

Investigación de Operaciones

Informe Árboles de Clasificación

Profesores: Daniel Quinteros - Nicolás Rojas M.
Ayudantes: Felipe Orellana, Diana Gil, Joaquín Gatica

Primer Semestre 2023

1. Objetivo

El objetivo de este trabajo es utilizar una herramienta para construir árboles de clasificación, evaluando el efecto de utilizar diferentes conjuntos de datos, atributos y tipos de atributos. Además, permitirá conocer y aprender a utilizar parte del software R.

2. Enunciado

La detección de spam es un problema importante en la industria de la tecnología, ya que el correo electrónico no deseado representa una gran cantidad de tráfico en Internet y puede ser perjudicial para los usuarios. Se han utilizado diferentes técnicas de clasificación de datos, entre ellas árboles de clasificación, para identificar el spam y filtrarlo de manera efectiva¹. En esta tarea, se utilizarán árboles de clasificación para desarrollar un modelo de detección de spam en correo electrónico. Se proporcionará un conjunto de datos de correo electrónico etiquetados como spam o no spam, el cual debe ser utilizado para construir/entrenar un árbol que sea capaz de discriminar esta situación.

Utilizando el software **R**, el paquete **tree**² y **seteando la semilla con el número de su rol, sin guión**³, conteste las siguientes preguntas:

- (10 puntos) Describa el conjunto de datos: Cantidad de datos, Tipo y Valores posibles de cada atributo⁴, Indique la cantidad de registros por clase.
 - (2 puntos) Cantidad de datos
 - (2 puntos) Tipos de datos
 - (3 puntos) Valores posibles de cada atributo
 - (3 puntos) Elementos por clase
- (6 puntos) Convierta a *factor* el atributo *target* (*'Prediction'*)⁵ y el atributo *'attachment_size'*, elimine los atributos que no se utilizarán (los que fueron convertidos), compruebe sus modificaciones.
 - (4 puntos) Convertir a factor
 - (2 puntos) Datos resultantes no contienen los atributos originales
- (20 puntos) Genere e imprima un árbol considerando el 70 % de los datos entregados para *training*, utilice como criterio de división **gini** y **deviance**, luego conteste lo siguiente:
 - Describa el árbol obtenido: cantidad de niveles (profundidad) y de hojas del árbol.
 - Evalúe el árbol utilizando la métrica *accuracy*⁶, imprima dicha métrica y comente acerca del desempeño del modelo.

¹<https://www.semanticscholar.org/paper/Designing-Spam-Model-Classification-Analysis-using-Rajput-Arora/9984a7d06e04a347718cb8c7f645b72195bb11ce>

²<https://cran.r-project.org/web/packages/tree/index.html>

³Por ejemplo `set.seed(2023001001)`

⁴Si su descripción aplica para varios atributos puede mencionarlo, para no repetirla.

⁵Tener precaución con el atributo *'predictions'*, que es distinto y se refiere a la frecuencia de dicha palabra.

⁶Casos correctamente clasificados / total de casos evaluados.

- (4 puntos) Seleccionar 70 % para *training*.
 - (3 puntos) Utilizar parámetros (criterio división) adecuados
 - (3 puntos) Cantidad de niveles
 - (3 puntos) Cantidad de hojas
 - (4 puntos) Cómputo de *accuracy*
 - (3 puntos) Comentario: evaluar que esté alineado a los resultados
4. (30 puntos) ¿Qué sucede si utiliza un 30 %, 50 % y 70 % de los datos entregados como *training*? Genere, imprima y evalúe los árboles obtenidos (utilice *accuracy*). Compare los resultados e indique qué árbol es el mejor.

Anteriormente se evaluó construcción de árbol y cómputo de *accuracy*, por eso ahora no se le asigna menos puntaje.

- (5 puntos) Construcción e impresión árbol 30 %
 - (5 puntos) Construcción e impresión árbol 50 %
 - (5 puntos) Construcción e impresión árbol 70 %
 - (4 puntos) *Accuracy* árbol 30 %
 - (4 puntos) *Accuracy* árbol 50 %
 - (4 puntos) *Accuracy* árbol 70 %
 - (3 puntos) Comentario: debe estar justificado
5. (34 puntos) Se desea mejorar el rendimiento temporal del árbol, sin disminuir su efectividad, considerando esta vez el 80 % de los datos para *training*. Construya, imprima y evalúe los siguientes árboles, utilizando los siguientes atributos para cada caso:
- a) Solo atributo '*thanks*'
 - b) Atributos '*thanks*' y '*subject*'
 - c) '*thanks*', '*subject*' y '*attachment_size*'
 - d) Todos los atributos

Compárelos, indique cuál es el mejor y por qué.

Nuevamente, ya se evaluó dos veces construcción, impresión y evaluación de árboles, quedando el puntaje:

- (6 puntos) Construcción, impresión y evaluación árbol primer item
- (6 puntos) Construcción, impresión y evaluación árbol segundo item
- (6 puntos) Construcción, impresión y evaluación árbol tercer item
- (6 puntos) Construcción, impresión y evaluación árbol cuarto item
- (10 puntos) Comparación y conclusión: debe estar justificada en términos de la métrica (*accuracy*), indicar que a mayor *accuracy* es mejor, pero que no se relaciona necesariamente con la cantidad de atributos utilizados para la construcción del árbol.

3. Requerimientos e Informaciones

- En cada pregunta, mencione qué comandos y funciones utilizó.
- El trabajo es individual.
- Los informes deben ser subidos a plataforma aula **con plazo máximo** el día 5 de Junio hasta las 23:59 horas. **Informes entregados fuera de plazo serán calificados con nota 0.**
- Debe subir una carpeta, que tenga por nombre su ROL, comprimida en .zip que contenga: (1) un archivo jupyter notebook con el código y (2) un pdf con el informe.
- El software R está disponible en <https://www.r-project.org/>
- Es obligación detallar y fundamentar adecuadamente cada respuesta, así como comentar el código que escriba.

- Debe incluir imágenes de los árboles en toda pregunta que se solicite.
- **Los datos del trabajo están disponibles en aula, en la unidad correspondiente.** Considere como *hint* la siguiente descripción del *dataset*:
 - La columna que indica si un caso es o no *spam* se llama *Prediction*.
 - Existe un atributo llamado '*attachment_size*' que indica el tamaño del archivo adjunto en el correo, en caso que posea.
 - Las demás columnas representan la frecuencia de apariciones de cada palabra en correos electrónicos. Por ejemplo: Columna '*subject*' con valor 10, indica que dicha palabra aparece 10 veces en ese correo.

Cualquier duda relacionada sobre el enunciado o con los contenidos, favor publicar en aula o escribir un mail directo a sus ayudantes.