DIPLOMADO
Ciencia de Datos

Módulo 7
Aprendizaje de máquina no supervisado
4. Detección de novedades y valores atípicos

Eduardo Espinosa Avila

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Detección de anomalías, definición

Detección de anomalías (*outliers*) es la identificación de eventos u observaciones que no son acordes a los patrones esperados en cierto conjunto de datos.

Se utiliza comúnmente para eliminar valores *anómalos* en conjuntos de datos, buscando mejorar el rendimiento de modelos de ML.

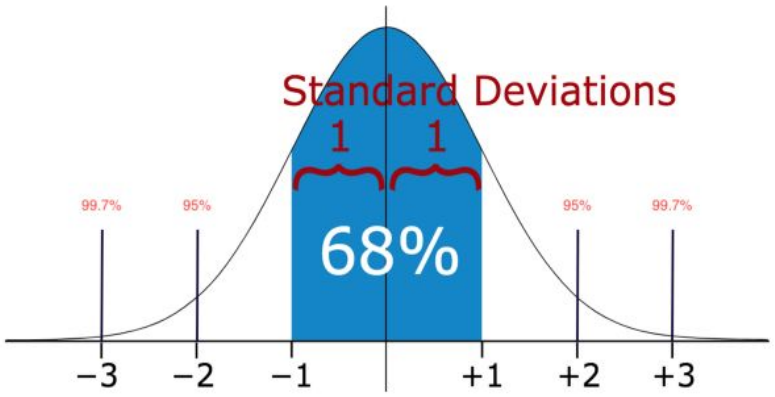
Detección de anomalías, desviación estándar

Es una forma simple de detectar valores anómalos en un conjunto de datos.

Si los datos tienen una distribución normal, entonces el 68% de los mismos se encuentran a una desviación estándar de la media.

El 95% estarán a dos desviaciones

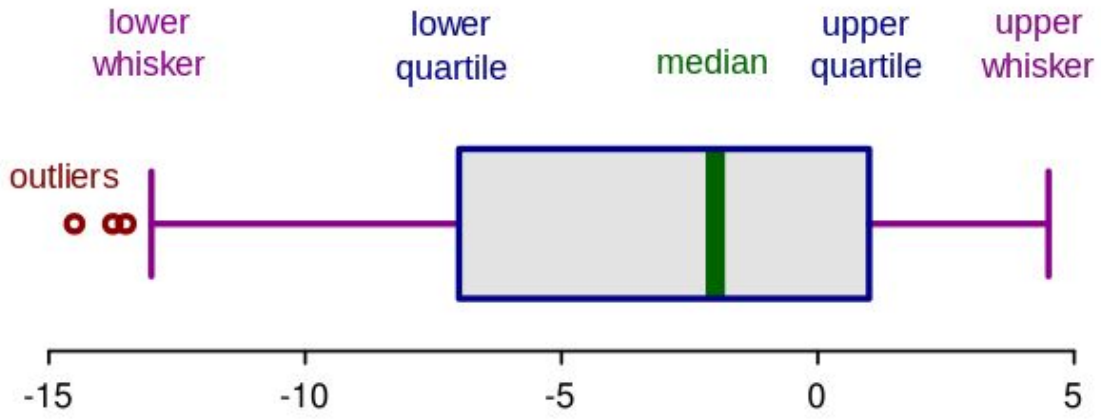
y 99.7%, a tres desviaciones



Detección de anomalías, *boxplot*

Un boxplot es una representación gráfica de datos números con sus cuartiles. Es una forma simple y muy efectiva de encontrar anomalías.

Los *bigotes* (*whiskers*) marcan los límites inferior y superior de la distribución de datos: cualquier dato fuera de esos límites se puede considerar anómalo.

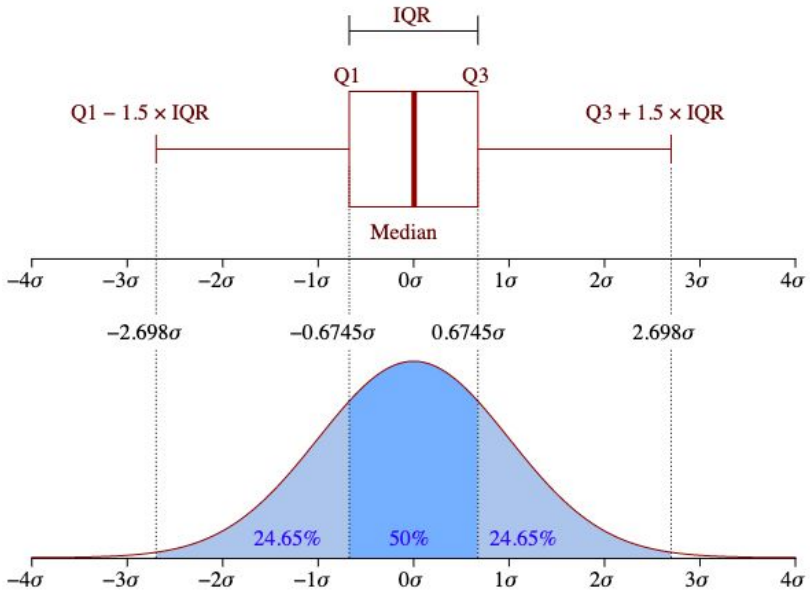


Detección de anomalías, *boxplot*

Para construirlo se utiliza el concepto de rango intercuartil (IQR), una medida de dispersión que divide un conjunto de datos en cuatro subgrupos.

Dados *n* datos:
$IQR = Q_3 - Q_1$
Q_1 = mediana de los *n*/2 datos menores
Q_3 = mediana de los *n*/2 datos mayores
Q_2 = mediana de todos los datos

$lw = Q_1 - 1.5*IQR$
$uw = Q_3 + 1.5*IQR$



De Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285

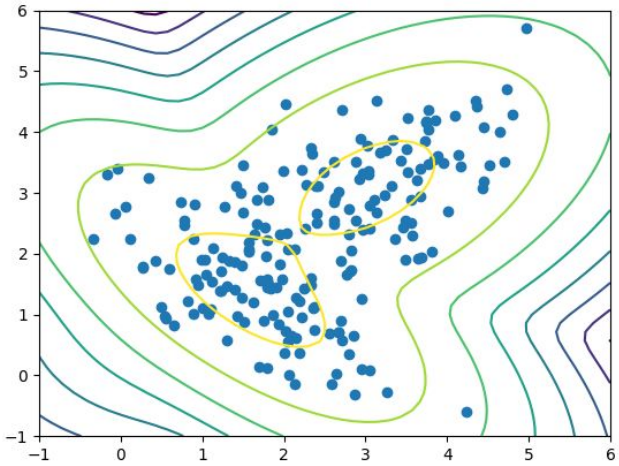
Detección de anomalías, agrupamientos

El análisis de agrupamientos puede usarse como método para detectar valores anómalos en conjuntos de datos.

La idea es simple:

Si un elemento tiene poca afinidad con todos los grupos, entonces es muy probable que se trate de una anomalía.

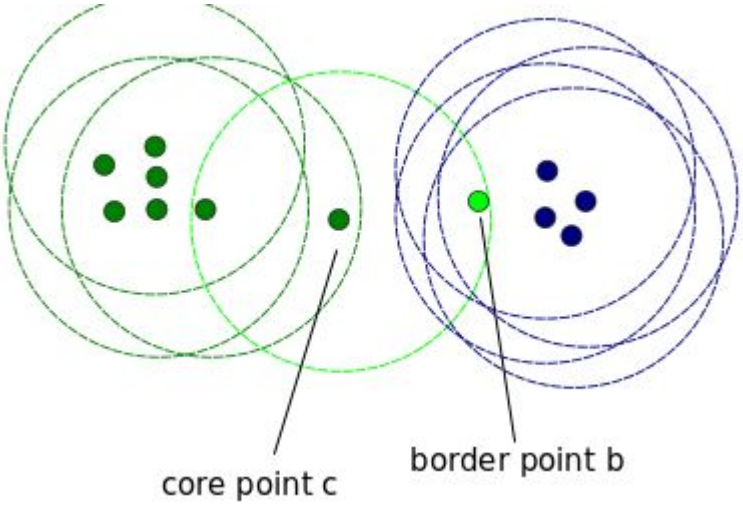
Se puede realizar con *k*-medias, pero es recomendable utilizar algo más sofisticado y que tome en cuenta la densidad, como un modelo de mezcla gaussiana o DBScan.



Detección de anomalías, DBScan

Define los grupos como regiones continuas de alta densidad:

- Para cada elemento, contar cuántos elementos se encuentran a una distancia pequeña *E* (vecindario-*E*)
- Si un elemento contiene más de cierto umbral en su vecindario-*E*, se considera un *core point*.
- Todos los elementos en el vecindario de un *core point*, pertenecen al mismo grupo. Puede incluir otros *core points*.



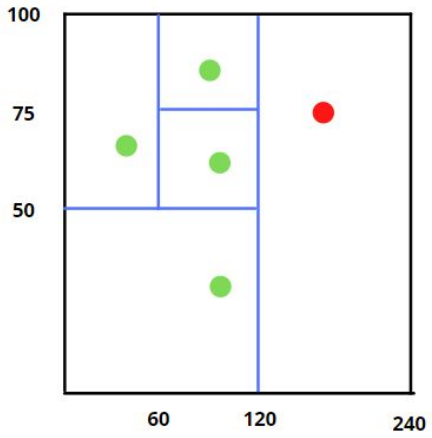
https://bit.ly/3wxcltR

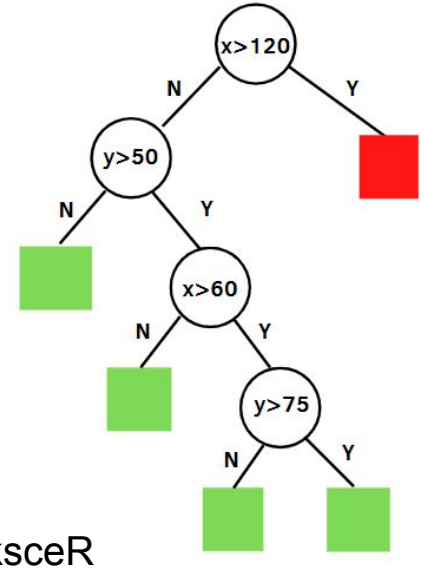
- Cualquier elemento que no pertenezca a algún grupo, se considera una anomalía.

Valores atípicos, bosque de aislamiento

Los métodos anteriores construyen perfiles de *elementos normales* y consideran anómalos aquellos que no se ajustan a la norma.

En cambio, este método aísla explícitamente las anomalías, aprovechando que son pocas y sus valores difieren mucho con respecto a los normales





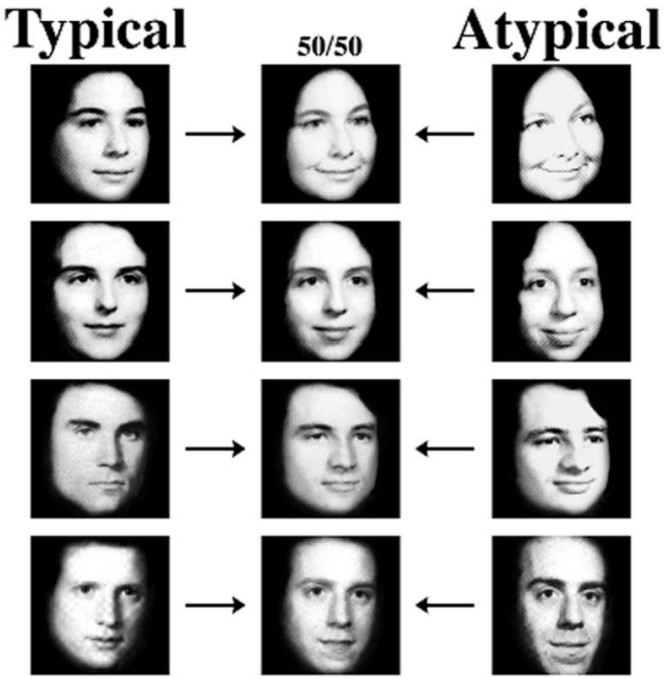
https://bit.ly/3fksceR

Valores atípicos, aplicaciones

Detectar anomalías es muy útil cuando se está realizando el preprocesamiento de datos. Puede ayudar a evitar sesgos no deseados.

Además, puede resultar útil en múltiples escenarios:

- Medicina
  - Monitoreo de salud en vuelos espaciales
- Seguridad en aeropuertos
- Detección de ataques en redes de computadoras
- Detección de fraudes
- Actividad inusual en redes móviles



https://bit.ly/3fqeG9O

Referencias

- Géron, A.
  Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow / Aurélien Géron -- USA : O'Reilly Media Inc., 2019 (484 páginas)

- Lishuai LI and R. John Hansman
  Anomaly Detection in Airline Routine Operations Using Flight Data Recorder Data
  https://core.ac.uk/download/pdf/16520235.pdf

- Jesús Burgueño, Isabel de-la-Bandera, Jessica Mendoza, David Palacios ,Cesar Morillas and Raquel Barco
  Online Anomaly Detection System for Mobile Networks
  https://www.mdpi.com/1424-8220/20/24/7232

- Luis Basora, Xavier Olive, Thomas Dubot
  Recent Advances in Anomaly Detection Methods Applied to Aviation
  https://hal.archives-ouvertes.fr/hal-02470453/document

- Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou
  Isolation Forest
  https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

https://twitter.com/laloea

https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/