Árboles de decisión - Teoría

Un árbol tiene muchas analogías en la vida real, y resulta que ha influido en una amplia área del aprendizaje automático o Machine Learning. Los árboles de decisión son una técnica de aprendizaje supervisado que predice valores de respuestas mediante el aprendizaje de reglas de decisión derivadas de características. Se pueden utilizar tanto en una regresión como en un contexto de clasificación.

Los árboles de decisión funcionan al dividir el espacio de la característica en varias regiones rectangulares simples, divididas por divisiones paralelas de ejes. Para obtener una predicción para una observación particular, se utiliza la media o el modo de las respuestas de las observaciones de entrenamiento, dentro de la partición a la que pertenece la nueva observación.

Veamos de manera matemática la función del árbol de decisión:

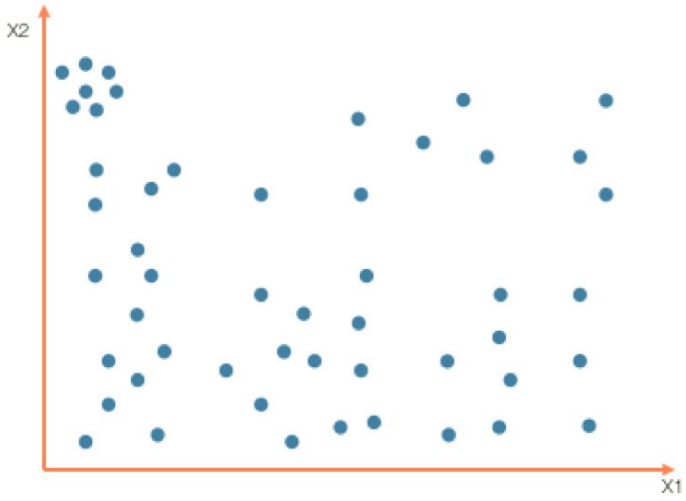
$$f(x) = \sum_{m=1}^{M} w_m \phi(x; v_m)$$

Donde:

wm es la respuesta media en una región particular (Rm).

vm representa cómo se divide cada variable en un valor de umbral particular.

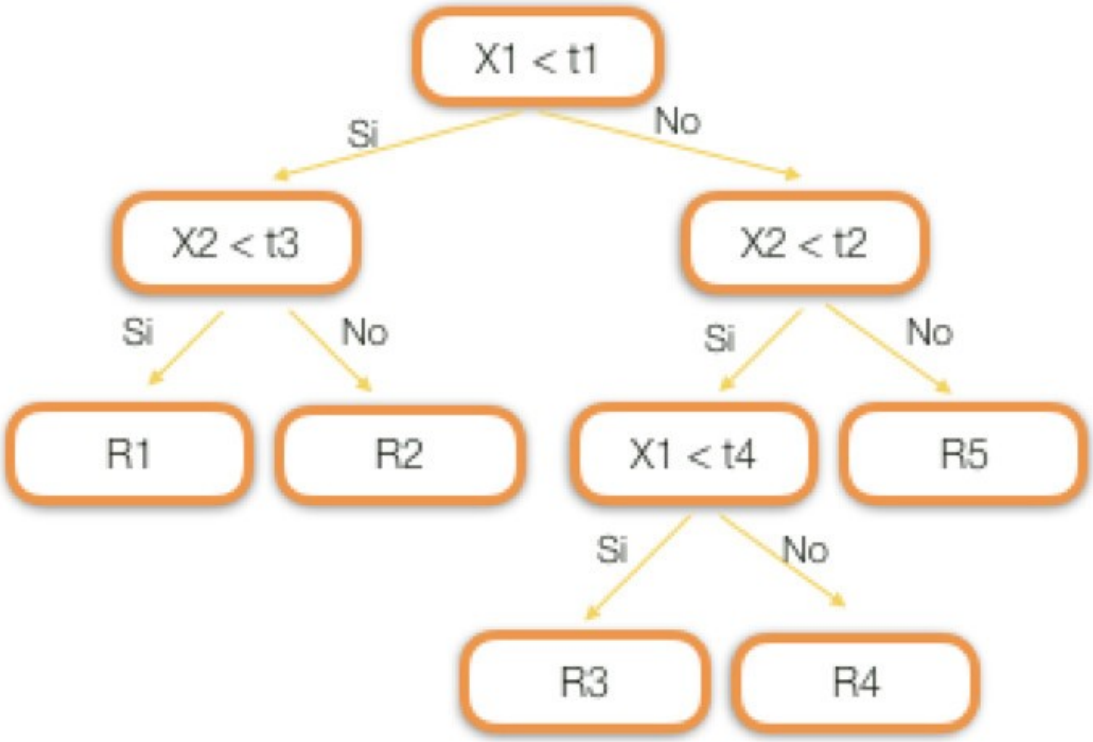
Estas divisiones definen cómo el espacio de características en R^2 en M regiones separadas, hiperbloques.

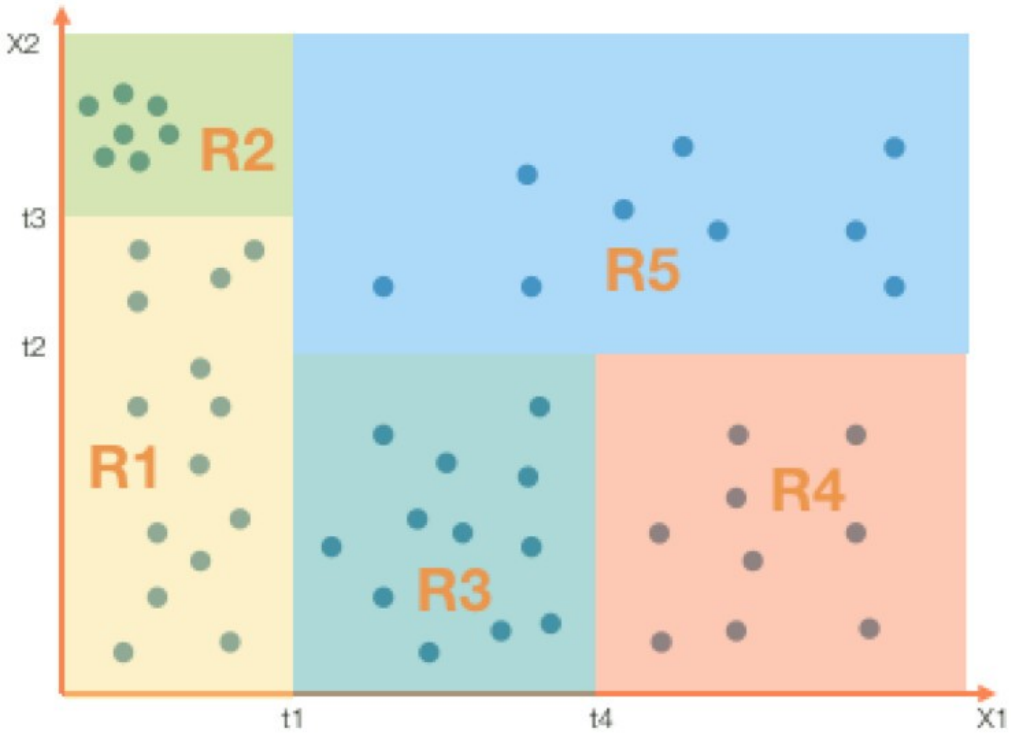


Árboles de decisión - Teoría

Consideremos un ejemplo con dos variables de características (X1 y X2) y una respuesta numérica “y”.

En la siguiente figura podemos ver un árbol desarrollado para este ejemplo en particular, con seis separadas.





Árboles de decisión - Teoría

Pero, ¿cómo se corresponde esto con una partición del espacio de características?, bueno en la siguiente figura se muestra un subconjunto que contiene nuestros datos de ejemplo. Observa cómo se divide el dominio mediante divisiones paralelas de eje, es decir, cada división del dominio se alinea con uno de los ejes de características.

El concepto de división paralela de ejes se generaliza directamente a dimensiones superiores a dos. Para un espacio de características de tamaño p, un subconjunto de , el espacio se divide en regiones M, , cada una de las cuales es un hiperbloque p-dimensional.

Creemos ahora un árbol de regresión y hagamos predicciones con él. La heurística básica para crear un árbol de decisión es la siguiente:

- Dadas las características p, divide el espacio de características p-dimensional, en M regiones mutuamente distintas que cubren completamente el subconjunto del espacio de características y no se superponen. Estas regiones están dadas por R1, ..., Rm .
- Cualquier observación nueva que caiga en una partición particular tiene la respuesta estimada dada por la media de todas las observaciones de entrenamiento con la partición denotada por .

Sin embargo, este proceso no describe realmente cómo formar la partición de una manera algorítmica. Para eso necesitamos usar una técnica conocida como división binaria recursiva.

Nuestro objetivo para este algoritmo es minimizar algún tipo de criterio de error. En este particular, deseamos minimizar la suma de cuadrados residual (RSS), una medida de error también utilizada en la configuración de regresión lineal.

Desafortunadamente, es demasiado costoso computacionalmente considerar todas las particiones posibles del espacio de la característica en rectángulos M, por lo tanto debemos utilizar un enfoque de búsqueda menos intensivo en computación, pero más sofisticado, aquí es donde entra la división binaria recursiva.

La división binaria recursiva aborda el problema comenzando en la parte superior del árbol y dividiendo el árbol en dos ramas, lo que crea una partición de dos espacios. Lleva a cabo esta división en particular en la parte superior del árbol varias veces y elige la división de las características que minimiza la suma de cuadrados residual (RSS).

En este punto, el árbol crea una nueva rama en una partición particular y lleva a cabo el mismo procedimiento, es decir, evalúa el RSS en cada división de la partición y elige el mejor.

Esto lo convierte en un algoritmo codicioso, lo que significa que lleva a cabo la evaluación para cada iteración de la recursión, en lugar de «mirar hacia adelante» y continuar bifurcándose antes de realizar las evaluaciones. Es esta naturaleza «codiciosa» del algoritmo la que lo hace computacionalmente factible y, por lo tanto, práctico para su uso.

Árboles de decisión - Teoría

Ahora, el principal problema con el árbol de decisión es que es propenso a sobreajuste. Podríamos crear un árbol que pudiera clasificar los datos a la perfección o no nos queda ningún atributo para dividir. Esto funcionaría bien en el conjunto de datos de entrenamiento, pero tendrá un mal resultado en el conjunto de datos de prueba. Existen dos enfoques populares para evitar esto en los árboles de decisión: detenga el crecimiento del árbol antes de que sea demasiado grande o pode el árbol después de que sea demasiado grande.

Por lo general, un límite para el crecimiento de un árbol de decisión se especificará en términos del número máximo de capas, o la profundidad, que puede tener. Los datos disponibles para entrenar el árbol de decisión se dividirán en un conjunto de entrenamiento y un conjunto de prueba y se crearán árboles con varias profundidades máximas en función del conjunto de capacitación y se compararán con el conjunto de prueba. La validación cruzada también se puede utilizar como parte de este enfoque.

La poda del árbol, por otro lado, implica probar el árbol original contra versiones podadas de él. Los nodos de la hoja se retiran del árbol siempre que el árbol podado funcione mejor contra los datos de prueba que el árbol más grande.

Si bien los modelos de árbol de decisión en sí mismos tienen un rendimiento de predicción no tan eficiente, son extremadamente competitivos cuando se utilizan en una configuración de conjunto, de esto hablaremos en otra entrada.