Machine Learning

Especialización en Ciencia de Datos.
Aplicaciones del aprendizaje profundo en el análisis de grandes volúmenes de datos.

Agenda

- ¿Qué es Machine Learning?
- Clasificación ML.
- K-Means
- Mixtura Gaussiana

¿Qué es Machine Learning?

Machine Learning consiste en extraer conocimiento desde los datos. Podríamos pensar en aprender a partir de los datos.

Es un campo que involucra estadística, conceptos de inteligencia artificial (viene de) en el contexto de ciencias de la computación.

El término Machine Learning es conocido como a*nálisis predictivo* o *aprendizaje estadístico*.

¿Qué es Machine Learning?

También, podemos decir que Machine Learning es el conjunto de métodos y técnicas que permiten la detección automática de patrones en un conjunto de datos.

Como fin, se busca encontrar patrones ocultos que permitan predecir o tomar decisiones a futuro.

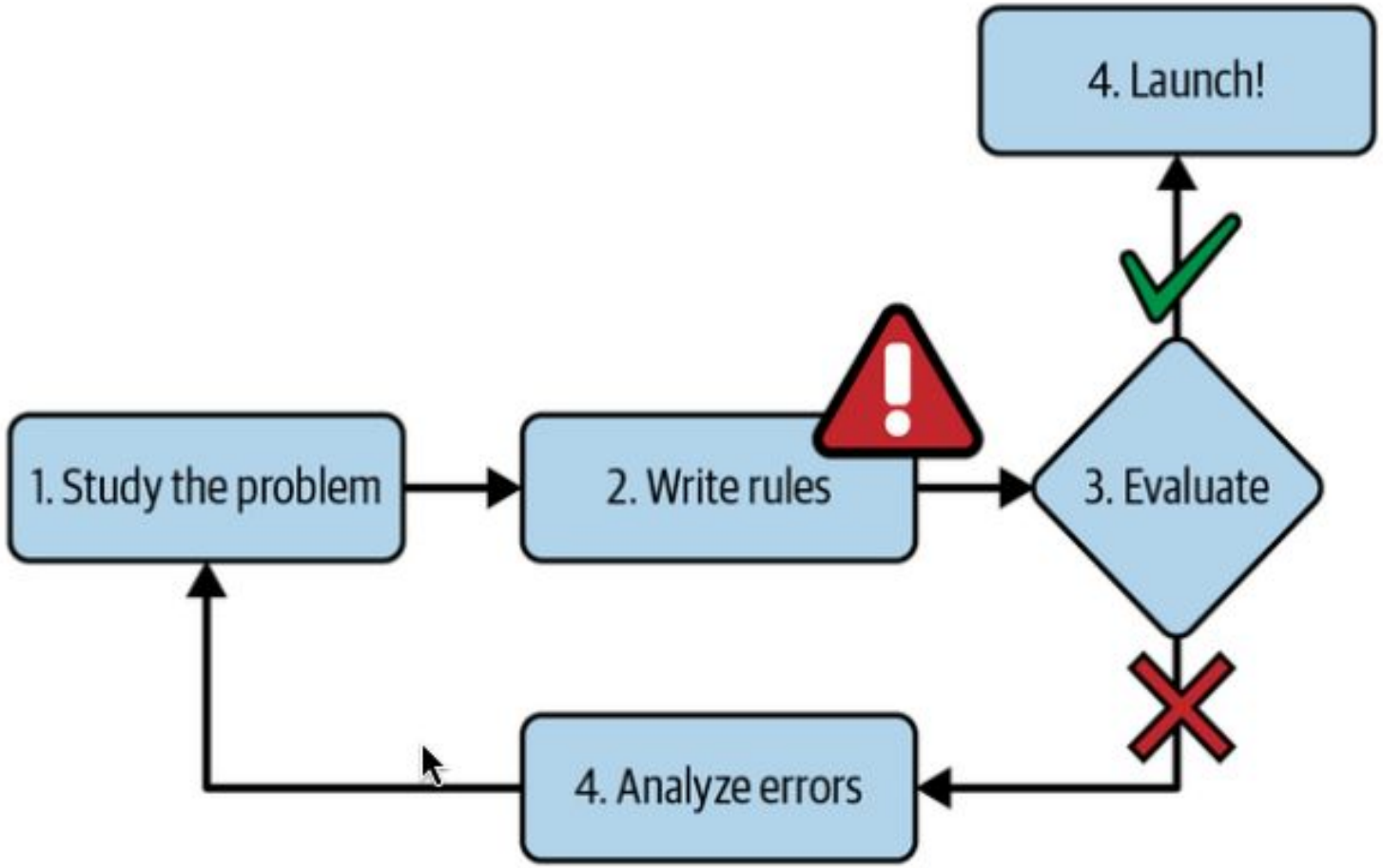
¿Qué es Machine Learning?

Ejemplo: Tenemos que diseñar un sistema que filtre correos spam del estilo, “Tenemos una gran oferta”, “Caja de Ahora en peligro”, “Ganó 1 millon de dolares”.

Un enfoque podría implicar crear una lista de Asuntos con frases a filtrar. Es decir, definir reglas que indiquen qué correo es spam y cuál no a partir del campo Asunto.

Problema: Este enfoque implica que la lista de filtros se debe actualizar constantemente (manualmente) y además se deben contemplar todas las combinaciones posibles. Por ejemplo: “Tenemos una gran Oferta”, “Tenemos una oferta grandiosa” , etc.

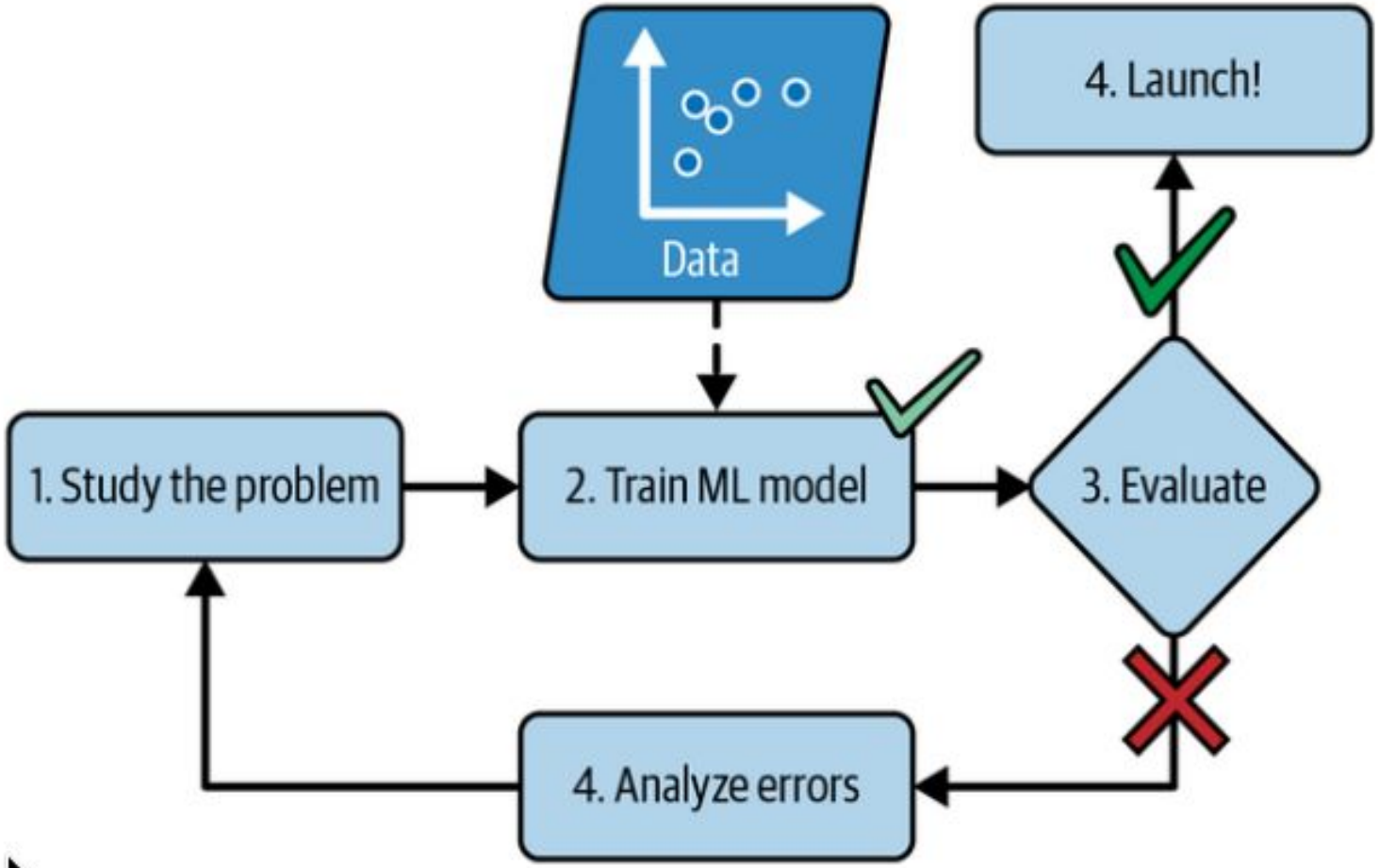
¿Qué es Machine Learning?



¿Qué es Machine Learning?

En contraparte, un filtro de spam basado en técnicas de aprendizaje automático *aprende automáticamente* qué palabras y frases son buenos predictores de spam, al *detectar patrones* de palabras inusualmente frecuentes en los ejemplos de spam.

¿Qué es Machine Learning?



Clasificación

Los sistemas de ML se pueden clasificar de acuerdo al nivel de "supervisión" durante la fase de entrenamiento del modelo.

En este sentido, existen diferentes categorías pero hay dos que son claves: el aprendizaje supervisado y el no supervisado.

Clasificación - Aprendizaje Supervisado

En el aprendizaje supervisado, el conjunto de entrenamiento que se introduce en el algoritmo incluye las soluciones deseadas, llamadas etiquetas (labels).

Por ejemplo, podemos tener un conjunto de datos de entrenamiento con asuntos de mails de la siguiente forma:

Asunto	Spam
Gran Oferta	Si
CV	No
Proyecto Investigación	No
Ganó 1 millon	Si

Clasificación - Aprendizaje No Supervisado

En Aprendizaje No Supervisado el conjunto de datos de entrenamiento no está etiquetado.

Por ejemplo, se tiene un conjunto de datos referido a visitantes de un blog. Con aprendizaje no supervisado podríamos detectar, con un algoritmo de clasificación, los grupos de usuarios que son similares.

Clasificación - Ejemplo Supervisado

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3	5.9	2.1	Iris-virginica

Clasificación - Ejemplo No Supervisado

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
7	3.2	4.7	1.4
5.8	2.7	5.1	1.9
7.1	3	5.9	2.1
6.3	2.9	5.6	1.8

K-Means

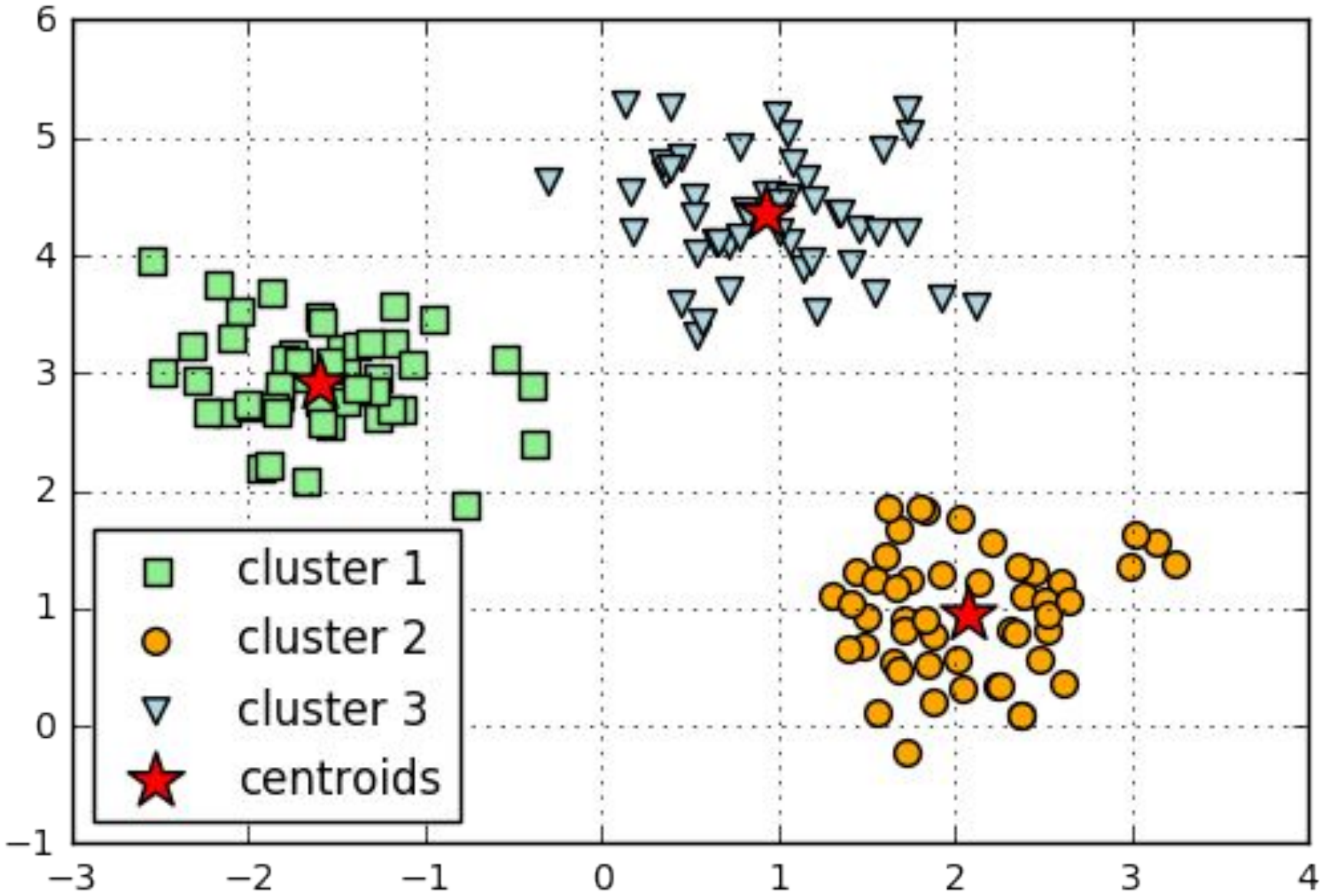
Es un algoritmo de aprendizaje no supervisado utilizado para la agrupación en clústeres de datos, que agrupa puntos de datos no etiquetados en grupos o clústeres.

En K-Means los datos pueden existir en un solo clúster. Este tipo de análisis se utiliza habitualmente en la ciencia de datos para la segmentación de mercados, la agrupación de documentos, la segmentación de imágenes.

K-Means

Es un algoritmo iterativo basado en centroides que divide un conjunto de datos en grupos similares en función de la distancia entre sus centroides. El centroide, o centro del clúster, es la media o la mediana de todos los puntos dentro del clúster, según las características de los datos.

K-Means



K-Means

Categoriza puntos de datos en clústeres utilizando una medida matemática de distancia, generalmente euclidiana, desde el centro del clúster.

El objetivo es minimizar la suma de distancias entre los puntos de datos y sus clústeres asignados. Los puntos de datos más cercanos a un centroide se agrupan dentro de la misma categoría.

Un valor k más alto, o el número de conglomerados, significa conglomerados más pequeños con mayor detalle, mientras que un valor k más bajo da lugar a conglomerados más grandes con menos detalle.

K-Means

Existen diferentes implementaciones y librerías que permiten utilizar K-Means.

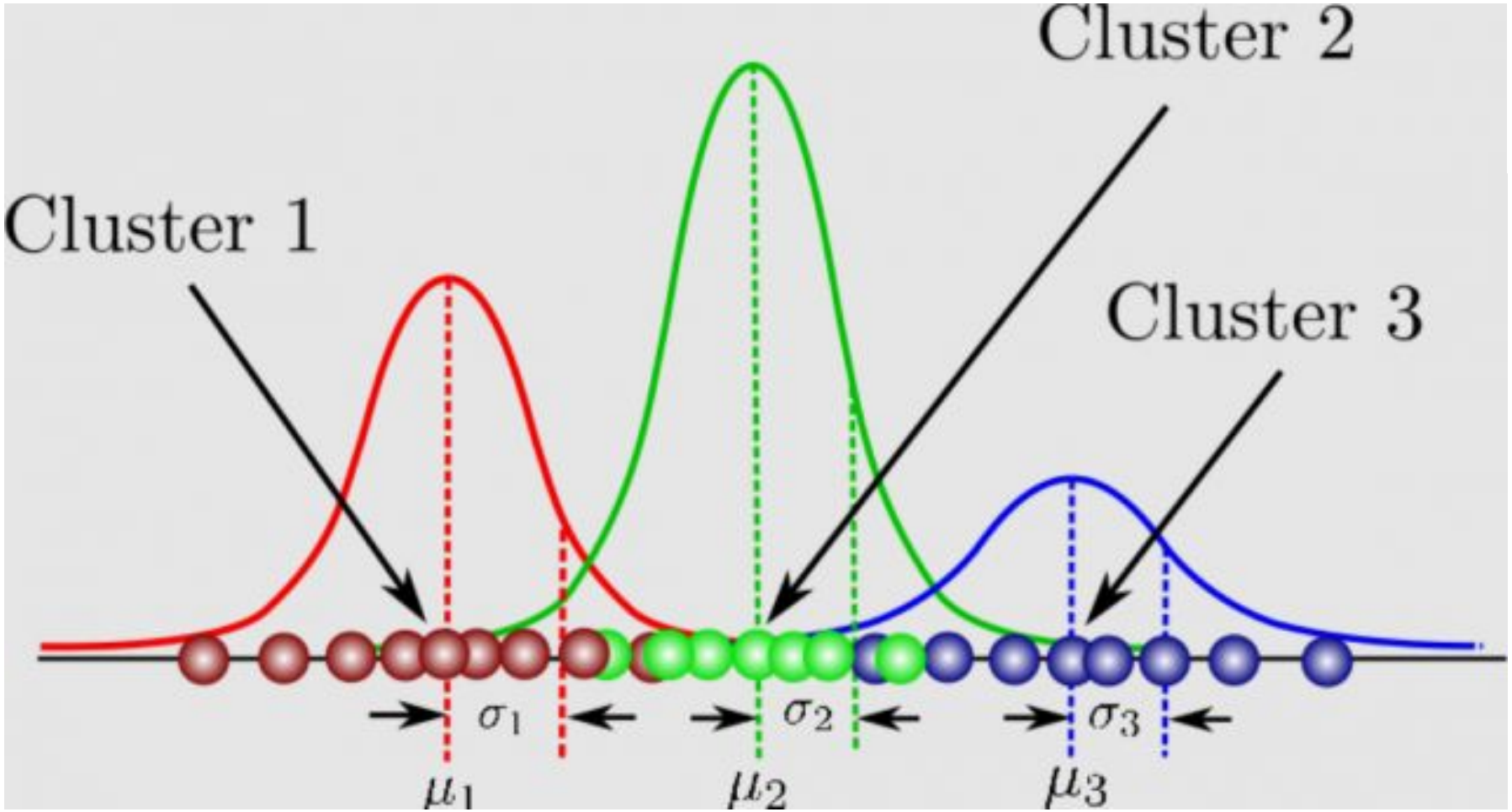
En la gran mayoría de las implementaciones es necesario definir una semilla (seed en PySpark) o un valor random inicial (random_state en SciKit-Learn). Esto se debe a que el algoritmo K-Means parte de centroides ubicados aleatoriamente (luego iterativamente se van ajustando). El valor seed o random_state permite que los experimentos sean reproducibles y a su vez, permiten que el algoritmo puede "iniciar".

Mixtura Gaussiana

Los Modelos de Mezclas Gaussianas (GMM por sus siglas en inglés) son modelos probabilísticos construidos a partir de una suma ponderada de distribuciones de probabilidad Gaussianas.

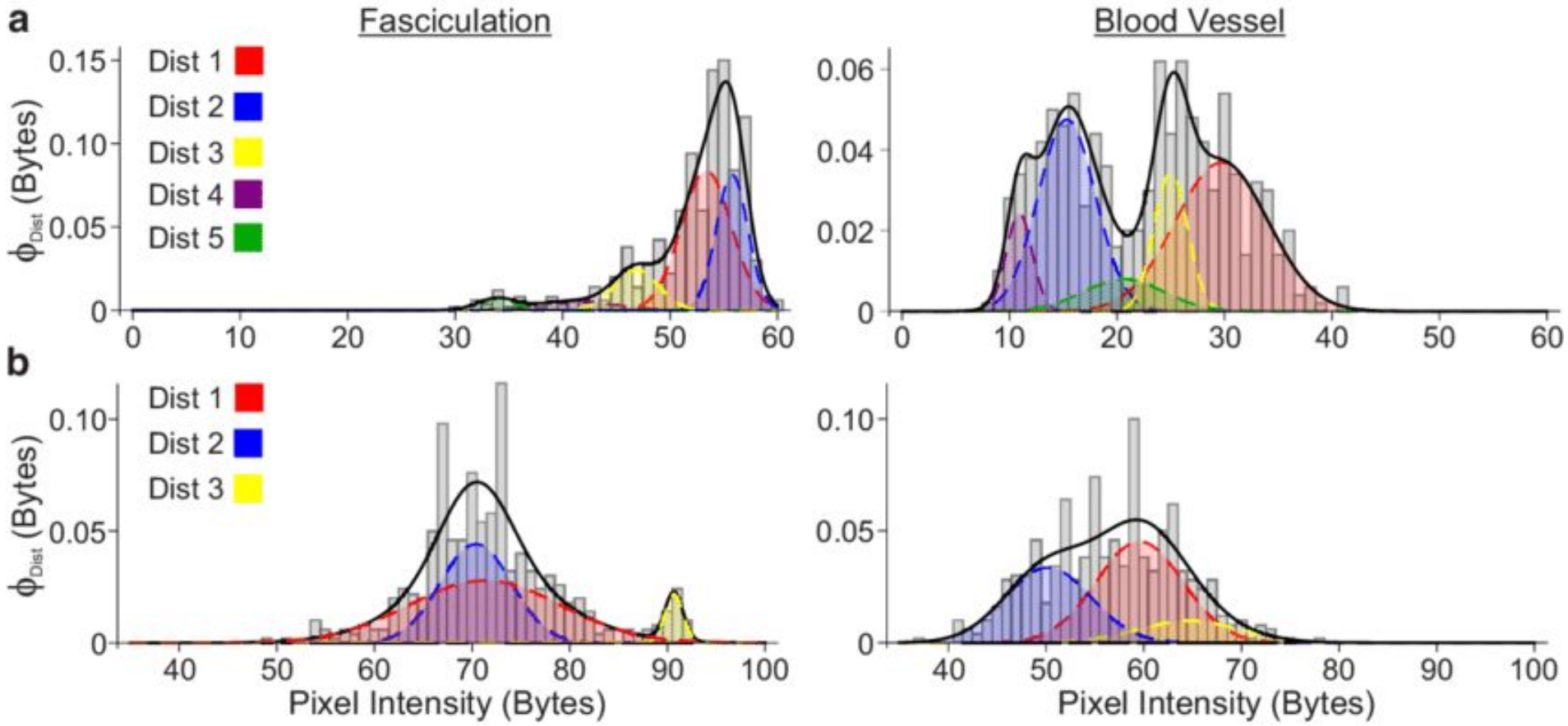
Los GMM son útiles para modelar fenómenos aleatorios, pero su aplicabilidad se puede extender a contextos como Machine Learning (ML). Dentro de este contexto los GMM se pueden emplear como clasificadores.

Mixtura Gaussiana



Se asocia a cada campana un peso junto con la especificación de cada normal: media y varianza

Mixtura Gaussiana



https://www.researchgate.net/publication/331623263_Foreground_Detection_Analysis_of_Ultrasound_Image_Sequences_Identifies_Markers_of_Motor_Neurone_Disease_across_Diagnostically_Relevant_Skeletal_Muscles

Mixtura Gaussiana

A fines prácticos, es muy similar a K-Means. Sin embargo, muchas veces resulta útil asignar probabilidades de pertenencia a cada cluster, en vez de directamente asignar una etiqueta. Lo anterior lo podemos hacer con Modelos Mixtos Gaussianos (GMM por sus siglas en inglés).

fuente: https://rpubs.com/dapivei/705612

Ejemplos!!!