Fundamentos de Ciencias de Datos

Semana 14 - Outliers

Outliers

Un **outlier** es un punto considerablemente "diferente" al resto de los datos

Ya hemos hablado de la presencia de ciertos outliers en los datasets que hemos revisado en el curso

Pero, ¿cómo detectamos los outliers?

Outliers

Búsqueda

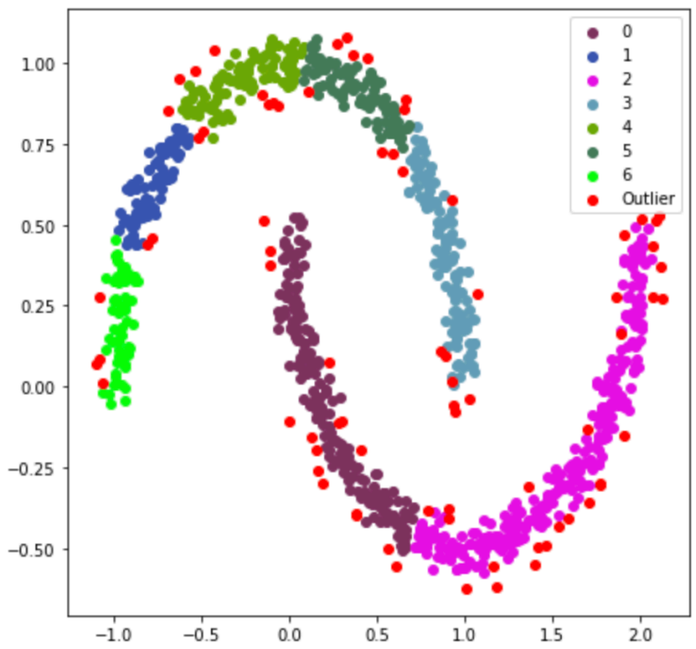
Si tenemos una columna, podemos definir un **outlier** como puntos con valor por sobre/bajo cierto threshold

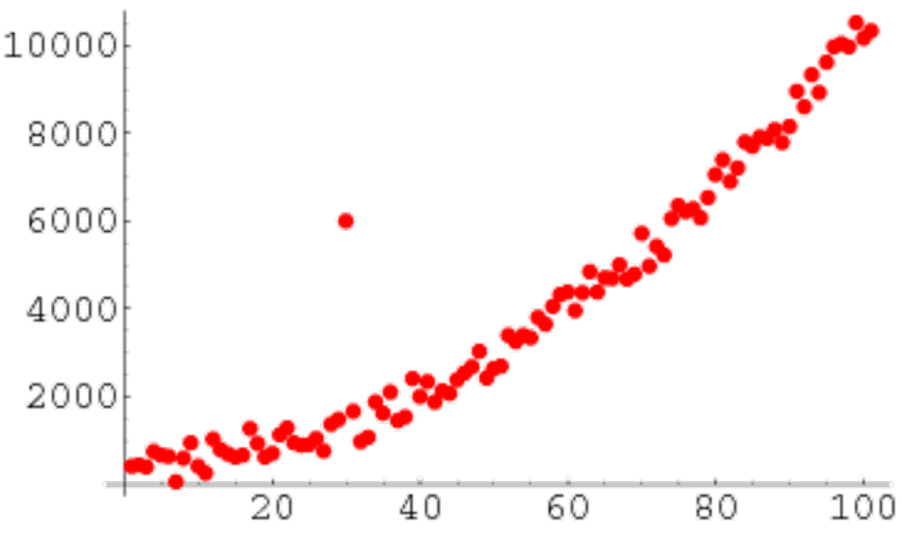
Podemos usar técnicas visuales para ver puntos alejados de donde se encuentra "la mayoría de los puntos"

También existen técnicas estadísticas para detectar outliers

Outliers

Ejemplos





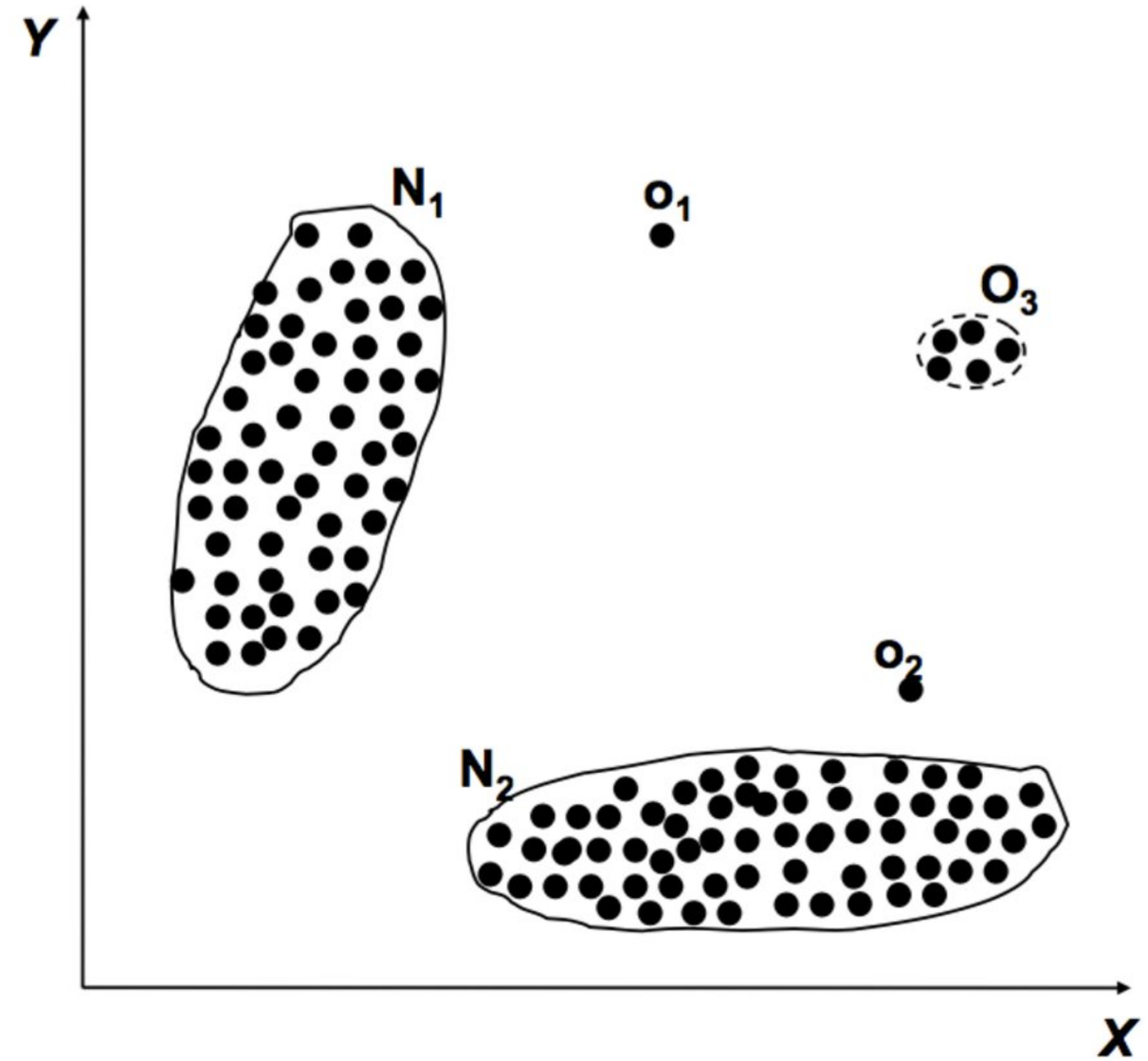
Outliers

Ojo. La noción de outlier es subjetiva, y depende del dominio de nuestro problema

Los métodos de detección buscan un patrón para la mayoría de los datos, y luego se buscan anomalías para esa distribución

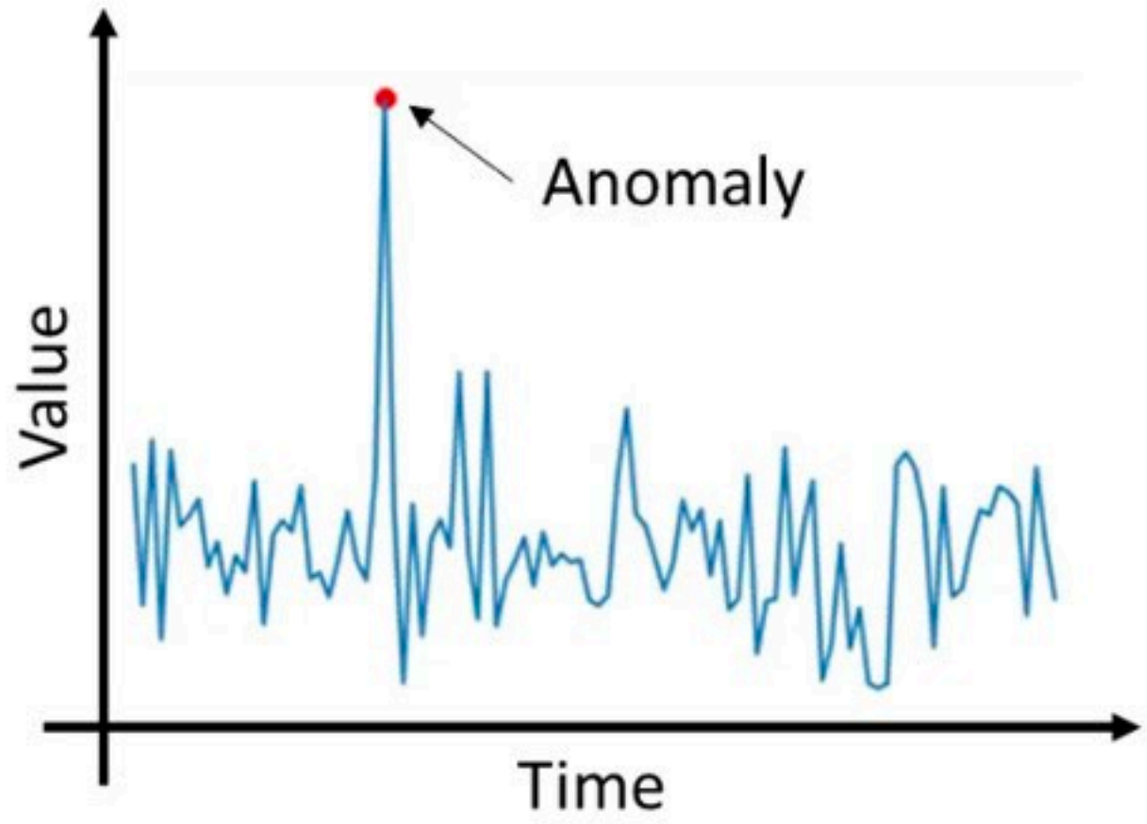
Outliers

Ejemplo - puntos anómalos



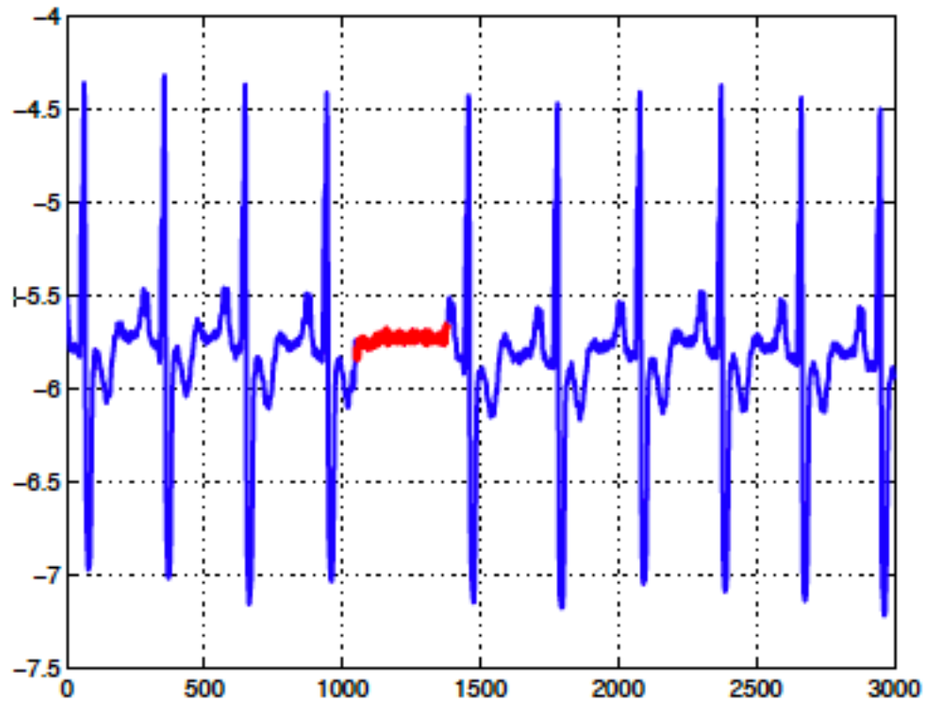
Outliers

Ejemplo - anomalía en el contexto



Outliers

Ejemplo - secuencia anómala



Outliers

Pensemos por ejemplo en las ventas de alcohol gel, mascarillas, desinfectantes antes del 2020 y durante el 2020

¿Crees que cambia la distribución?

Detección de Outliers

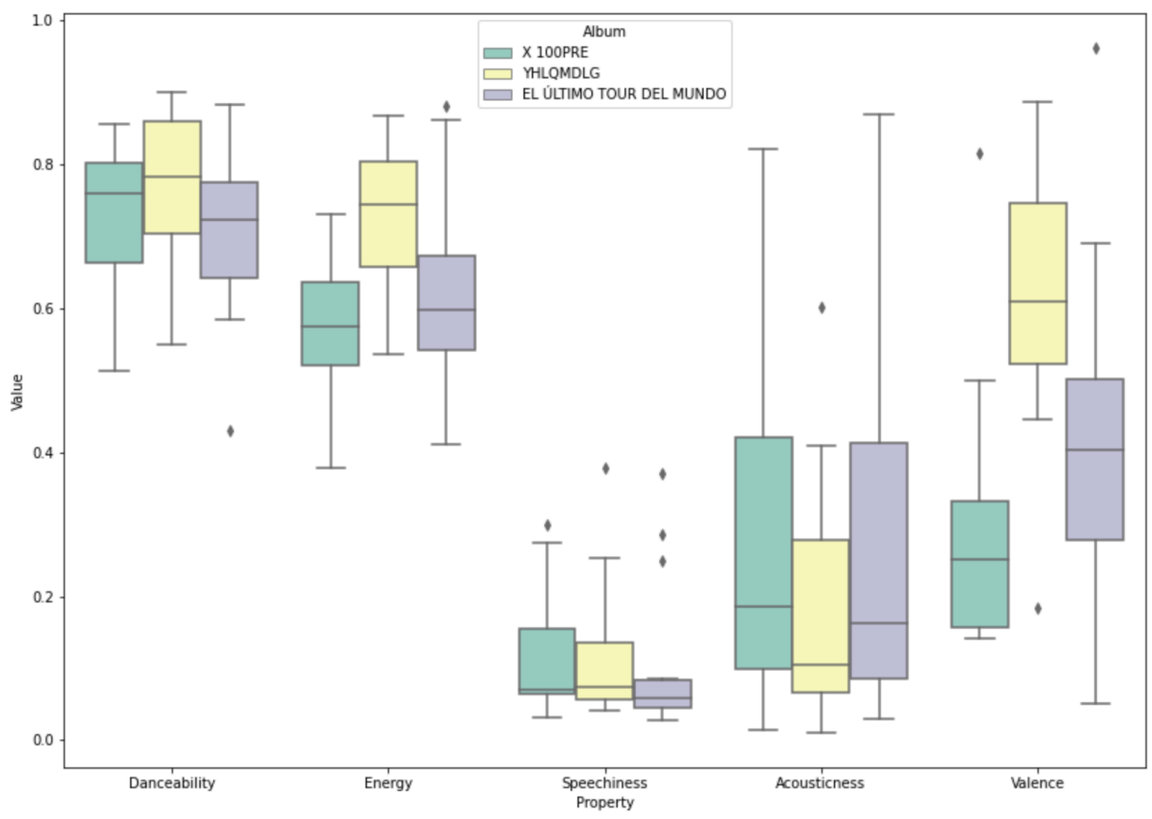
Métodos para encontrar outliers:

- Visuales
- Estadísticos
- Basados en distancia
- ...

Detección de Outliers

Boxplot

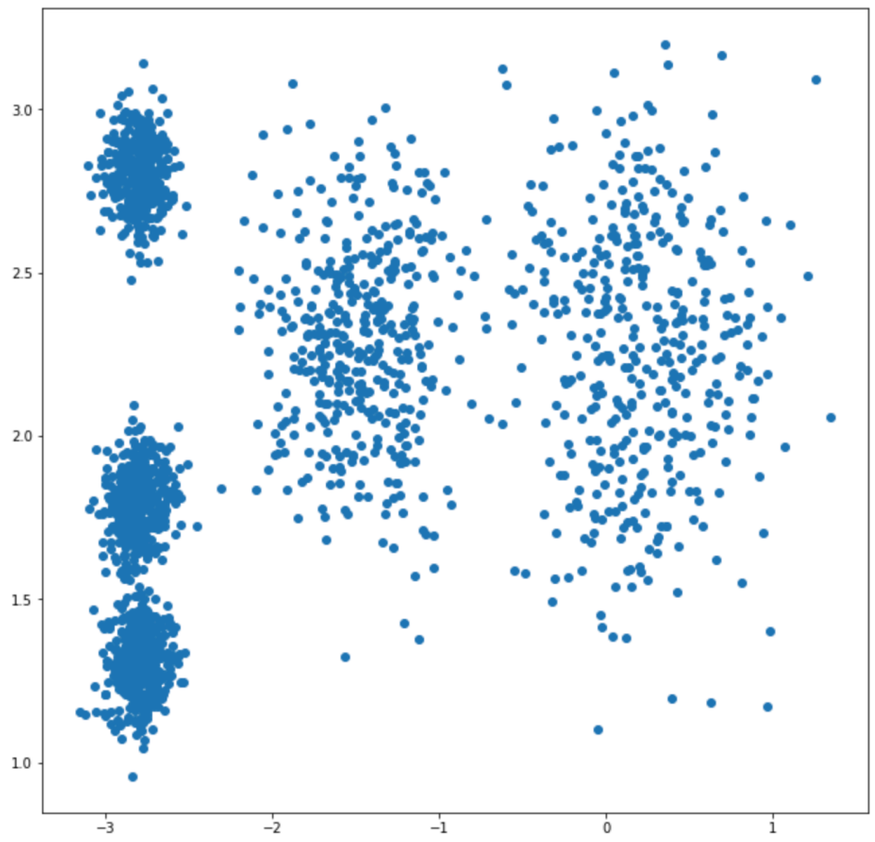
En un boxplot los puntos más allá de las barras indican un outlier; en general un outlier está a más de 1.5 veces el rango intercuartil



Detección de Outliers

Scatterplot

También podemos ver outliers en los scatterplot, por ejemplo, ¿cuáles creen que son outliers en la siguiente imagen?



Detección de Outliers

GMM

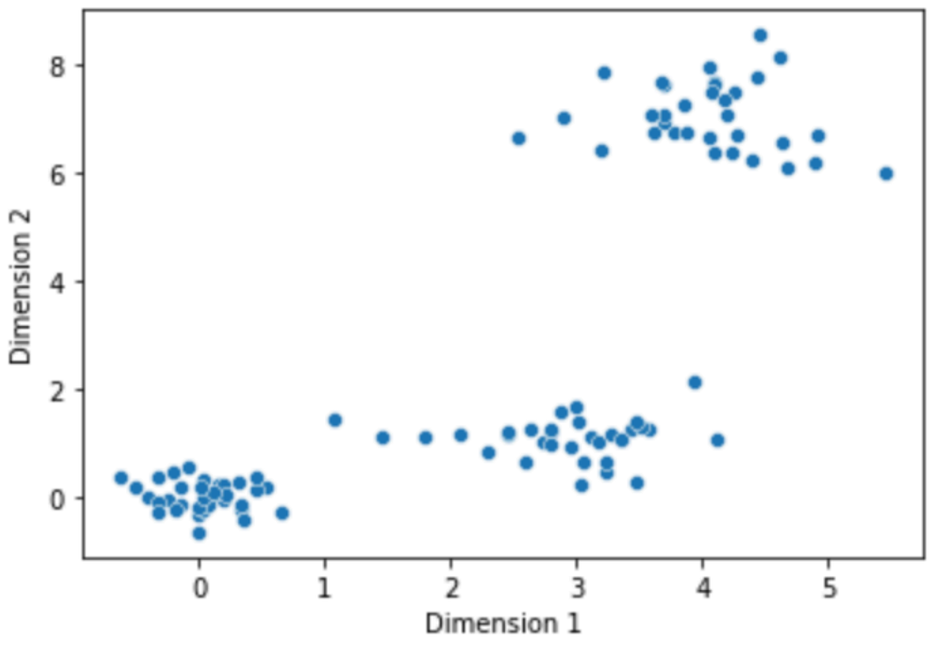
Una técnica famosa para encontrar outliers es usar un modelo parecido a K-Means llamado Gaussian Mixture Model (GMM)

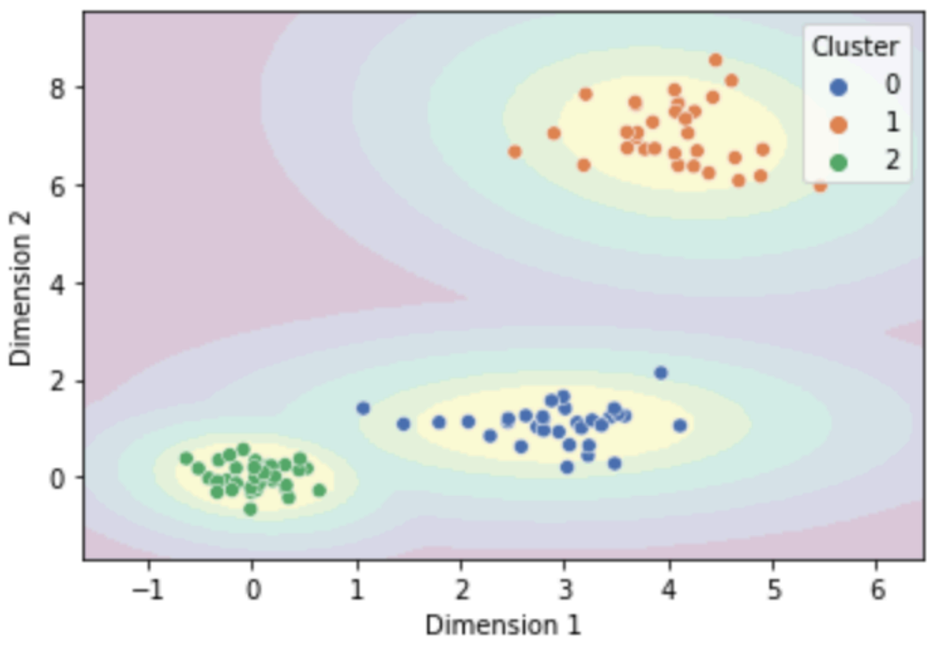
Este algoritmo, en vez de trabajar solamente con un centroide, también trabaja con la matriz de covarianza

Este modelo, a diferencia de K-Means, captura mejor clusters que son "ovalados"

Detección de Outliers

GMM





Detección de Outliers

GMM

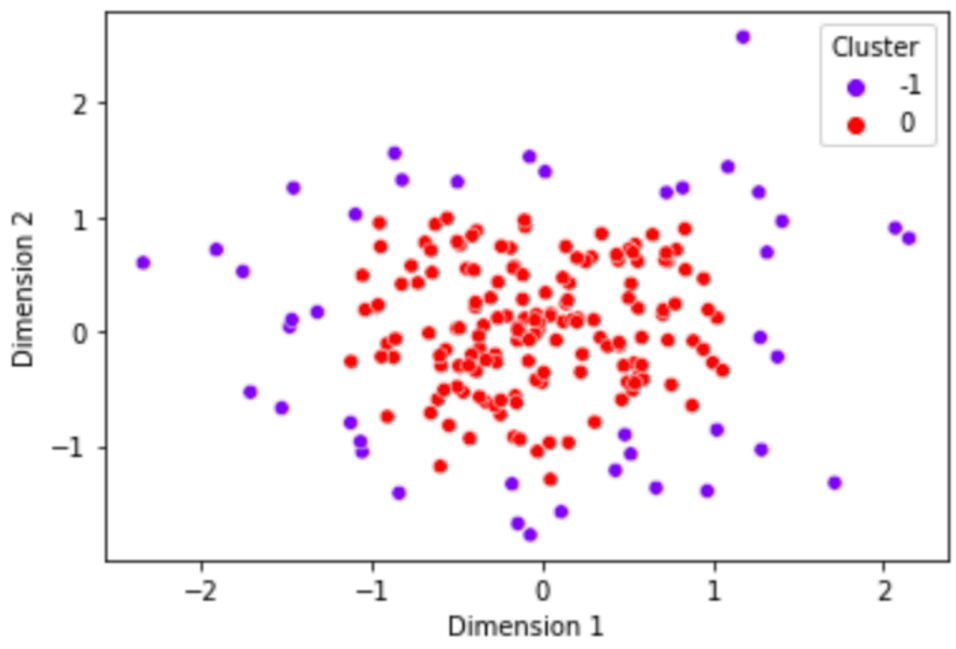
Aquí asumimos que cada *cluster* viene de una distribución gausiana (generalizada para **n** dimensiones)

Si hay un punto en las zonas donde la densidad de probabilidad es baja, se clasifica como outlier

Detección de Outliers

DBSCAN

En la clase de clustering vimos el algoritmo DBSCAN, que busca zonas de alta densidad



Detección de Outliers

DBSCAN

- Para cada instancia, contamos cuantas instancias están dentro de un rango ϵ
- Si una instancia tiene al menos m instancias cerca (según ϵ), se considera una instancia *core*
- Todas las instancias en el vecindario de una instancia *core* pertenecen al mismo cluster; esta vecindad puede tener otra instancia *core*
- Una secuencia de instancias core adyacentes forman un cluster
- Toda instancia no core es un outlier

Detección de Outliers

KNN

Otra técnica utilizada para detectar outliers es KNN

Este es un modelo de clasificación supervisado que funciona de la siguiente manera: veo mis K vecinos más cercanos, y mi etiqueta será la etiqueta más repetida entre mis vecinos

Detección de Outliers

KNN

Ahora, ¿qué pasa si mis K vecinos más cercanos están muy lejos?

Detección de Outliers

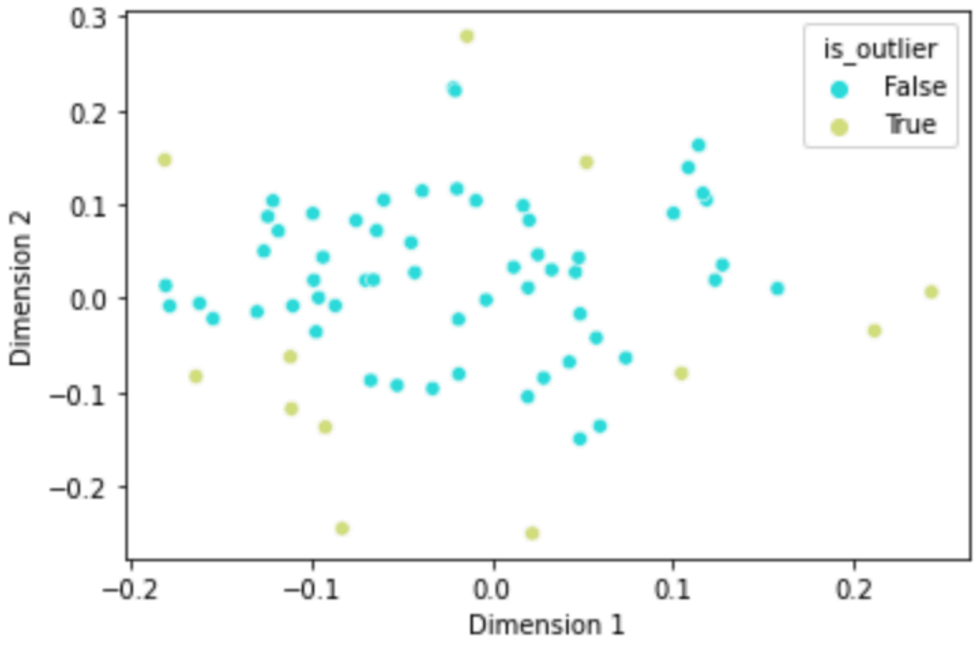
KNN

KNN puede clasificar como outlier en alguno de los siguientes casos:

- Calculo la distancia al **k** vecino más cercano para cada punto y los ordeno de mayor a menor según esta distancia; los primeros **p** puntos (yo escojo **p**) son outliers
- En vez de usar la distancia al **k** vecino ocupo el promedio de los primeros **k** vecinos
- Aquellos puntos que tienen menos de **p** vecinos dentro de una distancia **d**

Detección de Outliers

KNN



Detección de Outliers

Métodos basados en distancia

Otra forma de detectar outliers es tomar la media de los datos y calcular la distancia a cada punto desde la media

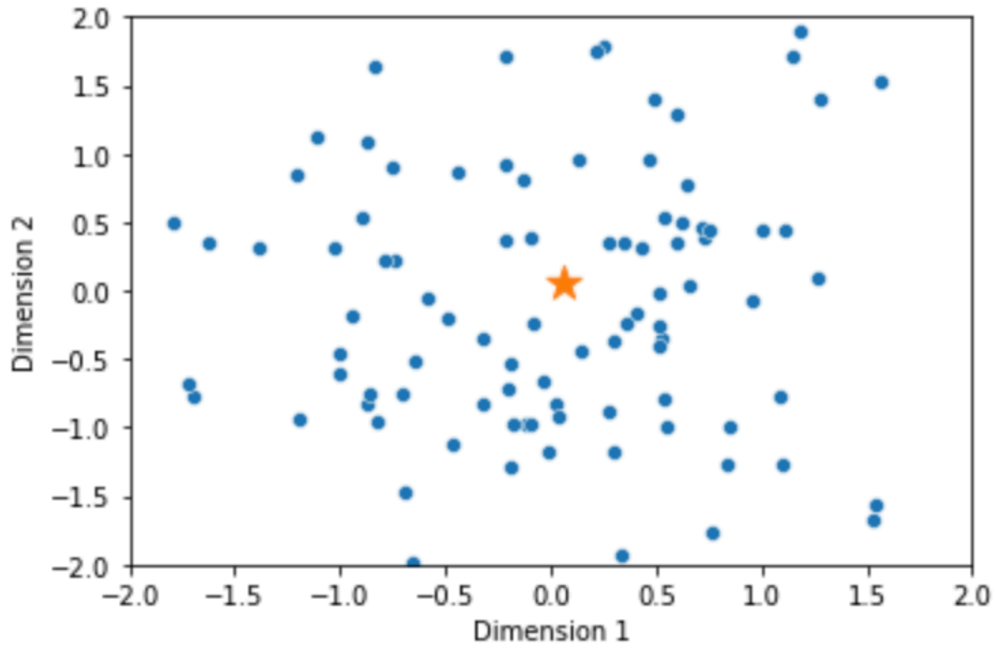
Dejamos como outliers los puntos más lejanos

Podemos hacer esto para cada *cluster*, en caso de tener más de uno

Detección de Outliers

Métodos basados en distancia

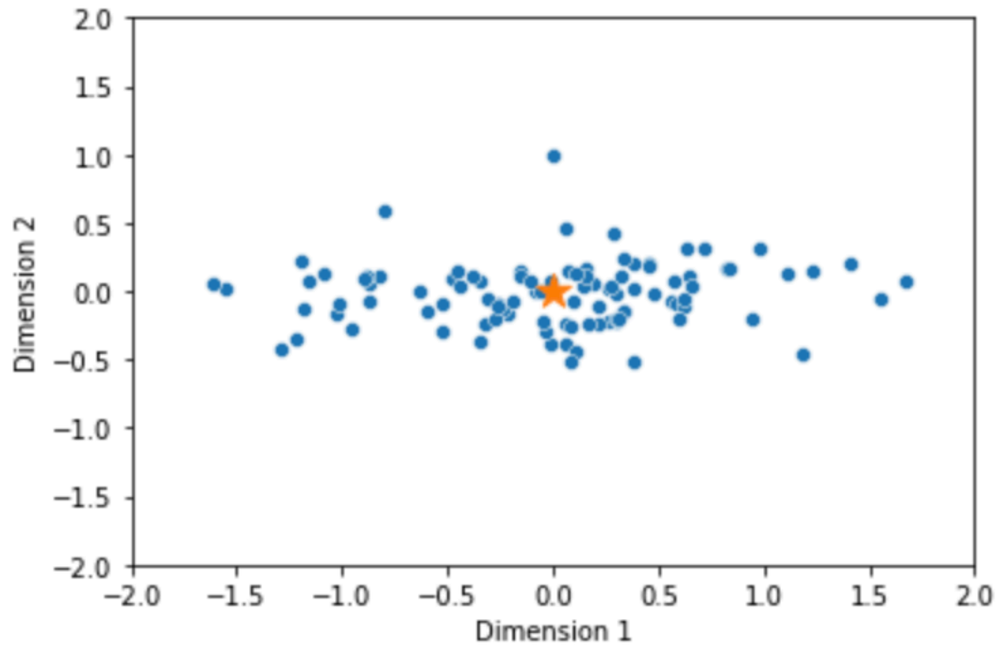
¿Cuáles son los puntos más lejanos a la media en este caso?



Detección de Outliers

Métodos basados en distancia

Ahora, considera este dataset



¿Es lo mismo desviarse de la media en una unidad en el eje **x** y en el eje **y**?

Detección de Outliers

Distancia de Mahalanobis

La distancia de Mahalanobis tiene en cuenta la "forma" de los datos

La distancia entre dos puntos que vienen del mismo *dataset* es:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \, S^{-1} \, (\vec{x} - \vec{y})}$$

Donde S^{-1} es la inversa de la matriz de covarianza

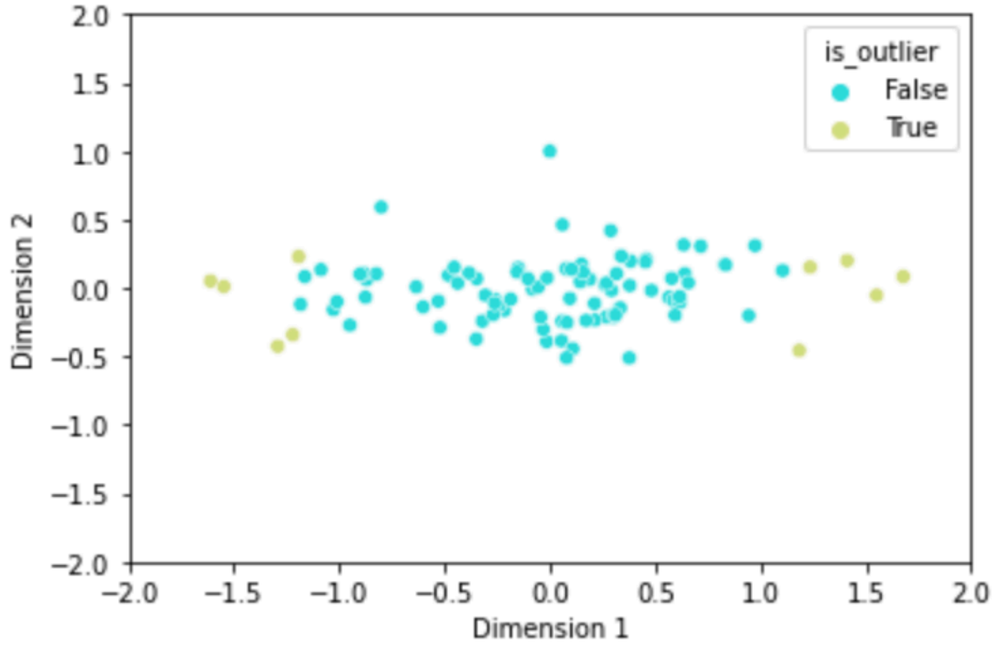
Detección de Outliers

Distancia de Mahalanobis

En la formula anterior, la matriz de covarianza nos ayuda entregar la información de la forma de la distribución de los datos

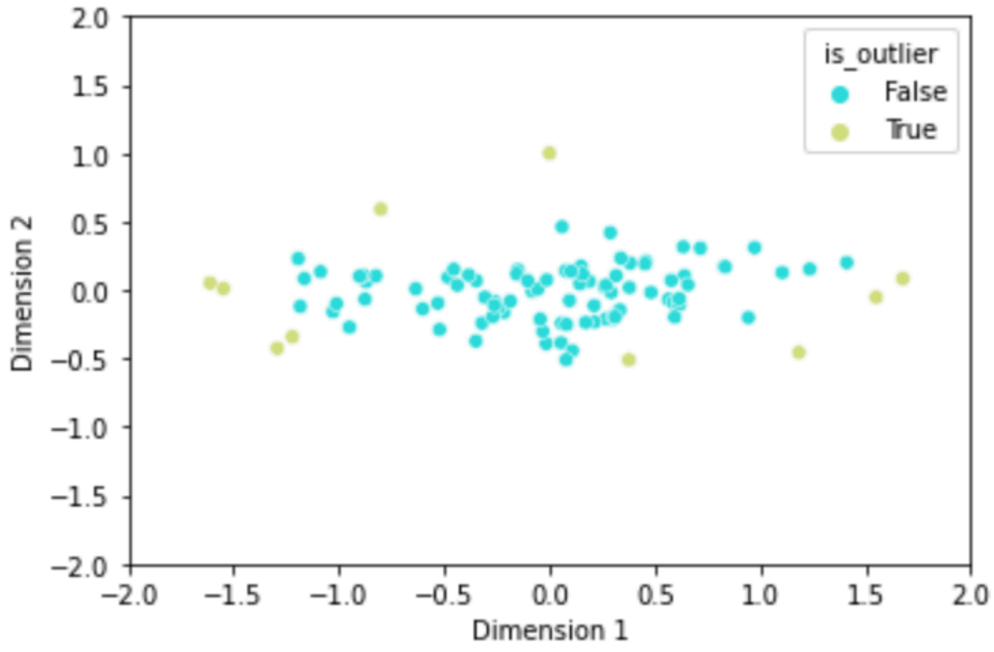
Detección de Outliers

Outliers con distancia euclidiana



Detección de Outliers

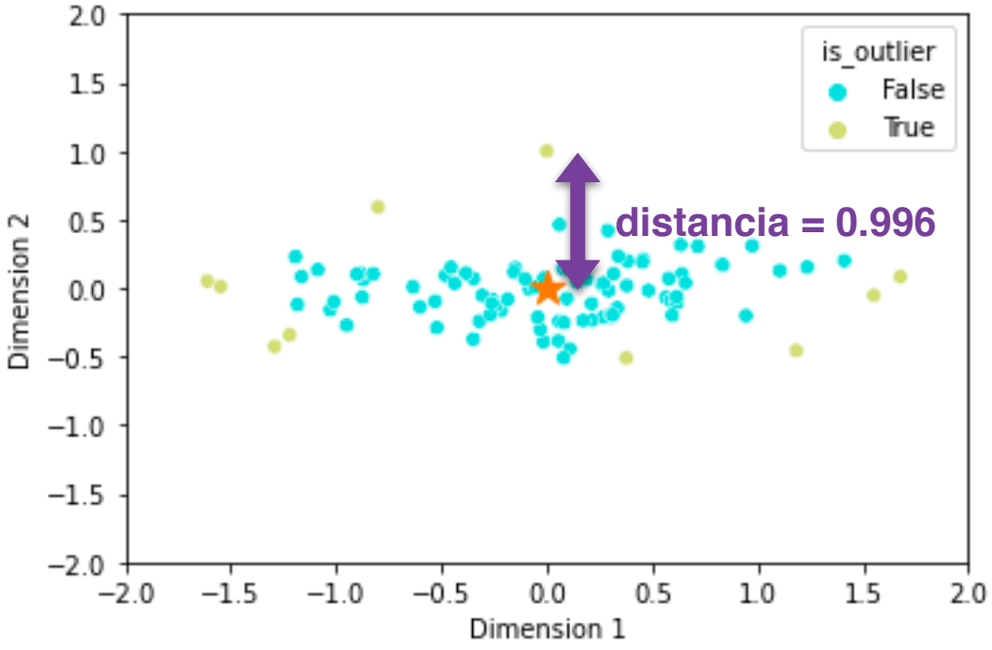
Outliers con distancia mahalanobis



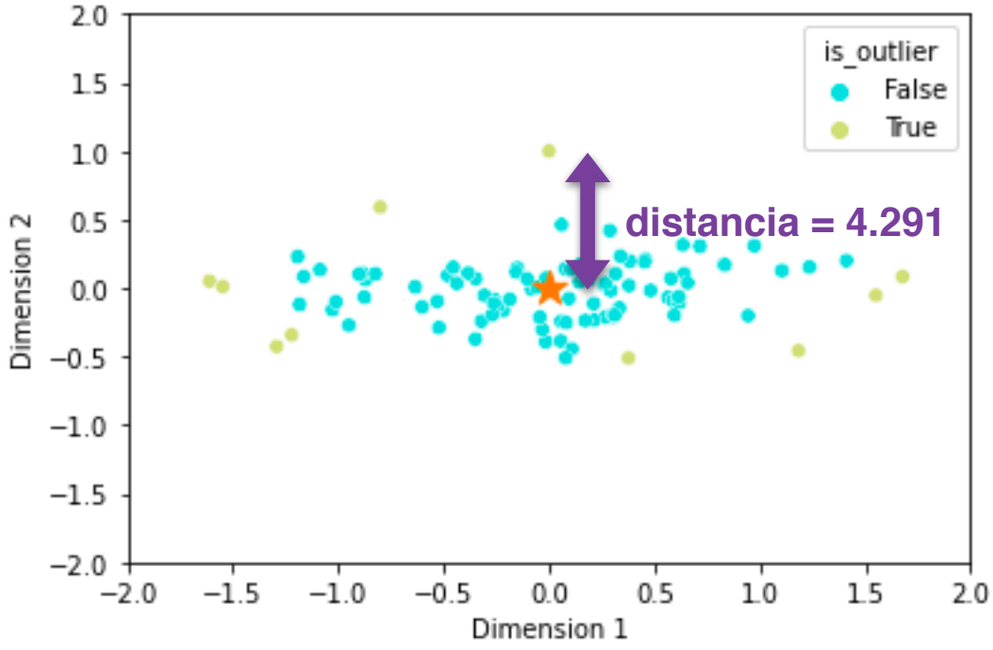
Detección de Outliers

Ejemplo - Diferencia entre distancias

Euclidiana



Mahalanobis



Fundamentos de Ciencias de Datos

Semana 14 - Outliers