UNIVERSIDAD DE GRANADA

/ UGR / decsai

Departamento de Ciencias de la Computación e Inteligencia Artificial

Sistemas Inteligentes para la Gestión en la Empresa

Práctica 1: Pre-procesamiento de datos y clasificación binaria

Curso 2017-2018

Objetivos y evaluación

En esta primera práctica de la asignatura Sistemas Inteligentes para la Gestión en la Empresa estudiaremos cómo realizar diversas tareas de pre-procesamiento de datos, como paso previo e imprescindible para el aprendizaje automático.

La práctica consistirá en la resolución de un problema de pre-procesamiento y aprendizaje automático. **Junto a la solución del problema (código R), se entregará una memoria explicativa de las tareas realizadas.**

La práctica se desarrollará de forma individual. La calificación constituirá el 15% de la nota final de la asignatura (1.5 puntos). Se evaluará, en este orden: (1) la calidad de la memoria presentada; (2) la precisión y la exactitud obtenida por el clasificador. Se valorará especialmente la claridad en la redacción y en la presentación.

La entrega se realizará a través de la plataforma docente de DECSAI, en el enlace que se habilitará al efecto.

Descripción del problema

Se trabajará sobre el conjunto de datos de préstamos proporcionado en la web *Lending Club* (https://www.lendingclub.com/info/download-data.action) denominado *LoanStats_2017Q4.csv*. La descripción de las variables de este conjunto de datos se encuentra en la sección DATA DICTIONARY de la misma web.

El problema consiste en predecir el estado de un préstamo (*loan_status*) a partir del resto de variables. Trataremos el conjunto de datos como un problema de clasificación binaria, con dos posibles salidas: {Pago, Impago}.

- Impago:
  - Late (16-30 days)
  - Late (31-120 days)
  - In Grace Period
  - Charged Off
- Pago:
  - Current
- No relevantes:
  - Otros valores

La memoria explicará qué **tareas de pre-procesamiento** se han llevado a cabo y con qué objetivo, así como los resultados obtenidos. Se enumeran a continuación de forma no exhaustiva algunas de las tareas que pueden realizarse:

- Eliminación de instancias pertenecientes a clases no relevantes
- Transformación y limpieza de valores de texto y numéricos
- Eliminación de variables sin información
    - Variables con todos los valores perdidos
    - Variables con 'mucha' diversidad en sus valores
    - Variables con 'poca' diversidad en sus valores
- Detección de conflictos e inconsistencias en los datos
    - Identificación y tratamiento de 'outliers'
    - Identificación e imputación de valores perdidos
    - Tratamiento del ruido
- Normalización y discretización
- Reducción y ampliación de datos
    - Selección de características
    - Selección de ejemplos
    - Tratamiento de clases no balanceadas
- Visualización de datos

La memoria explicará qué **técnicas de clasificación se han utilizado**. Al menos se utilizarán dos técnicas diferentes, cuya selección deberá justificarse. También se detallará el proceso de generación de los conjuntos de entrenamiento, validación y test. Se analizarán los resultados obtenidos con las técnicas utilizadas, haciendo especial énfasis en las medidas de precisión y exhaustividad. También se discutirá el impacto del pre-procesamiento en los resultados de clasificación.

Se recomienda apoyar las explicaciones con gráficos, diagramas, etc.

Entrega

Límite: 23 de abril de 2017

Contenidos: Un fichero .zip, incluyendo:

- Código en R (.R, .Rmd)
- Memoria
    - Portada: nombre, título
    - Índice
    - Contenidos
        - Exploración
        - Pre-procesamiento
        - Clasificación
        - Discusión de resultados
        - Conclusiones
    - Bibliografía