

Unidad 4: Aprendizaje no supervisado y detección de anomalías

Adarsh KUMAR

Universitat Politècnica de Catalunya
Department of Computer Science

Coordinador del proyecto:
Prof. Ilker Demirkol

Proyecto MERiT
3 de septiembre de 2025

Co-funded by the European Union

[Diapositiva 2: Índice]

Aprendizaje no supervisado | Detección de anomalías | Casos de uso

Índice

1. Aprendizaje no supervisado

2. Detección de anomalías

3. Casos de uso

[Diapositiva 3: Aprendizaje no supervisado frente a detección de anomalías]

Aprendizaje no supervisado frente a detección de anomalías

Aprendizaje no supervisado:

- Aprende patrones a partir de datos sin etiquetar
- No hay salidas predefinidas (a diferencia del aprendizaje supervisado)

Detección de anomalías:

- Identifica observaciones raras o inusuales
- Aplicaciones: detección de fraude, monitorización del estado de los sistemas

---

[Diapositiva 4: Ejemplo numérico: conjunto de datos sintético]

Ejemplo numérico: conjunto de datos sintético

Puntos de datos en 2D (características x1 y x2):

- (1.0, 1.1)
- (1.1, 1.0)
- (1.2, 1.2)
- (1.0, 1.2)
- (5.0, 5.5)
- (1.1, 1.1)
- (5.2, 5.0)
- (1.2, 1.0)

Gráfico: Clúster (azul) frente a anomalías (rojo)

Co-funded by the European Union

Observación

- La mayoría de los puntos se agrupan cerca de (1.0, 1.0)
- (5.0, 5.5) y (5.2, 5.0) están lejos del clúster ⇒ anomalías

---

[Diapositiva 5: Detección de anomalías: visualización y métodos]

Detección de anomalías: visualización y métodos

Gráfico: Clúster (azul) frente a anomalías (rojo)

Cómo detectan las anomalías los algoritmos

- **K-Means:**
  - Mide la distancia al centroide más cercano
  - Señala los puntos que superan un umbral
- **Isolation Forest:**
  - Necesita menos divisiones para aislar las anomalías
  - Usa la longitud del camino como puntuación de anomalía
- **DBSCAN:**
  - Identifica regiones de baja densidad
  - Puntos núcleo frente a puntos frontera

[Diapositiva 6: Conceptos clave]

Conceptos clave

Clustering frente a detección de anomalías:

- Clustering: agrupa datos similares (p. ej., K-Means).
- Detección de anomalías: señala los valores atípicos (p. ej., Isolation Forest).

Técnicas comunes:

- Basadas en densidad (DBSCAN, Local Outlier Factor).
- Basadas en reconstrucción (Autoencoders).

[Diapositiva 7: Ejemplo: clustering K-Means]

Ejemplo: clustering K-Means

Puntos de datos en 2D (características x1, x2):

- (1.1, 1.2)
- (1.3, 1.1)
- (1.2, 1.3)
- (5.1, 5.2)
- (5.3, 5.0)
- (5.2, 5.3)
- (3.0, 3.0)

Pasos de K-Means:

1. Inicializar los centroides (k=2)

2. Asignar cada punto al centroide más cercano

3. Recalcular los centroides

4. Repetir hasta la convergencia

Gráfico: Clustering K-Means (k=2) con un valor atípico

[Diapositiva 8: Clustering K-Means: cálculo matemático (parte 1)]

Clustering K-Means: cálculo matemático (parte 1)

Datos de partida (k=2):

X = (1.1, 1.2), (1.3, 1.1), (1.2, 1.3), (5.1, 5.2), (5.3, 5.0), (5.2, 5.3), (3.0, 3.0)

Paso 1: Inicializar los centroides

Se eligen al azar los centroides iniciales:

$\mu 1 = (1.1, 1.2), \quad \mu 2 = (5.1, 5.2)$

Paso 2: Asignar los puntos (distancia euclídea)

$d(xi, \mu 1) = \sqrt{((xi1 - 1.1)^2 + (xi2 - 1.2)^2)}$

$d(xi, \mu 2) = \sqrt{((xi1 - 5.1)^2 + (xi2 - 5.2)^2)}$

Resultados de la asignación:

- Clúster 1: puntos 1, 2, 3 (distancia < 0.22 a $\mu 1$)
- Clúster 2: puntos 4, 5, 6 (distancia < 0.22 a $\mu 2$)
- Valor atípico (3.0, 3.0): $d(\mu 1) = 2.63$, $d(\mu 2) = 3.28 \Rightarrow$ Clúster 1

Co-funded by the European Union

[Diapositiva 9: Clustering K-Means: cálculo matemático (parte 2)]

Clustering K-Means: cálculo matemático (parte 2)

Paso 3: Recalcular los centroides

$\mu 1$ nuevo = ((1.1 + 1.3 + 1.2 + 3.0) / 4, (1.2 + 1.1 + 1.3 + 3.0) / 4) = (1.65, 1.65)

$\mu 2$ nuevo = ((5.1 + 5.3 + 5.2) / 3, (5.2 + 5.0 + 5.3) / 3) = (5.2, 5.17)

Paso 4: Reasignar los puntos

Nuevas distancias para el valor atípico (3.0, 3.0):

$d(\mu 1 \text{ nuevo}) = \sqrt{((3.0 - 1.65)^2 + (3.0 - 1.65)^2)} = 1.91$

$d(\mu 2 \text{ nuevo}) = \sqrt{((3.0 - 5.2)^2 + (3.0 - 5.17)^2)} = 3.11$

Sigue asignado al Clúster 1.

Comprobación de convergencia:

- Centroides anteriores: $\mu 1 = (1.1, 1.2)$, $\mu 2 = (5.1, 5.2)$
- Nuevos centroides: $\mu 1$ nuevo = (1.65, 1.65), $\mu 2$ nuevo = (5.2, 5.17)
- Movimiento significativo ⇒ Continuar las iteraciones

[Diapositiva 10: Ejemplo: Isolation Forest]

Ejemplo: Isolation Forest

Cómo funciona:

- Divide los datos de forma aleatoria para aislar las anomalías más rápido.

[Diapositiva 11: Caso práctico: detección de fraude con tarjetas de crédito]

Caso práctico: detección de fraude con tarjetas de crédito

Problema: detectar transacciones fraudulentas entre millones.

Solución: Autoencoders (el error de reconstrucción como puntuación de anomalía).

[Diapositiva 12: Caso práctico: detección de intrusiones en la red]

Caso práctico: detección de intrusiones en la red

Problema: identificar ciberataques en los registros de los servidores.

Co-funded by the European Union

Solución: DBSCAN (valores atípicos basados en densidad).

Resultado

Detectó el 95 % de los ataques con pocos falsos positivos.

[Diapositiva 13: Desafíos]

Desafíos

- **Datos desbalanceados:** las anomalías son raras.
- **Ajuste del umbral:** decidir qué se considera «anómalo».
- **Alta dimensionalidad:** la maldición de la dimensionalidad.

[Diapositiva 14: Herramientas del mundo real]

Herramientas del mundo real

- **Scikit-learn:** Isolation Forest, DBSCAN.
- **TensorFlow:** Autoencoders.
- **PyOD:** biblioteca de detección de valores atípicos en Python (Python Outlier Detection).

[Diapositiva 15: Tendencias futuras]

Tendencias futuras

- Aprendizaje profundo para la detección de anomalías (p. ej., GAN).
- IA explicable (Explainable AI) para facilitar la interpretabilidad.
- Computación en el borde (edge computing) para la detección en tiempo real.

Co-funded by
the European Union