Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO III: Minería de datos

PARTE 06: Criterios de Selección de Modelos

Introducción:

En este trabajo se abordarán diferentes métricas de evaluación utilizadas para la selección de modelos de minería de datos ante un problema determinado.

En primera instancia, se utilizará software, como **R** y **Weka**, con el objetivo de aplicar un algoritmo de aprendizaje automático a un dataset, para luego calcular métricas de evaluación a la clasificación/predicción resultante a efectos de poder concluir acerca de la efectividad del algoritmo utilizado con relación a ese dataset.

Consignas:

1. **Medidas de evaluación para técnicas de clasificación.** A partir del archivo *asado.csv*, genere dos modelos utilizando los algoritmos de minería de datos vistos en clase para clasificación. Luego, en función de las clasificaciones realizadas, complete las siguientes actividades:
    a. **Exploración de la clasificación.**
        1. Explore y documente las clasificaciones realizadas por ambos modelos.
        2. ¿Cuáles instancias fueron clasificadas de manera incorrecta en cada uno? ¿A qué se debe?
    b. **Curva ROC.**
        1. Genere la matriz de confusión de las clasificaciones y grafique la curva ROC.
        2. ¿Qué información brinda la técnica del punto anterior?
        3. ¿Qué permite concluir con respecto a la clasificaciones de ambos modelos?

c. **Accuracy.**
    1. Ahora, calcule el accuracy de ambos modelos.
    2. ¿Cómo se interpreta la métrica anterior?
    3. ¿Qué aporta el accuracy?

d. **Recall/Precision.**
    1. Calcule las métricas recall y precisión para ambos modelos.
    2. ¿Cuál es la diferencia entre ambas?
    3. ¿Qué aspectos aborda cada una?

e. **F-score.**
    1. Calcule la métrica F-score en función de la matriz de confusión resultante.
    2. ¿Qué haría en caso de querer dar mayor importancia a la precisión del modelo? ¿Y en caso de querer ponderar la exhaustividad?
    3. Compare y documente los resultados.

f. **Kappa.** Calcule el Coeficiente de Kappa. ¿En qué casos resulta conveniente?

2. **Medidas de evaluación para técnicas de regresión.** A partir del archivo *llamadas.csv*, aplique el algoritmo de regresión lineal simple estudiado en clase. Luego, en función de la clasificación realizada por el algoritmo, realice las siguientes actividades:

    a. **Exploración de la clasificación.** Explore y documente la fórmula de regresión realizada por el algoritmo.

    b. **MSE** y **RMSE.**
        1. Calcule el MSE y la raíz del error cuadrático de la media (RMSE) de esta clasificación.
        2. ¿Cómo se interpretan estas medidas?

3. ¿En qué casos es útil? ¿Cuál es la diferencia entre ambas?

c. **AIC.**

1. Calcule el criterio de información de Akaike e interprete el resultado.
2. Calcule nuevamente el criterio, suponiendo un k=3.
3. Compare los resultados obtenidos en los dos puntos precedentes y documente.

3. Incorpore a una herramienta de minería de datos el dataset *bank-full* y realice las siguientes operaciones:
   a. Aplique el algoritmo *Naive Bayes* y documente las métricas de evaluación obtenidas y la clasificación realizada para cada instancia.
   b. Aplique el algoritmo *C-4.5 (árboles de decisión)* y documente las métricas de evaluación obtenidas y la clasificación realizada para cada instancia.
   c. Aplique el algoritmo *Random Forest*[1] y documente las métricas de evaluación obtenidas y la clasificación realizada para cada instancia.
   d. Calcule y compare las métricas obtenidas en cada caso. ¿Qué le permite concluir la comparación precedente en términos de las métricas de evaluación estudiadas?

Referencias sugeridas:

Random Forest (R):

https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Data Mining: Concepts and Techniques. Jiawei Han & Micheline Kamber. Morgan Kaufmann. Second Edition. 2006.

[1] Para utilizar Random Forest en R deberá instalar la librería **randomForest**.