LEARNING MACHINE LEARNING 2018
UNIVERSIDAD DEL ROSARIO

LAB 5: RANDOM FORESTS AND GRADIENT BOOSTING FOR CLASSIFICATION
Instructores: Luis Felipe Giraldo y Dora Suarez
Julio 25 de 2018

PROBLEMA:

El proceso de administración de los recursos en un hospital depende de varios aspectos, entre los cuales se resalta uno muy importante: la cantidad de tiempo que un paciente va a estar hospitalizado una vez es admitido. La incertidumbre en esta variable afecta en gran medida la utilización eficiente de los recursos del hospital, como la disponibilidad de mano de obra e instalaciones. Una predicción precisa del tiempo de estancia de los pacientes permitiría una mejor administración de los recursos del hospital, que resultaría en un mejor servicio. En este laboratorio el objetivo es entrenar y comparar los modelos de clasificación Random Forests y Gradient Boosting para predecir el tiempo de estancia de un paciente en caso de que éste sea admitido y calcular la importancia de las variables para la predicción.

BASE DE DATOS

La base de datos fue suministrada por el Ministerio de Salud y Protección Social de Colombia, y los datos fueron pre-procesados y analizados por Samir Char en su proyecto de grado de Ingeniería Electrónica titulado "Machine Learning Tecniques to predict hospital Length of Stay (LOS) at the time of patient admission," Universidad de los Andes, 2017, bajo la asesoría del profesor Luis Felipe Giraldo.

Esta base de datos contiene información de 219.440 pacientes que fueron admitidos en hospitales en la ciudad de Bogotá. Por paciente hay mediciones de nueve (9) variables: ocho (8) predictores, y una (1) variable a predecir.

Predictores:

- Causa externa: Variable categórica que toma valores entre 15 posibilidades. Las categorías incluyen accidente de trabajo, accidente en automóvil, accidente debido a rabia con animal o humano, accidente por serpiente, evento catastrófico, heridas debido a agresión, heridas auto infligidas, sospecha de abuso físico, sospecha de abuso sexual, enfermedad debido al trabajo, entre otras.
- Medio de entrada al hospital: Variable categórica que toma valores entre 4 posibilidades. Las categorías incluyen unidad de cuidados intensivos, consulta externa, nacimiento en las instalaciones, y referido.
- Diagnóstico que da origen a la admisión: Variable categórica con el diagnóstico dado por un médico antes de la admisión. Las categorías incluyen dolor abdominal, infección del tracto urinario, apendicitis, entre otras.
- Código de la IPS: Variable categórica asociada a la IPS.
- Código de la administradora de salud: Variable categórica asociada a la administradora de salud.
- Ocupación del hospital: Variable con el número de pacientes en el hospital en el momento de ser admitido.
- Número de reinserciones: Número de veces que el paciente ha sido internado en el hospital.

- Edad

Variable a predecir:

Tiempo de estancia en el hospital: Variable categórica que toma valores entre 3 posibilidades. La categoría es 1 es asignada si el paciente estuvo menos de 1.24 días en el hospital, 2 si estuvo entre 1.24 y 3.54 días, y 3 si el paciente estuvo más de 3.54 días.

Para información más detallada sobre la base de datos, por favor consultar el documento:
Char, Samir. "Machine Learning Tecniques to predict hospital Length of Stay (LOS) at the time of patient admission," Proyecto de grado, Ingeniería Electrónica, Universidad de los Andes, 2017.

ACTIVIDAD DE LABORATORIO

Random Forests

1. Considere el archivo `ejercicio_RF.py`. Entienda cada línea de código y su función en la solución del problema.

2. Determine el porcentaje de datos utilizados para entrenamiento y prueba.

3. Explique qué es una matriz de confusión en un problema de clasificación.

4. Análisis de resultados:
   - Observe la importancia de las variables en el proceso de predicción e intente dar una explicación a estos resultados.
   - Analice la matriz de confusión obtenida. Determine qué información útil provee esta matriz sobre este problema de predicción en particular.

Gradient Boosting

1. Considere el archivo `ejercicio_RF.py`. Complete este archivo al que resuelva el problema de clasificación utilizando método de predicción Extreme Gradient Boosting (xgboost en pyton https://xgboost.readthedocs.io/en/latest/python/index.html) siguiendo la estructura del ejercicio anterior (entrenamiento, prueba, precisión en la clasificación, matriz de confusión, importancia de las variables). Considere la siguiente tabla con los parámetros del algoritmo en Python:

Parameter	Value
eta	0.1
n_estimators	300
eval_metric	merror
max_depth	8
min_child_weight	8
nthread	4
num_class	4
objective	multi:softmax
subsample	1
colsample_bytree	0.6

2. Analice los resultados obtenidos (importancia de las variables). Compare con los resultados obtenidos con el método de Random Forests.
3. Compare de forma cualitativa el costo computacional (tiempo y memoria) y la capacidad de predicción (precisión y matriz de confusión) de los dos métodos.

EJERCICIOS

- Estudie el funcionamiento del Extreme Gradient Boosting, y entienda las características de esta implementación que lo hacen eficiente y preciso.
- Selecciones algunos de los parámetros de ambos métodos, y asígnele diferentes valores. Estudie el efecto de esta variación sobre la precisión y eficiencia de los algoritmos.
- Lea esta discusión y el material citado allí comparando ambos métodos: https://stats.stackexchange.com/questions/173390/gradient-boosting-tree-vs-random-forest