Fundamentos de Aprendizaje Automático

Máster Universitario en Ciencia de Datos aplicada a las Ciencias Sociales
por la Universidad de Granada y la Universidad de Salamanca

Métodos no supervisados

02. Detección de Anomalías

UNIVERSIDAD
DE GRANADA

Índice

Sesión 3

- Detección de Anomalías
  - Definición
  - Tipos
  - Métodos basados en distancia
  - Métodos basados en densidad

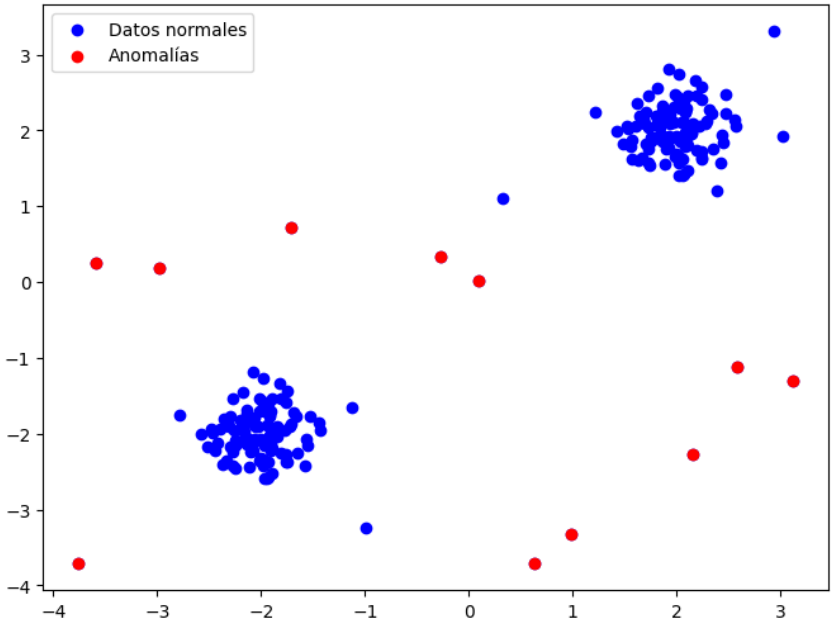
- Uso de algoritmos de detección de anomalías en Python

1. Detección de Anomalías

Definición

¿En qué consiste?

Trata de identificar patrones o comportamientos inusuales en un conjunto de datos



1. Detección de Anomalías

Definición

¿En qué consiste?

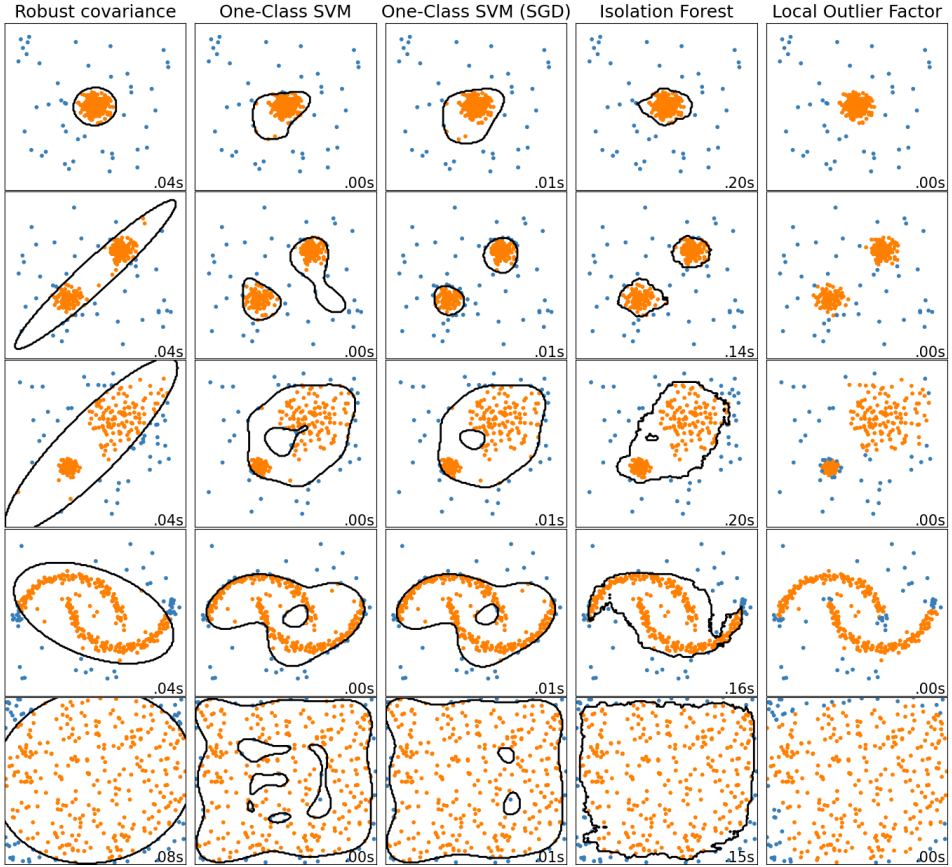
Trata de identificar patrones o comportamientos inusuales en un conjunto de datos

Hay técnicas supervisadas y otras no supervisadas

- La definición tanto de datos *"normales"* como de datos *anómalos* varía significativamente en función del contexto
- La presencia de anomalías en los datos **distorsionan** los análisis estadísticos al introducir patrones inexistentes, lo que lleva a conclusiones erróneas y predicciones poco fiables.

1. Detección de Anomalías

Tipos



https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html#sphx-glr-auto-examples-miscellaneous-plot-anomaly-comparison-py

1. Detección de Anomalías

Tipos

- Métodos estadísticos
  - Se basan en la desviación estándar de los datos o en sus distribuciones de probabilidad.
- Métodos basados en distancia
  - Utilizan distancia para ver los puntos más alejados de (sus k vecinos más cercanos) o de cualquier clúster.
- Métodos basados en densidad
  - Se analiza la densidad “local” de los puntos: Los puntos que no pertenecen a ninguna región densa son considerados anomalías
- Métodos basados en modelos
  - Utilizan métodos supervisados para detectar los puntos anómalos.
- Métodos basados en árboles
  - Los puntos que requieren menos divisiones para ser aislados en el árbol se consideran anómalos

1. Detección de Anomalías

Métodos basados en distancia

Se basan en calcular las **distancias** entre puntos de datos. Estos métodos asumen que las anomalías estarán a una distancia significativa de sus vecinos en el espacio de características.

- **k-Nearest Neighbors (k-NN):** Calcula la distancia entre cada punto de datos y sus *k* vecinos más cercanos. Los puntos con distancias mayores a un umbral son considerados anómalos.

- **Clustering** (como K-means): Los puntos que están lejos de cualquier centro de clúster pueden ser marcados como anomalías.

1. Detección de Anomalías

k-NN

El algoritmo **k-NN** es un método de aprendizaje supervisado que puede adaptarse para la detección de anomalías en un entorno **no supervisado**. La idea principal es medir la distancia entre cada punto de datos y sus *k* vecinos más cercanos.

Procedimiento:

- Para cada punto, calcular la distancia a sus *k* vecinos más cercanos.
- Obtener la distancia promedio del *k*-ésimo vecino más cercano.
- **Definir un umbral:** los puntos cuya distancia promedio sea mayor que este umbral se consideran anomalías.

1. Detección de Anomalías

k-NN

Ventajas:

- Sencillo de implementar.
- No requiere suposiciones sobre la distribución de los datos.

Desventajas:

- El rendimiento puede verse afectado por la alta dimensionalidad de los datos.
- La elección del valor de *k* puede influir en los resultados.

1. Detección de Anomalías

Clustering (K-Means)

Los puntos que están lejos de cualquier centro de clúster pueden considerarse anomalías.

Procedimiento:

- Aplicar el algoritmo de clustering, por ejemplo K-Means, para agrupar los datos en *k* clústeres.
- Calcular la distancia de cada punto de datos a su centroide de clúster más cercano.
- **Definir un umbral:** los puntos cuya distancia al centroide sea mayor que este umbral se consideran anomalías.

1. Detección de Anomalías

Clustering (K-Means)

Ventajas:

- Fácil de entender e implementar.
- Escalable a grandes conjuntos de datos.

Desventajas:

- Sensible a la elección de *k* y a la inicialización de los centroides.
- No siempre captura formas de clústeres no esféricas.

1. Detección de Anomalías

Métodos basados en densidad

Se basan en la **densidad local** de los puntos. Estos métodos asumen que las anomalías residen en regiones de menor densidad en comparación con la mayoría de los datos.

- **Usando DBSCAN:** Se Agrupan los puntos en clústeres basados en la densidad y considera como anomalías los puntos que no pertenecen a ningún clúster.

- **Usando OPTICS:** similar a DBSCAN, pero puede encontrar clústeres en diferentes niveles de densidad.

- **LOF (Local Outlier Factor):** Mide la densidad local de un punto de datos en relación con sus vecinos, y los puntos con densidad significativamente menor que la de sus vecinos se consideran anomalías.

1. Detección de Anomalías

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Agrupa puntos de datos en clústeres basados en la densidad y considera como anomalías los puntos que no pertenecen a ningún clúster.

Procedimiento:

- Definir dos parámetros clave: *eps* (radio máximo de un punto para ser considerado vecino) y *minPts* (número mínimo de puntos necesarios para formar un clúster).
- Para cada punto de datos, encontrar los puntos vecinos dentro de eps.
- Si el número de vecinos es ≥ minPts, formar un clúster.
- Los puntos que no se pueden asignar a ningún clúster se consideran anomalías.

1. Detección de Anomalías

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Ventajas:

- No requiere especificar el número de clústeres a priori.
- Puede encontrar clústeres de forma arbitraria y detectar anomalías de manera efectiva.

Desventajas:

- Sensible a la elección de *eps* y *minPts*.
- Puede tener dificultades con datos de alta dimensionalidad.

1. Detección de Anomalías

OPTICS (Ordering Points To Identify the Clustering Structure)

Similar a DBSCAN, pero puede encontrar clústeres en diferentes niveles de densidad.

Procedimiento:

- Ordenar los puntos de datos en función de su accesibilidad (similar al parámetro *eps* en DBSCAN).
- Generar un gráfico de accesibilidad para identificar clústeres en diferentes niveles de densidad.
- Los puntos que no forman parte de ningún clúster se consideran anomalías.

1. Detección de Anomalías

OPTICS (Ordering Points To Identify the Clustering Structure)

Ventajas:

- Puede encontrar clústeres en diferentes niveles de densidad.
- No requiere especificar un valor global para *eps*.

Desventajas:

- Puede ser más complejo de implementar y comprender en comparación con DBSCAN.
- Puede tener dificultades con datos de alta dimensionalidad.

1. Detección de Anomalías

LOF (Local Outlier Factor)

Mide la densidad local de un punto de datos en relación con sus vecinos, y los puntos con densidad significativamente menor que la de sus vecinos se consideran anomalías.

Procedimiento:

- Para cada punto de datos, calcular la densidad local basada en la distancia a sus *k* vecinos más cercanos.
- Comparar la densidad local de cada punto con la densidad local de sus vecinos.
- **Asignar un factor LOF a cada punto:** los puntos con un valor de LOF significativamente mayor que 1 se consideran anomalías.

1. Detección de Anomalías

LOF (Local Outlier Factor)

Ventajas:

- Puede detectar anomalías en datos con diferentes densidades.
- No requiere la definición de un umbral fijo.

Desventajas:

- La elección del número de vecinos (k) puede influir en los resultados.
- Puede ser computacionalmente costoso en grandes conjuntos de datos.

Índice

Sesión 3

- Detección de Anomalías
  - Definición
  - Tipos
  - Métodos basados en distancia
  - Métodos basados en densidad

- Uso de algoritmos de detección de anomalías en Python