Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Tarea 2: Minería de Datos IIC2433

Fecha de entrega : 31 de octubre de 2018, 23:59 hrs.

Introducción

El árbol de decisión es uno de los algoritmos de aprendizaje supervisado más conocidos en Machine Learning [Quinlan, 1986]. Estos permiten realizar tareas de clasficación o regresión. Un árbol de clasificación retorna categorías/clases mientras que un árbol de regresión corresponde a predicciones en intervalos reales continuos [Breiman et al., 1984]. En esta tarea serán guiados para que implementen sus propios árboles de regresión y posteriormente la prueben sobre una base de datos real.

Árboles de Regresión

Como fue visto en clases, un árbol es una estructura que divide el espacio de variables regresoras/clasificadoras en secciones y asigna un valor a la predicción a cada una, como se ilustra en la siguiente figura:

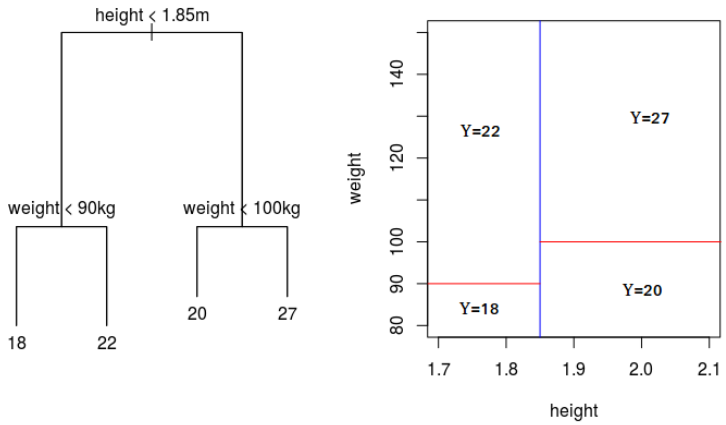

Fig 1. Puntaje basquetbolista según altura y peso

Para la construcción del árbol, por cada nodo se debe decidir la variable de división según una métrica que indique la "homogeneidad" de los hijos resultantes de ella. Una con la cual ustedes ya se encuentran familiarizados es según "Ganancia de Información" que emplea el concepto de Entropía [Shannon, 1948]. Sin embargo, esta medida no es válida para el caso donde la predicción puede tomar infinitos valores, por lo que: ¿qué medida utilizamos para definir cuál es la mejor variable?

Por otra parte, el árbol para clasificación retorna la clase más común una vez alcanzada una hoja: ¿cómo se define el valor de la predicción para cada hoja cuando el valor a predecir es un número Real?

Una posible solución a las preguntas anteriores consiste en el uso de medidas de varianza. Su primera tarea consistirá en averiguar cómo abordar estas problemáticas que surgen al usar árboles de regresión en lugar de árboles para clasificación. Para ello pueden ser útiles los siguientes links:

- http://www.appliedaisystems.com/papers/RegressionTrees.pdf
- http://www.stat.cmu.edu/cshalizi/350-2006/lecture-10.pdf
- http://uc-r.github.io/regression_trees

Aparte del criterio de selección de variables (o *splits*) y la forma de *predicción*, el algoritmo de construcción de un árbol de regresión opera de manera similar al árbol de decisión para clasificar.

Pasos a seguir

Para completar la tarea ustedes deben seguir los siguientes pasos:

1. **Preprocesar** la base de datos. Esta consiste en una base de datos astronómica *FATS_GAIA.csv*, que se encuentra adjunta al enunciado. Contiene *features* de series de tiempo astronómicas del *survey* GAIA [1]. El **target** (variable a predecir) corresponde a la columna *PeriodLS* y **data** (variables predictoras) a todas las otras columnas. Deben dividir esta base de datos en un set **train** (%80) y otro **test** (%20).

2. **Implementar** el algoritmo de árboles de regresión. En este punto el código debe contar con funciones *fit* y *predict*:

   - *fit*: Debe aplicar el algoritmo a la base de datos. Debe recibir **data**, **target** y opcionalmente permitir definir la profundidad máxima *max_depth* (para evitar overfitting).

     Considere que las variables con que su algoritmo deberá trabajar son continuas, esto implica implementar una forma automática de encontrar el punto de split de cada nodo.

   - *predict*: Debe recibir **data** y retornar la predicción. Para la implementación solo podrá utilizar las librerías Numpy y Pandas.

3. **Aplicar** el algoritmo a la base de datos de prueba. Entrene con **data_train** y haga predicciones sobre **data_test**. Debe entregar una medida de qué tan buena es su predicción sobre **data_test**.

[1] http://sci.esa.int/gaia/

4. **Visualizar** un árbol. En este paso puede usar *matplotlib* o *graphviz*.

5. **Explicar** y analizar los resultados obtenidos (puede apoyar su explicación en la visualización), comentar los experimentos realizados ¿Cómo espera que cambie la predicción al setear distintos niveles de profundidad del árbol?

1 Entrega

- La entrega debe ser realizada en un .zip con todos los archivos necesarios.
- El archivo de entrega debe ser subido al cuestionario abierto en el sistema SIDING específicamente para esta tarea hasta el día y hora señalada con el nombre $[numero\_alumno]\_T2$.
- En caso de atraso, se aplicará un descuento lineal de nota 7 a 1 en 24 horas.
- La tarea es estrictamente individual y el algoritmo debe ser implementado 100% (no usar funciones previamente implementadas o re-utilizar código).
- El documento principal debe ser un jupyter notebook con el código.
- Cualquier instrucción adicional y necesaria para la revisión debe ser escrita en un archivo README.txt contenido en el .zip

References

J. R. Quinlan. Induction of decision trees. *MACH. LEARN*, 1:81–106, 1986.

Leo Breiman, J. H Friedman, R. A Olshen, and C. J. Stone. Classification and regression trees. *Wadsworth Brooks/Cole Advanced Books Software*, 1984.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.