Almacenes y Minería de Datos

Clasificación y Reglas de Asociación

Joaquín Ángel Triñanes Fernández
Instituto de Investigacións Tecnolóxicas
Joaquin.Trinanes@usc.es Ext: 16001 Externo: 881816001

Clasificación/Regresión

- Dentro de las técnicas de minería de datos/aprendizaje automático supervisado encontramos la clasificación y la regresión
- •La regresión permite predecir un valor determinado dentro de un conjunto contínuo o muy extenso.
- •La clasificación predice su pertenencia a una clase discreta o categórica.
- Múltiples métricas para evaluar los modelos: matriz de confusión, precisión, F1-score, recall, exactitud, AUC, MSE, MAE, etc.
- Múltiples algoritmos: Regresión logística, K-NN, Random Forests, SVM, ANN,…
- Optimización hiperparámetros: Grid search, Random Search

Reglas de asociación

- Buscamos encontrar asociaciones o estructuras causales entre conjuntos de elementos de transacciones.
- Análisis de la cesta de la compra: ¿qué productos compramos juntos?
- Diagnósticos médicos: ¿qué enfermedades están asociadas?
- Navegación Web: ¿Qué página es más probable visitar dado nuestro historial?
- Concentración de actividad económica: ¿Qué firmas suelen coincidir en polígonos industriales/centros comerciales/…?
- No es lo mismo que el filtrado colaborativo: ¿Qué productos compran los usuarios parecidos a ti? Usando las reglas de asociación, a 2 usuarios que han comprado los mismos productos, se les recomendará los mismo, independientemente de su historial de compras.

Reglas de asociación

- Algoritmos: ECLAT, FP Growth, SETM, **Apriori**, CARMA, DMA, etc.
- Indicadores: **lift, soporte y confianza**.
  - Soporte(A=>B)= $P(A\ U\ B)$
  - Confianza(A=>B)= $P(A\ U\ B)/P(A)$
  - Lift(A=>B)=$P(A\ U\ B)/P(A)P(B)$
    - ¿Existe una asociación entre los productos? ¿Esa asociación es positiva o negativa?

Reglas de asociación

- Algoritmos: ECLAT, FP Growth, SETM, **Apriori**, CARMA, DMA, etc.
- Indicadores: **lift, soporte y confianza**.
  - Lift(A=>B)=$P(A\ U\ B)/P(A)P(B)$
    - ¿Existe una asociación entre los productos? ¿Esa asociación es positiva o negativa?
  - Soporte(A=>B)= $P(A\ U\ B)$
  - Confianza(A=>B)= $P(A\ U\ B)/P(A)$
- ¿Qué buscamos?

  Lift >1 o lift <1, valores cercanos a 1 indica falta de asociación

  Soporte: suele ser un valor bajo dependiendo del conjunto de datos

  Confianza: mejor cuanto más alta, pero también depende del conjunto de datos

Ejercicio#8

1. Clasificación:

https://www.dropbox.com/s/juyw9312251hrbs/clasificar_train.csv?dl=0
https://www.dropbox.com/s/kinkmcscflkfsc1/clasificar_test.csv?dl=0

Empleando como variables independientes los valores de las bandas, cúal es el mejor resultado que obtenéis para el conjunto de test?

2. Regresión. Datos:

https://www.dropbox.com/s/b5swwqmpim5k4bu/AMD_regresion.csv.gz?dl=0

La columna que queremos predecir es fC.

2.1. Aplicar Random Forest o una variante

2.2 Entrenad una red neuronal y haced predicciones sobre el conjunto de entrenamiento.

3. Extraer reglas de asociación sobre este conjunto de datos:

https://www.dropbox.com/s/7rfh9hgvwj0nok0/cesta_compra2.csv?dl=0

(Paquetes de python: mlxtend, apyori, apriori_python, etc.)

Se valorará la claridad de la descripción y código, así como el análisis de resultados. Esta práctica será realizada en equipos de 3 componentes y consistirán en un documento PDF con el código y la descripción de la práctica.