
Aprendizaje de Máquinas, Laboratorio #1

Clasificación de SPAM: Árboles de Decisión, Naïve Bayes

Fecha de entrega **Primer parte L1a: 7 de mayo, 23:59 hs. Segunda parte Bayes L1b: 14 de mayo**

Nota: (1) Estas preguntas requieren pensar, pero no requieren largas respuestas. Por favor se tan conciso como sea posible. (2) Cuando envíes una pregunta al foro, por favor asegúrate de escribir el número de laboratorio y el número del problema, tal como L1 P2. (3) Para problemas que requieran programación, por favor incluye en tu envío el código (con comentarios) y cualquier figura que se haya solicitado graficar. Ten en cuenta que el código debe poder correr desde cualquier máquina (4) Si escribes tus soluciones a mano, por favor escribe claramente y utilizando una birome de color oscuro. **Los apartados teóricos ofrecen los contenidos mínimos para realizar las actividades del laboratorio. Hay muchos temas que no se incluyen, sin embargo no dejan de ser importantes. No descarte consultar la bibliografía ofrecida en este curso.**

§ Ejercicio 1. [10 puntos] Motivación de Decision Tree y Naïve Bayes: resolviendo clasificación de spam

Las consignas a desarrollarse en este punto se encuentran en el archivo llamado *cajaNegra.Rmd*. Abra dicho archivo en Rstudio, lea, interpreta, resuelva en el archivo *cajaNegra.Rmd*.

Decision Tree (L1a)

§ Ejercicio 2. [30 puntos] ID3 + Ganancia de información

Para este ejercicio debe utilizar el template `dt.R`, en el mismo se indican las funciones que deben implementarse, como `learn.tree()` (algoritmo ID3 tabla 3.1 de [1]), `best.attribute()` (*Gain Information*, fórmula 3.4 de [1]) y `classify.example()`.

Use `MATRIX.TRAIN` para entrenar el clasificador. Tenga en cuenta limpiar el dataset de spam antes de usarlo para entrenamiento, omitiendo los tokens *spam*, *news* y *httpaddr*, tal como se explicó en el punto anterior. Puede probar su algoritmo primero con otros datasets más chicos, como *PlayTennis* o *Restaurant* que se encuentran en `data`. Todas las tareas de carga y procesamiento de los datasets deben realizarse en la función `load.data`

Se adjunta el archivo `dt.tutorial.R` que explica como utilizar funciones bases para manejar la estructura árbol. Se recomienda revisar este tutorial para implementar el algoritmo ID3.


Recuerde que NO debe modificar los nombres de las funciones de los templates, ni los nombres de las variables o datos regresados por las funciones, simplemente debe completar el código faltante indicado como `# ADD YOUR CODE HERE`.

§ Ejercicio 3. [10 puntos]

Explique y detalle como implementaría *Gain Information* para el caso en el que existan missing values (sección 3.7.4 de [1]). Use y de al menos un ejemplo. Puede crear un notebook en R para documentar su respuesta.

§ Ejercicio 4. [15 puntos]

Implemente *Split Information* (sección 3.7.3 de [1]). Evalúe la performance del clasificador usando *Gain Information* y *Split Information* para el dataset de SPAM y compare los resultados.

Naïve Bayes (L1b)  Parte 2

§ Ejercicio 5. [35 puntos] Naïve Bayes

Implemente un clasificador naïve Bayes para clasificar spam. Aprenda sus parámetros usando la matrix palabra-documento en `MATRIX.TRAIN`, y entonces reporte el error al clasificar el conjunto de testeo en `MATRIX.TEST`. Para esto debe usar el template de código en `nb.R`. Lea atentamente los comentarios en `nb.R`. La función principal es `run.experiment`.

Referencias

[1] Tom Mitchell. *Machine Learning*. McGRAW-HILL, 1997.