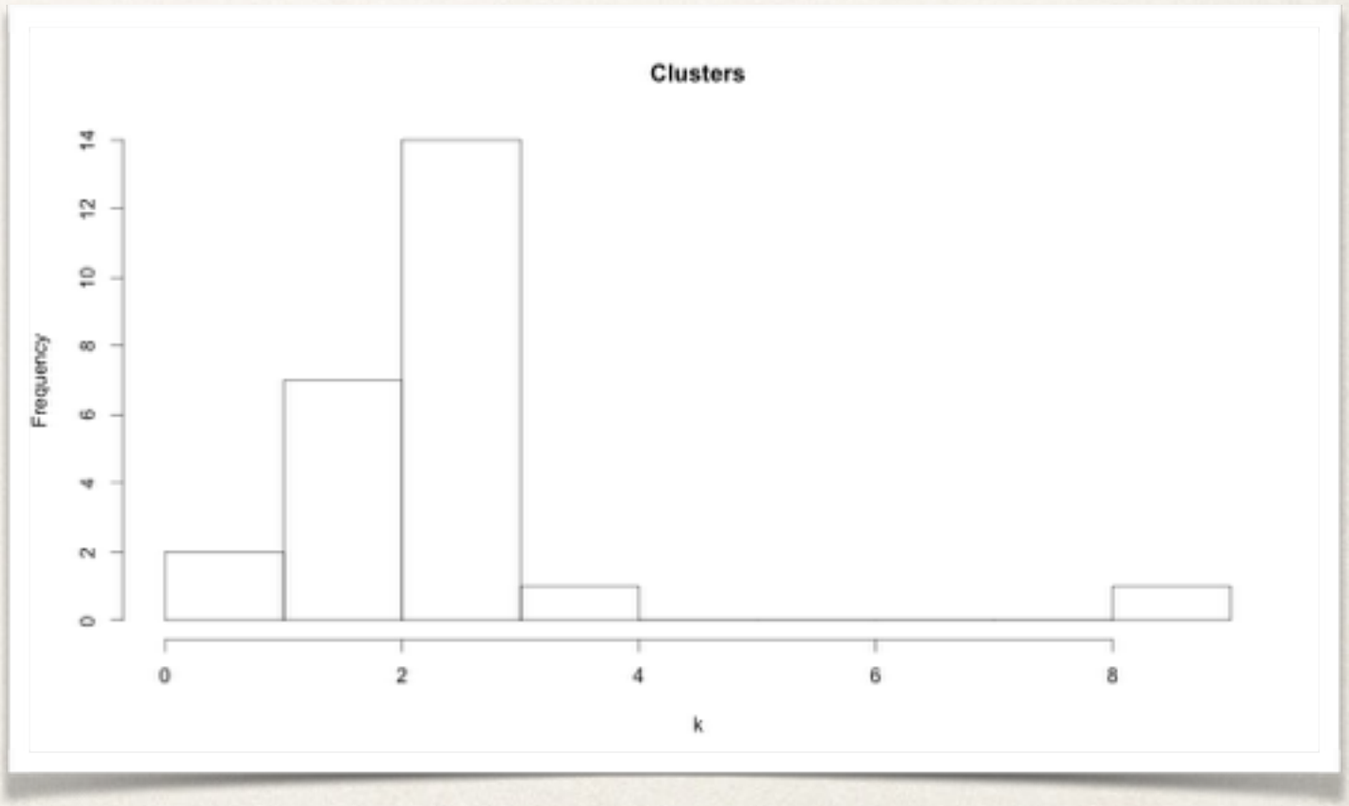
Existe una amplia variedad de métodos, en R existe un paquete `NbCluts` implementa 23:

```
library(NbClust)


x <- matrix(c(rnorm(100, 2, .1), rnorm(100, 3, .1),
              rnorm(100, -2, .1), rnorm(100, 1, .1),
              rnorm(100, 1, .1), rnorm(100, -3, .1)),
            300, 2)

res <- NbClust(x, min.nc = 2, max.nc = 10,
               method = ‘kmeans’)

hist(res$Best.nc[1, ],
     breaks = 8,
     main = ‘Clusters’,
     xlab = ‘k’)
```



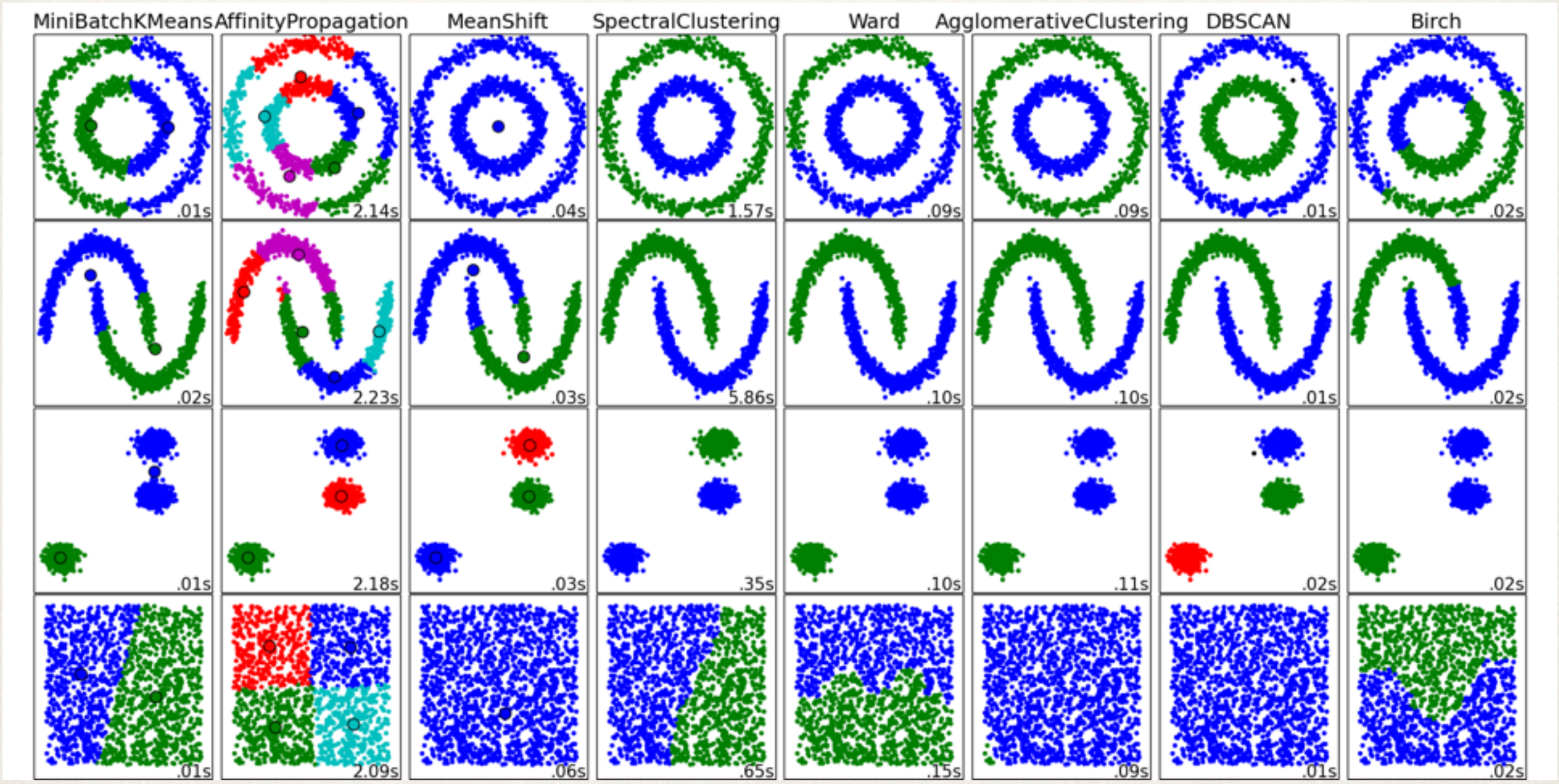
Otros algoritmos de clustering (I)

- Propagación de afinidades ("Affinity Propagation”)
  - En el algoritmo "Affinity Propagation" cada uno de los objetos a agrupar se considera un nodo de una red y todos ellos simultáneamente son considerados potenciales ejemplares. El funcionamiento general del algoritmo consiste en una transmisión de mensajes de forma recursiva con valores reales entre las aristas de la red, es decir, entre todos los pares de objetos hasta que los valores de los mensajes converjan.
- Mean Shift
  - El algoritmo de “Mean Shift” esta basado en la búsqueda de los centros de datos, a diferencia de k-means no necesita saber cuántos son. La búsqueda de los centros la realiza definiendo una región y calculando la posición en función de la media, repitiendo el proceso hasta que converge.
- Hierarchical clustering
  - Hierarchical clustering es una familia genérica de algoritmos de clustering que construyen grupos anidados mediante la fusión o división sucesiva. Existen tres tipos de vincular:
    - Ward: minimiza la suma de las diferencias al cuadrado dentro de todos los grupos.
    - Average: minimiza la media de las distancias entre todos los pares de observaciones de clusters.
    - Complete: minimiza la distancia máxima entre las observaciones de pares de grupos.

Otros algoritmos de clustering (y II)

- DBSCAN
  - El algoritmo DBSCAN ve agrupaciones como áreas de alta densidad separadas por zonas de baja densidad.
- Gaussian mixtures
  - Un modelo de mezcla gaussiana es un modelo probabilístico que asume todos los puntos de datos se generan a partir de una mezcla de un número finito de distribuciones gaussianas con parámetros desconocidos.
- Spectral Clustering
  - Son métodos de agrupamiento basados en la descomposición espectral de una matriz de afinidad
- BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)
  - Se basa en una estructura de datos jerárquica llamada CF Tree (Clustering Feature Tree), estos son árbol balanceado en el que los nodos internos almacenan las sumas de los CFs de sus descendientes
    - Fase 1: Se construye un árbol CF inicial
    - Fase 2: Se utiliza un algoritmo de clustering arbitrario para agrupar los nodos hoja del árbol CF

Comparación de lo algoritmos de clustering



Fuente: http://scikit-learn.org/stable/modules/clustering.html