Introducción a los métodos de aprendizaje automático

Universidad Católica del Uruguay

UT02 – Trabajo de Aplicación 3

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)

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

TA 3 - Ejercicio 1

- En Excel, carga los datos del archivo “telephone.csv”

- Grafica los datos.

- Intuitivamente, ¿qué situación se observa?

TA 3 - Ejercicio 1

Universidad Católica del Uruguay

- 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?

Bloques de RM para detección y gestión de outliers

Universidad Católica del Uruguay

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

TA3 - Ejercicio 2

Universidad Católica del Uruguay

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

Universidad Católica del Uruguay

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