

# Introducción a los métodos de aprendizaje automático



UT02 – Trabajo de Aplicación 3

1

---

---

---

---

---

---

---

---

## Outliers



- En general, son muestras que están excepcionalmente lejos del conjunto principal de los datos
- A menudo podemos identificarlos observando los gráficos
- Cuando se sospecha que un valor es un outlier,
  - verificar que los valores sean científica y físicamente válidos
  - verificar que no se tengan errores en la captura de los datos
- Cuidado al retirar valores!
- también pueden ser información específica sobre el tópico (ej. detección de anomalías)

2

2

---

---

---

---

---

---

---

---

## Outliers (2)



- Errores de datos
  - errores de medida, errores humanos o de recolección
  - a menudo son ignorados
- Varianza normal
  - en una distribución normal, 99.7% de los puntos de datos están dentro de un rango de +/- 3 desvíos estándar con respecto a la media
  - ej: 1 persona que gana más de 1 billon de USD por año, o alguien que tiene una altura mayor a 2 metros...
  - estos outliers producen un sesgo en las estadísticas descriptivas (ej: media)
  - sin embargo... pueden ser legítimos!
- Datos con otras distribuciones
- Suposiciones sobre las distribuciones

3

3

---

---

---

---

---

---

---

---

TA 3 - Ejercicio 1



- En Excel, carga los datos del archivo “telephone.csv”
- Grafica los datos.
- Intuitivamente, ¿qué situación se observa?

4

4

---

---

---

---

---

---

---

TA 3 - Ejercicio 1



- Busca información sobre este problema
  - International Telephone Calls - Belgium 1950 1973 (Data Mining - Ian Witten)
- Si esos datos son anómalos, ¿qué podemos hacer con ellos?

5

5

---

---

---

---

---

---

---

Bloques de RM para detección y gestión de outliers



- Detect Outlier (Distances)
- Detect Outlier (LOF)
- Detect Outlier (Densities)
- Detect Outlier (COF)

6

6

---

---

---

---

---

---

---

TA3 - Ejercicio 2



- Para cada bloque de detección de outliers de RM, generar un resumen:
- breve descripción
  - parámetros que acepta
  - característica de aplicación más importante

---

---

---

---

---

---

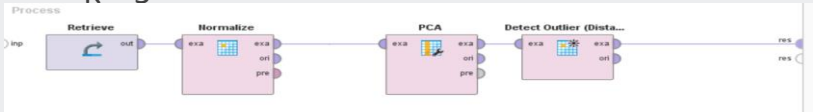
---

TA3 - Ejercicio 3 a)



Cargar el dataset “Iris”

- Analizar atributos (tipos de datos, rangos, distribuciones, etc.)
- Visualizar scatter plots
- Normalizar
- Aplicar modelo: PCA
- Agregar bloque de detección de outliers por distancia
  - K=1, outliers = 10 , distancia euclideana
  - K = 5



---

---

---

---

---

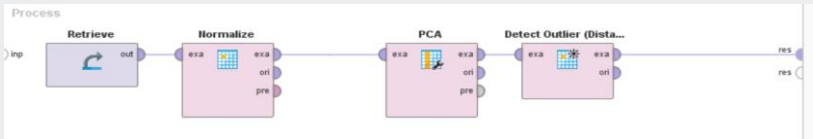
---

---

TA3 - Ejercicio 3 b)



- Generar otro canal en el modelo y utilizar, para los mismos datos normalizados y PCA aplicado, un operador de Detección de Outliers por Densidad
- Analizar los parámetros y probar con distintos valores
- Comparar con los resultados obtenidos en la parte a)



---

---

---

---

---

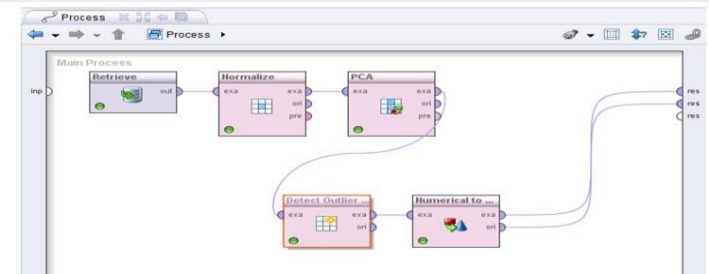
---

---

TA3 - Ejercicio 3 c)



- ¿Qué es el LOF (Local Outlier Factor) y para qué sirve?
- Generar otro canal utilizando LOF, Analizar los parámetros y probar con distintos valores
- Comparar con los resultados obtenidos en las partes anteriores



10

10

---

---

---

---

---

---

---

---