UT4. Aprendizaje automático no supervisado

IES Abastos

CE Inteligencia Artificial y Big Data

Sistemas de Aprendizaje Automático

Bloques de la unidad:

1. Introducción al clustering.
2. K-Means.
3. Clustering Jerárquico.
4. Introducción a técnicas de detección de anomalías.
5. Aplicaciones y retos detección de anomalías.
6. Tipos de anomalías.
7. Técnicas de detección de anomalías.
8. Prácticas.

0. Esquema


Aprendizaje Automático
Aprendizaje No Supervisado
Agrupamiento
Algoritmo K-Medias
Agrupamiento Jerárquico


Aprendizaje Automático
Aprendizaje Supervisado y No Supervisado
Técnicas de Detección de Anomalías
Desarrollando y Evaluando un Sistema de Detección de Anomalías
Detección de Anomalías vs. Aprendizaje Automático

- El clustering es una técnica de aprendizaje automático que divide los datos en clusters o grupos similares. Se lleva a cabo sin necesidad de indicar las características de cada uno de estos grupos.
- Las instancias dentro de un grupo deben ser muy **similares**, pero muy **distintas** entre los **grupos**. Es clave poder definir una medida de similitud entre elementos.

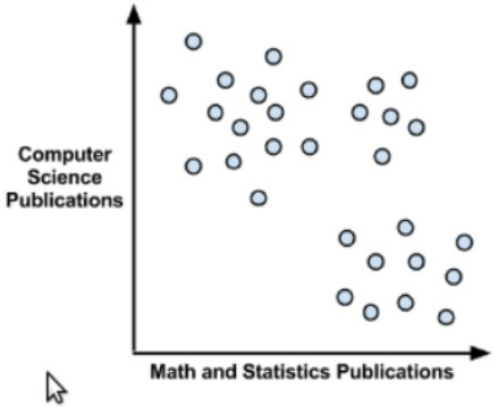
El Clustering se utiliza en Aplicaciones como:

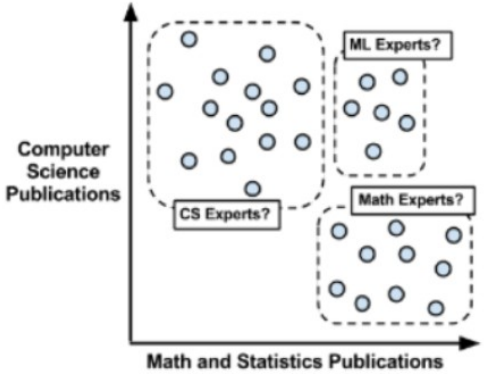
- Segmentar clientes entre grupos con patrones de compra similares.
- Detectar comportamiento anómalo, identificando patrones que caen fuera de los clusters habituales.
- Simplificar o resumir datasets muy grandes, agrupando usuarios con similares características.

Medidas de similitud.

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the Covariance matrix

En el clustering, instancias sin etiquetar son proporcionadas con el objetivo de inferir las relaciones en los datos.





Se pueden dividir en dos grandes grupos:

- **Agrupación:** Tienen definidos previamente un número de grupos. Son algoritmos iterativos que comienzan con una asignación inicial y se van modificando siguiendo un criterio de optimización.
- **Jerárquicos:** En cada iteración solo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no se vuelve a cambiar.

2. K-Medias

- Es el algoritmo de clustering más utilizado.
- Entendiendo su funcionamiento podemos entender casi cualquier algoritmo de clustering utilizado hoy en día.
- Existen diversas implementaciones del algoritmo KMeans

2. K-Medias

El algoritmo consiste en asignar cada uno de los n ejemplos a uno de los k clusters, donde k es un número definido previamente.

El objetivo es minimizar las diferencias entre los grupos de cada cluster y maximizar las diferencias entre clusters.

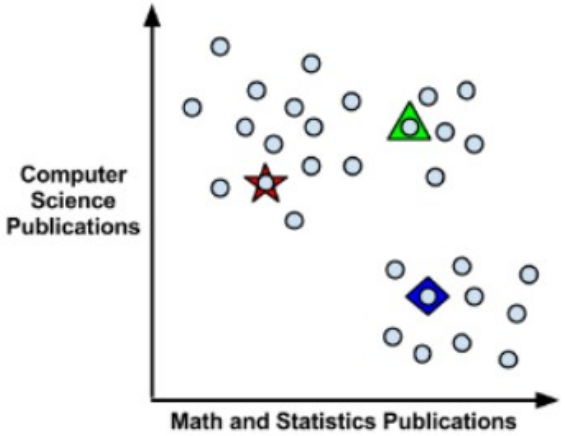
A menos que k y n sean extremadamente pequeños no es factible calcular los grupos óptimos entre todas las combinaciones posibles de ejemplos. En su lugar el algoritmo utiliza un proceso heurístico para calcular la solución óptima.

El Algoritmo comprende 2 fases.

Primero asigna ejemplos a un conjunto inicial de k clusters. Después actualiza las asignaciones ajustando los límites de los grupos de acuerdo con los ejemplos de cada cluster.

Este proceso de asignación y actualización ocurre varias veces hasta que los cambios no proporcionan mejoras en los clusters.

El algoritmo empieza eligiendo los k puntos iniciales.



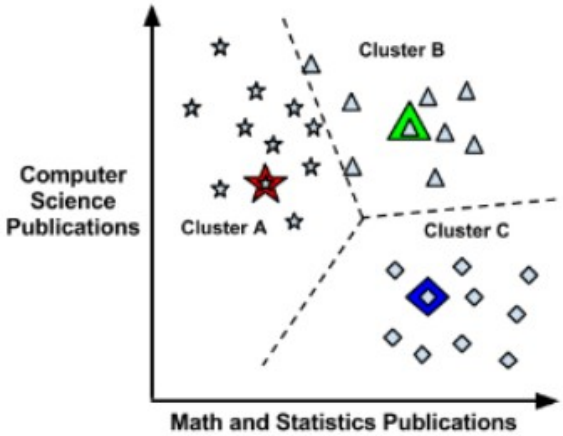
2. K-Medias

Para los puntos iniciales se suelen elegir 3 ejemplos al azar. Otras opciones son elegir puntos que pueden ocurrir en cualquier intervalo del feature space.

Después de elegir los puntos iniciales, los otros ejemplos se asignan al centroide del cluster más cercano de acuerdo a la función de distancia. Está función suele ser la distancia Euclídea.

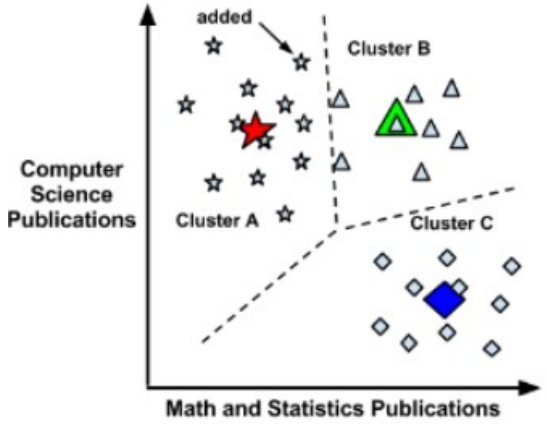
$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Debido a que estamos utilizando distancias, todos los puntos deben de ser numéricos y normalizados antes de utilizarlos.

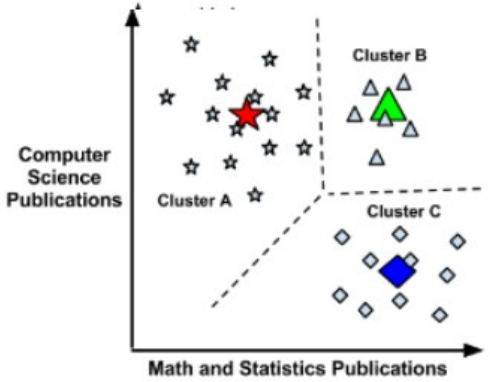


2. K-Medias

El primer paso de la actualización conlleva mover los centroides iniciales a una nueva posición, calculada como la media de los puntos asignados al cluster.



Como los límites de los centroides se han modificado es muy posible que haya que reasignar instancias a otros clusters.



Elegir un buen valor requiere de cierto balance, un número muy grande mejora la homogeneidad pero a la vez sobre-ajusta los datos.

Idealmente existe un conocimiento a-priori sobre el número de grupos apropiado. Por ejemplo, si estamos agrupando películas se pueden agrupar por los géneros existentes. Otras veces el número de clusters viene dado por los requisitos de negocio.

Sin conocimiento a priori se suele elegir K = sqrt(n/2)

También se pueden usar métricas para medir la homogeneidad versus la heterogeneidad. A través de un proceso de experimentación.

2. K-Medias

Debido a la naturaleza heurística del algoritmo los resultados pueden ser distintos en función de la inicialización del algoritmo.

Sin embargo, si los resultados difieren mucho puede resultar un problema. Por ejemplo los datos no se pueden separar bien en los k clusters.

K-Medias considera que los valores de las variables soncoordenadas en un espacio multi-dimensional.

El Algoritmo comprende 2 fases.

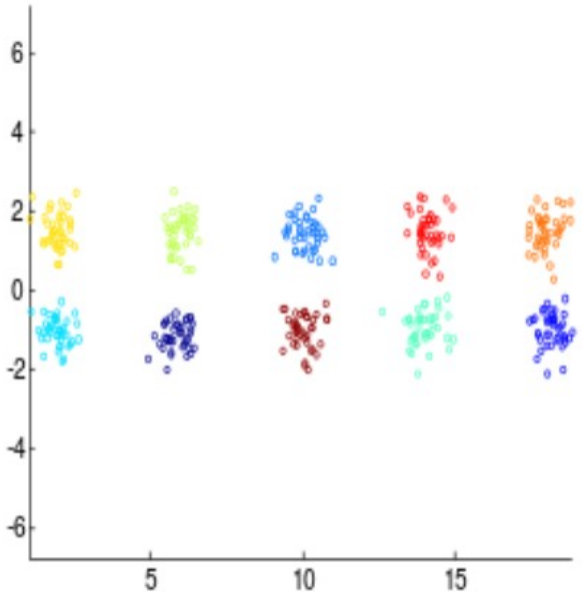
Primero asigna ejemplos a un conjunto inicial de k clusters. Después actualiza las asignaciones ajustando los límites de los grupos de acuerdo con los ejemplos de cada cluster.

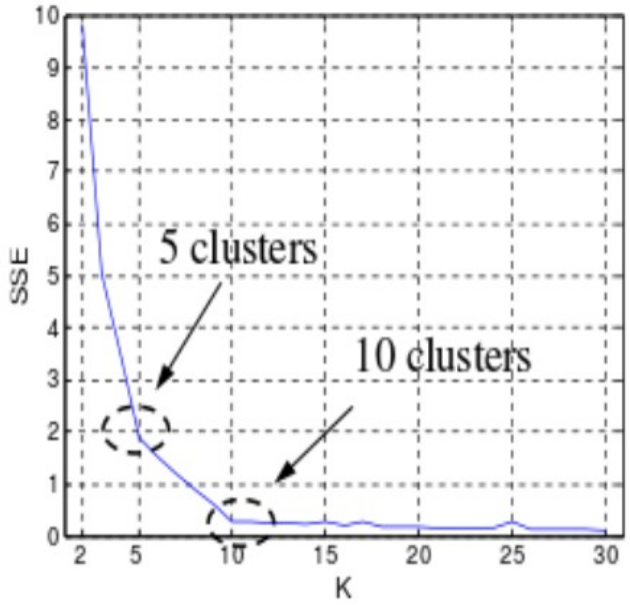
Este proceso de asignación y actualización ocurre varias veces hasta que los cambios no proporcionan mejoras en los clusters.

- Elegir un buen valor requiere de cierto balance, un número muy grande mejora la homogeneidad pero a la vez sobre-ajusta los datos.
- Idealmente existe un conocimiento a-priori sobre el número de grupos apropiado. Por ejemplo, si estamos agrupando películas se pueden agrupar por los géneros existentes. Otras veces el número de clusters viene dado por los requisitos de negocio.

- Sin conocimiento a priori se suele elegir K = sqrt(n/2).
- También se pueden usar métricas para medir la homogeneidad versus la heterogeneidad.
- Teniendo en cuenta la suma del error cuadrático medio para diferentes valores de K

Técnica del codo, ejemplo: Por mucho que hagamos más de 10 grupos el error ya no baja.





La inicialización también es un factor clave.

Fortalezas

• Utiliza principios simples para identificar clusters, los cuales se pueden explicar en términos no estadísticos.

• Es muy flexible y se puede adaptar para tener en cuenta muchas de sus limitaciones de forma simple.

• Es eficiente y se comporta bien al dividir los datos en clusters útiles.

Debilidades

- Debido a que tiene un componente aleatorio es posible que no encuentre el conjunto óptimo de clusters.
- Requiere una intuición razonable de los clusters que pueden existir en los datos.

Clustering Jerárquico

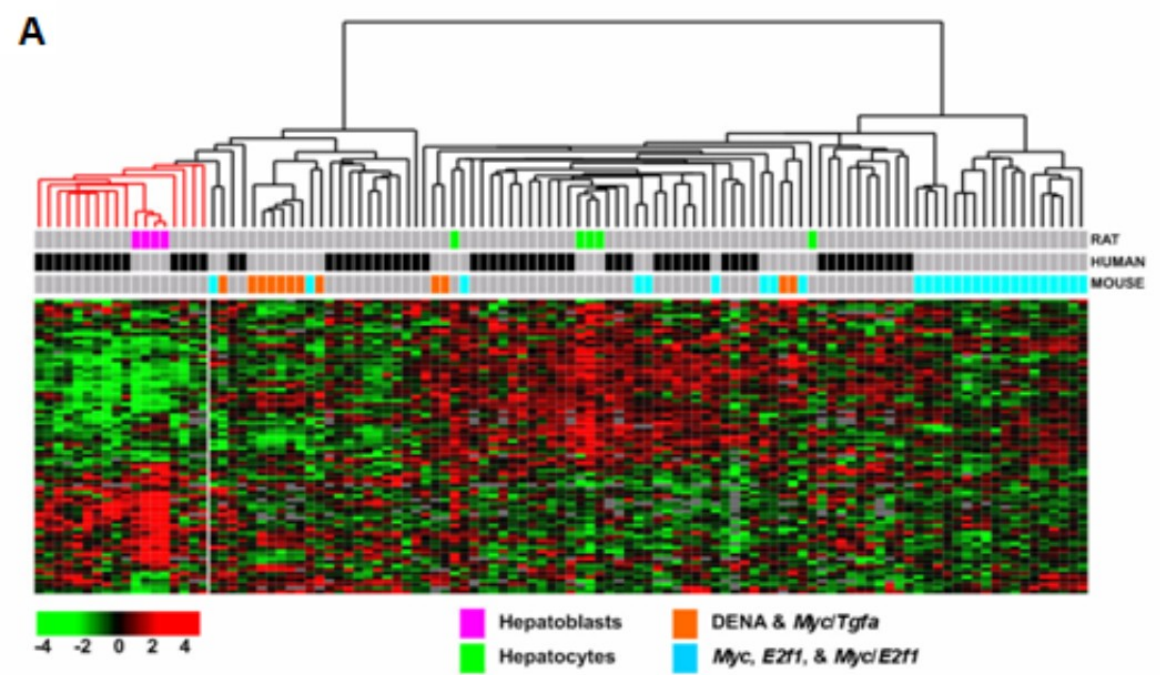
- No conocemos de antemano cuántos clusters queremos.

- Obtenemos una representación basada en árboles llamada **dendograma** que nos permite visualizar los grupos obtenidos para cada posible número de clusters de 1 a n.

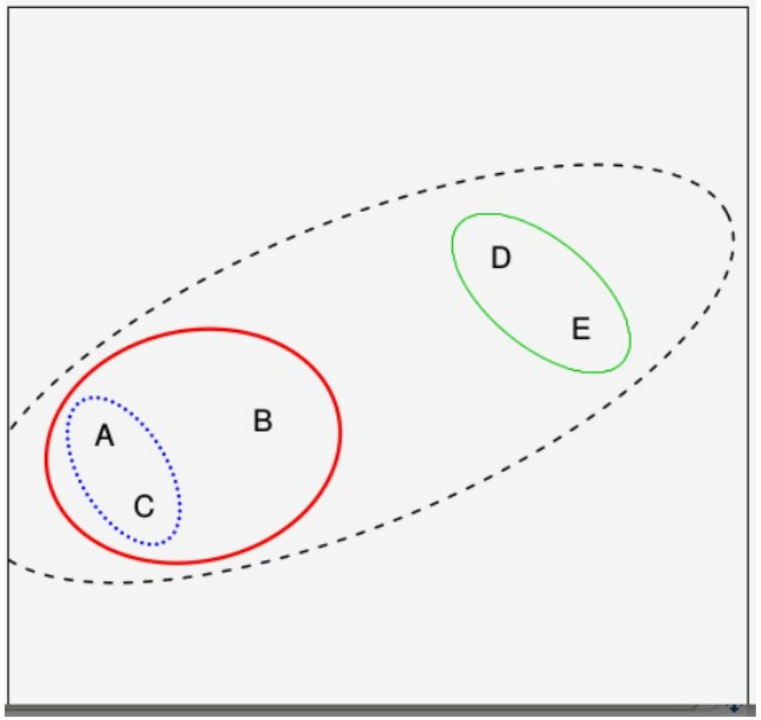
Pueden ser de dos tipos:

•Agglomerative: Es un método “bottom-up” cada observación empieza en un cluster y los pares de clusters se combinan cuando se avanza hacia arriba en la jerarquía.

•Divisive: Es un método “top-down” en donde todas las observaciones empiezan en un cluster y se van haciendo divisiones hacia abajo.


A
RAT
HUMAN
MOUSE
-4 -2 0 2 4
Hepatoblasts
Hepatocytes
DENA & Myc/Tgfa
Myc, E2f1, & Myc/E2f1

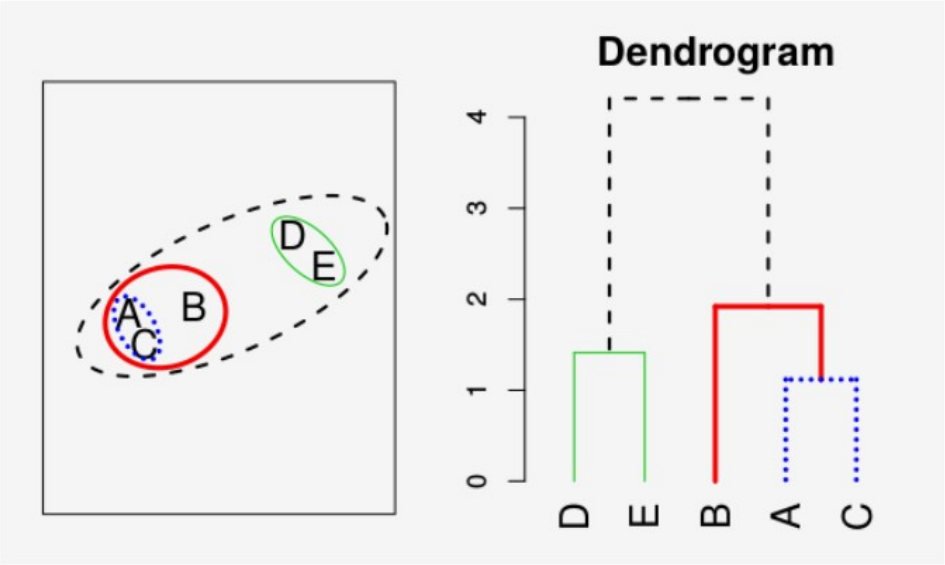
Clustering jerárquico (Concepto).



Algoritmo:

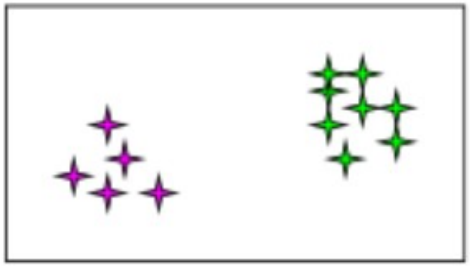
Empezar con cada punto en su propio cluster
  Identificar los dos clusters más cercanos y combinarlos
Repetir
Terminar cuando todos los puntos estén en un cluster

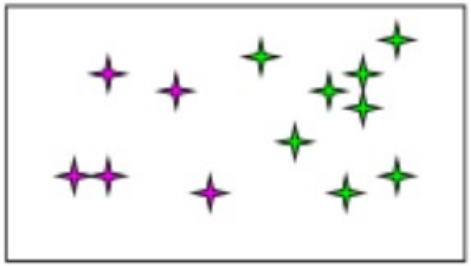
Algoritmo:



¿ Cómo evaluar el cluster ?

- Por interpretación: cada cluster corresponde a una agrupación válida?
- Criterio división: similitud intra-cluster alta y similitud intercluster baja.



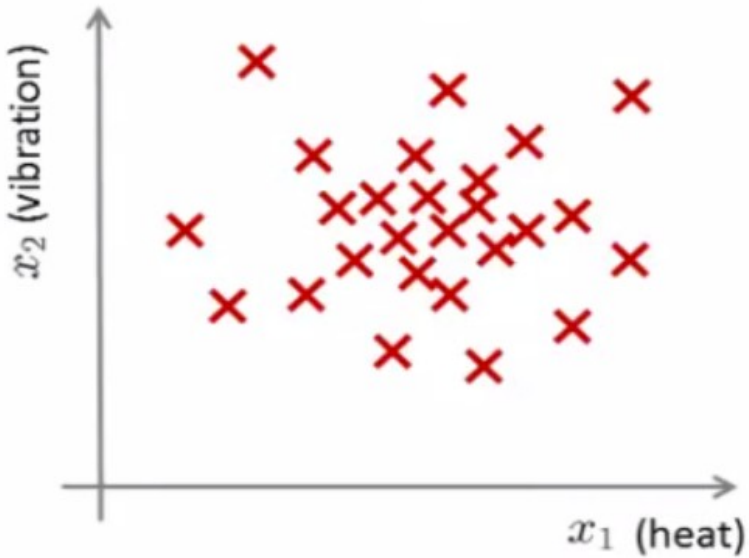


La métrica de calidad de un clustering depende tanto de la representación de los objetos como de la métrica de similitud utilizada.

Empresa que fabrica motores de aviones.

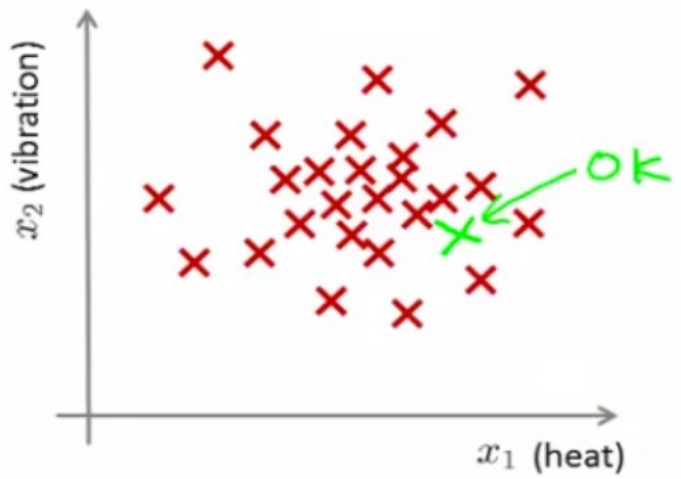
Se desea realizar un aseguramiento de la calidad donde se mide el calor generado y la vibración.

Después de evaluar un conjunto de m motores, tenemos la siguiente distribución.

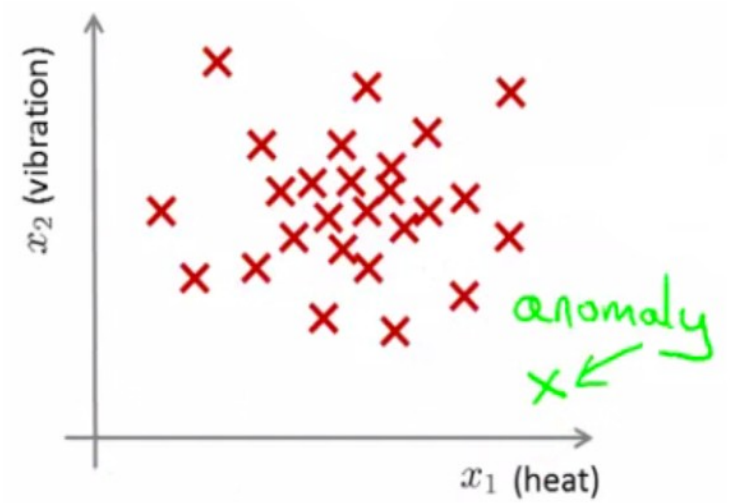


Cuando se genera un nuevo motor, se compara este con respecto a los motores previos.

Si obtenemos una gráfica como esta:



Sin embargo, si obtenemos una gráfica como esta:



Partimos de un conjunto de datos que contiene registros “normales”, o bien la mayoría de ellos lo son.

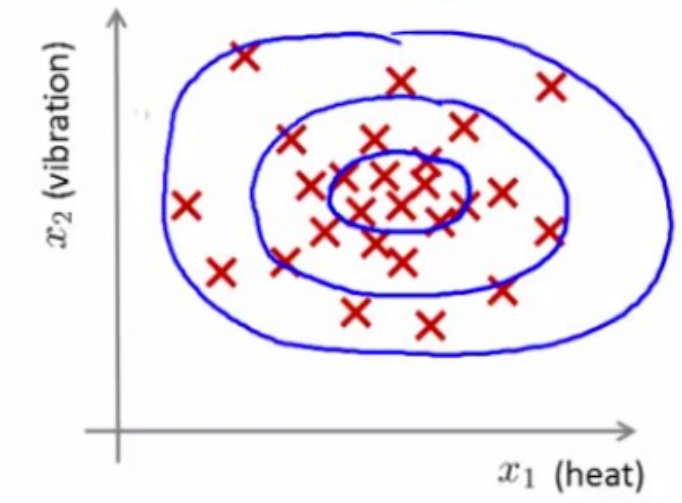
El objetivo es utilizar este conjunto como referencia para detectar anomalías en nuevos ejemplos.

Para ello se construye un modelo que indica la probabilidad de que un ejemplo sea normal.

- Una vez construido el modelo, buscamos respuesta a la pregunta de:

Si $p(X_{test}) < \varepsilon$ se trata de una anomalía.
Si $p(X_{test}) \geq \varepsilon$ se trata de un ejemplo normal.

En el caso de un modelo en 2 dimensiones:



Patrones de compra anómalos: horario de compra, localización de compra, importe de compra.

Patrones de acceso web anómalos: volumen de páginas visitas, horario de las visitas, distribución de frecuencia de las visitas.

Detección de fraude: localización de login, duración de tiempo on-line.

Detección de fallos en una cadena de montaje.

Monitorización de Data-Center: uso de la memoria, accesos al disco, carga de la CPU.

¿ Se te ocurre algo más ?

Retos:

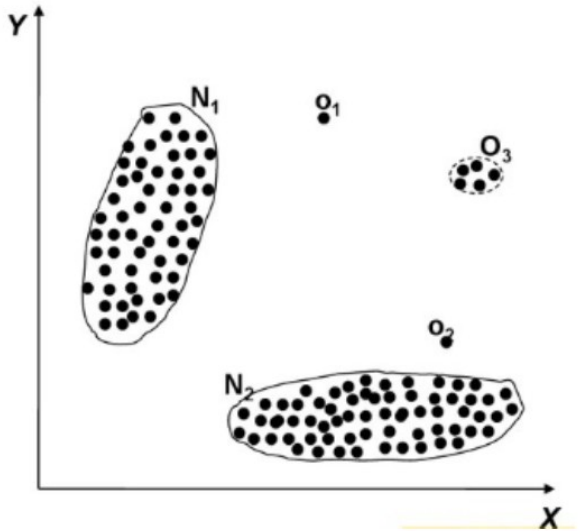
- Definir región normal
- El límite entre lo “normal” y el “outlier” a veces no está claramente definido.
- La noción de outlier es diferente para distintos dominios.
- Disponibilidad de datos para “aprender”

Suele haber muy pocos ejemplos de anomalías.

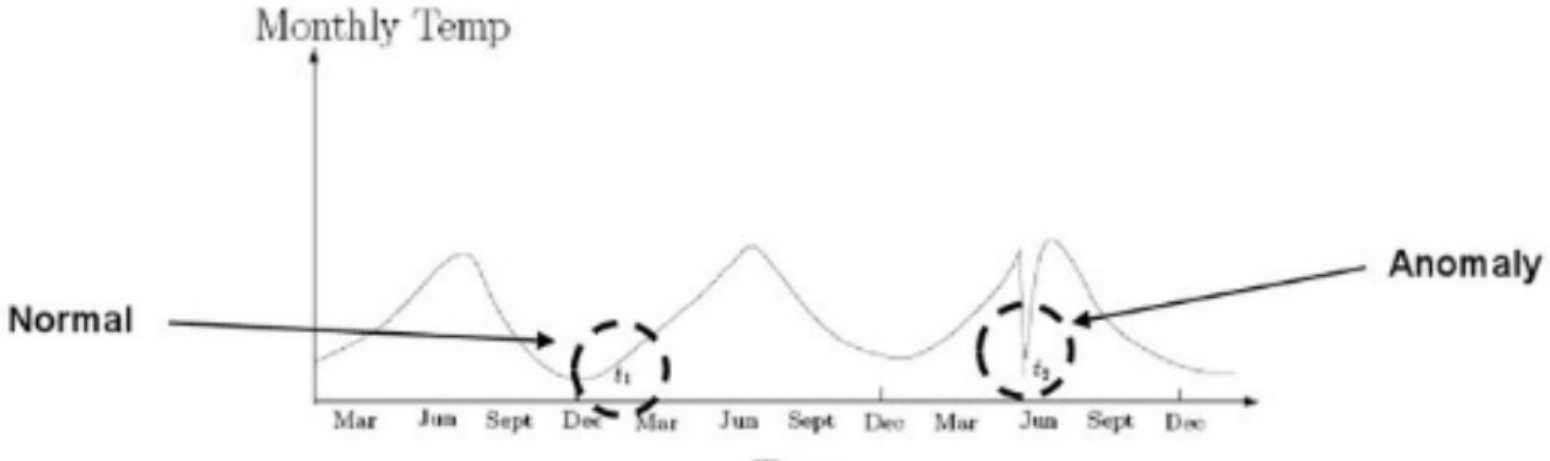
Si utilizaramos Aprendizaje supervisado.

¿ Qué pasaría ?

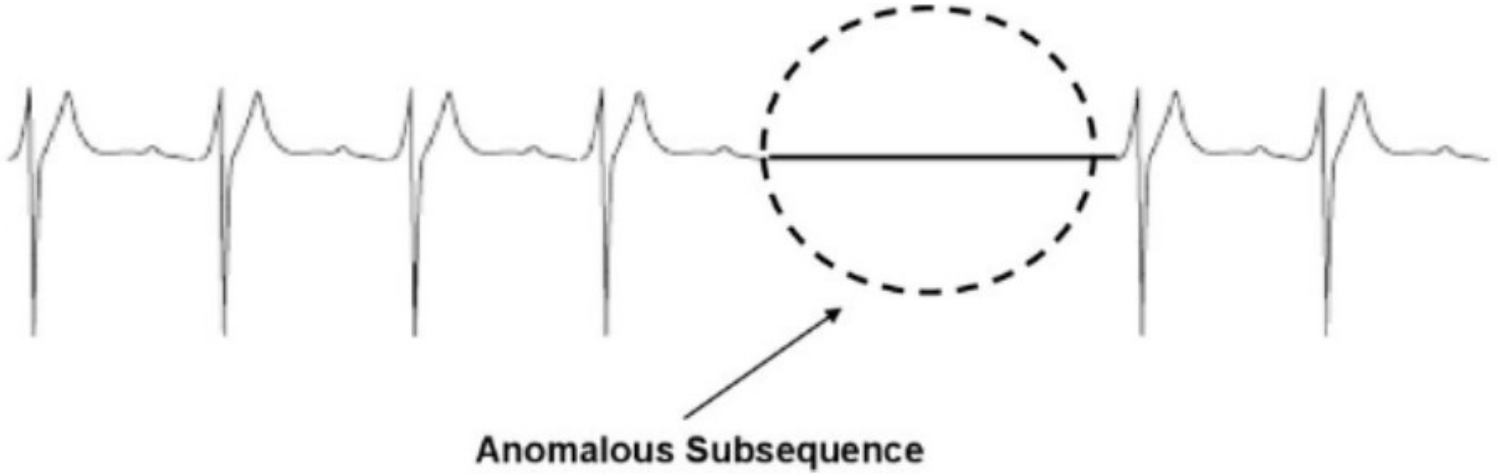
Se trata de una instancia que es anómala con respecto de los datos.



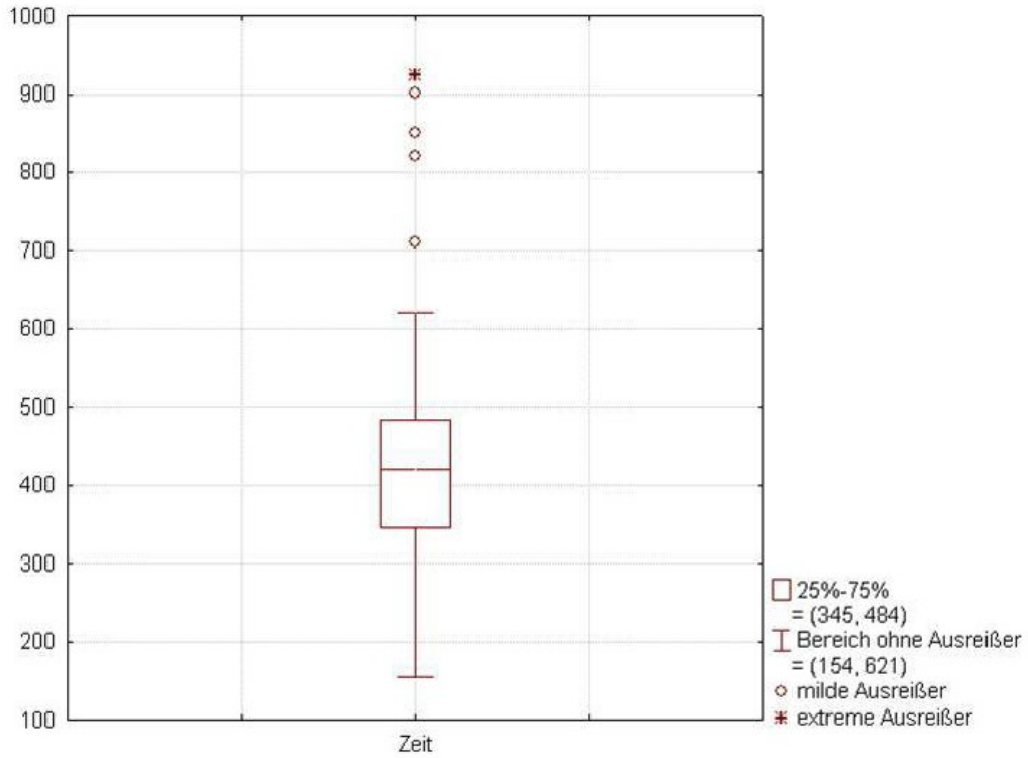
Una instancia es anómala con respecto a un contexto. También se conocen como anomalías condicionales.



Una colección de instancias son anómalas. Requiere una relación entre las instancias: secuencial, espacial, grafos.



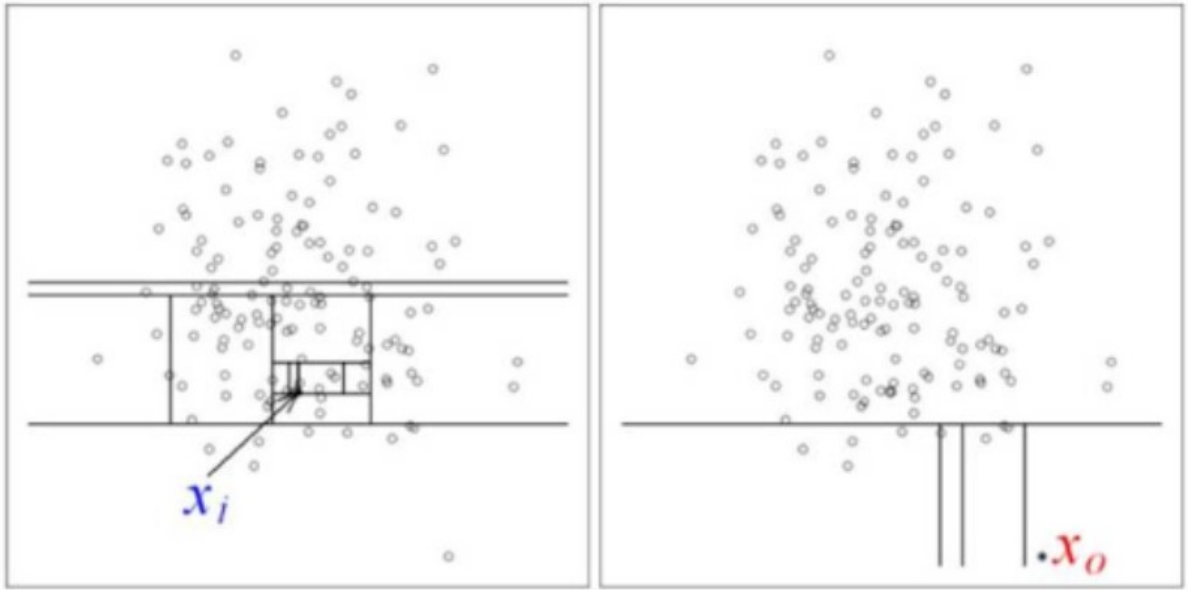
7. Técnicas de detección de anomalías



En el box plot o diagramas de diagrama de Box-Whisker estamos utilizando medidas.

Con Isolation Forest vamos a detectar anomalías aislando instancias sin tener en cuenta una medida de distancia o densidad.

7. Técnicas de detección de anomalías



(a) Isolating x_i (b) Isolating x_o

mas de