FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Tarea 3 - EYP3407

Profesor: Mauricio Castro
Ayudante: Bladimir Morales

1. **(60%)** En el contexto de los árboles y bósques aleatorios de clasificación muestre que:

   - **(20%)** Si $\phi(p_1, \ldots, p_K)$ es una función estrictamente cóncava en $0 \leq p_k \leq 1$, $k = 1, \ldots, K$, y $\sum_k p_k = 1$, entonces, para $\phi(p(1 \mid \tau), \ldots, p(K \mid \tau))$ y cualquier split s, se tiene que $\Delta i(s, \tau) \geq 0$, con igualdad, si solo si, $p(k \mid \tau) = p(k \mid \tau_L) = p(k \mid \tau_R)$, con $k = 1, \ldots, K$.
   - **(20%)** Demuestre que $R^{re}(\tau) \geq R^{re}(\tau_L) + R^{re}(\tau_R)$.
   - **(20%)** Se dice que en el contexto del *bagging*, en promedio, cada árbol basado en una muestra *bootstrap* usa al rededor de 2/3 de las observaciones. El 1/3 restante no es utilizado para ajustar dicho árbol. Ese grupo de observaciones se conoce como las observaciones OOB (*out-of-bag*). Suponga entonces que se desea realizar *bootstrap* de una muestra de n observaciones.
     - **(15%)**¿Cuál es la probabilidad de que la j-ésima observación de la muestra original no esté en la muestra *bootstrap*? Detalle sus cálculos.
     - **(5%)** Grafique la probabilidad de que la j-ésima observación de la muestra original esté en la muestra *bootstrap* en función de n (considere una grilla de valores de 1 a 100000). ¿Es cierta la afirmación que se hace sobre las observaciones OOB? Realice un experimento numérico usando simulación que reafirme o contradiga lo anterior (la afirmación sobre las observaciones OOB). Comente sus hallazgos.

2. **(40%)** Considere los datos de LendingClub.com *loan.csv* desde el año 2007 hasta el año 2015. Lending Club conecta a las personas que necesitan dinero (prestatarios) con las personas que tienen dinero (inversionistas). Para los inversionistas, la pregunta a responder es ¿a qué personas prestarles dinero? Las variables disponibles son:

   - *purpose*: El propósito del préstamo (toma los valores "credit_card", "debt_consolidation", "education", "high_purchase", "small_business", y "all_other").
   - *int_rate*: La tasa de interés del préstamo, como proporción (una tasa del 11% se almacenaría como 0.11).
   - *installment*: Las cuotas mensuales adeudadas por el prestatario si el préstamo es financiado.

- *annual_inc*: los ingresos anuales auto-reportados del prestatario.
- *dti*: La relación deuda-ingreso del prestatario (monto de la deuda dividido por el ingreso anual).
- *revol_bal*: el saldo rotativo del prestatario (monto no pagado al final del ciclo de facturación de la tarjeta de crédito).
- *revol_util*: tasa de utilización de la línea rotativa del prestatario (el monto de la línea de crédito utilizada en relación con el crédito total disponible).
- *inq_last_6mths*: El número de consultas crediticias del deudor realizadas por los prestatarios en los últimos 6 meses.
- *delinq_2yrs*: La cantidad de veces que el prestatario ha estado vencido por más de 30 días en un pago en los últimos 2 años.
- *pub_rec*: El número de registros públicos derogados del prestatario (declaraciones de quiebra, gravámenes fiscales o fallos).
- *loan_status*: estado del préstamo (fully paid es completamente pagado).

Ud. como consultor estadístico deberá pre-procesar los datos y considerar los siguientes clasificadores: árboles de clasificación, bagging, boosting, random forest, AdaBoost y otra técnica de clasificación su elección para contestar la pregunta de investigación. Reporte tasas de mala clasificación (evalúe la pertinencia de las métricas a considerar dependiendo del desbalance), error OOB, importancia de las variables, y realice un análisis de sensibilidad sobre el número de árboles a considerar (cuando se necesite definir este número). Comente los resultados y detalle todos los aspectos técnicos utilizados en los métodos considerados. Utilize la razón 70-30 para determinar el conjunto de entrenamiento y de testeo. Cualquier decisión que tome sobre los datos o métodos considerados deberá estar debidamente justificada en el reporte a presentar.