
Universidad
Tecmilenio®

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Métodos lineales generalizados

Introducción

Es importante que conozcas los modelos lineales generalizados, ya que estos representan la expansión de los modelos lineales (al incorporarle la capacidad de manejar datos no lineales), realizando transformaciones adecuadas que los puedan convertir a una forma lineal.

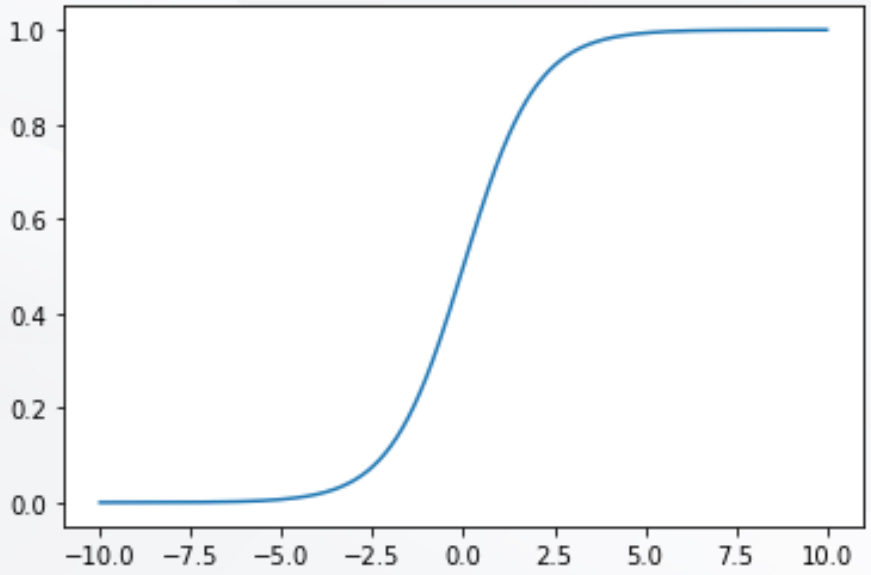
En muchos casos, una relación no lineal se puede convertir en una relación lineal, pero agregando un paso adicional para transformar uno de los datos (entrada o salida) en otro dominio. La función que realiza dicha transformación se denomina como función de base o función de enlace. La regresión logística es uno de esos ejemplos donde se utiliza la función logística como función base para transformar la no linealidad en linealidad.

En este tema verás varios de los detalles de la regresión logística y aprenderás lo fácil que es de implementarlo. También se dará un recorrido por las dos principales problemáticas a las que se enfrenta el aprendizaje automático supervisado: la regresión y la clasificación.

Explicación

La regresión logística introduce una forma simple de realizar la clasificación binaria cuando en una situación dada hay dos clases de valores bien definidos. Aunque en su nombre se incluye el término “regresión”, su uso principal ha sido en la resolución de problemas de clasificación.

La función base que utiliza este método es la función **sigmoide** o función **logística**, la cual ha sido ampliamente utilizada para modelar el crecimiento de poblaciones naturales que se desarrollan rápidamente y que tienden a saturar la capacidad del ambiente que las sostiene. Por ende, de manera gráfica representa una curva en forma de “S” que tiene los valores 0 y 1 como asíntotas.



Explicación

La regresión logística modela la probabilidad de que una salida corresponda a un grupo de referencia de manera indirecta, utilizando la razón de probabilidad.

Por ejemplo, supongamos que se está construyendo un modelo que sea capaz de diferenciar naranjas y toronjas con base en el diámetro de la fruta. En este caso, la clase principal es la toronja, por lo que el modelo ha sido elaborado a partir de cierto valor del diámetro medido (por ejemplo, 15 cm):

$$P(fruta = toronja|diámetro) \Leftrightarrow P(X) = P(y = 1|X)$$

Si en este ejemplo la probabilidad de que la fruta sea una toronja es de 0.75, entonces la probabilidad de que sea una naranja es 1 - 0.75 = 0.25.
La razón de probabilidad quedaría definida como:

$$\frac{P(fruta = toronja|diámetro)}{P(fruta = naranja|diámetro)} = \frac{0.75}{0.25} = 3$$

Explicación

Dado que las probabilidades están limitadas al rango [0, 1], la razón de probabilidad está acotada entre [0, ∞]. Al aplicar la operación logarítmica, el rango de valores pasa a ser de [-∞, +∞]. Para este ejemplo, la expresión del modelo quedaría de la siguiente forma:

$$P(X) = \frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}}$$

Para encontrar los coeficientes se replantea el modelo utilizando la razón de probabilidad:

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X$$

Explicación

Para realizar las predicciones se sustituyen los valores de los coeficientes encontrados y se evalúa el modelo para una muestra determinada.

Como la salida del modelo logístico es un valor de probabilidad, para conseguir la clasificación es necesario establecer un criterio (*threshold*) a partir del cual se considera que la variable pertenece a uno de los niveles:

$$0 \; si \; P(toronja) < 0.5$$

$$1 \; si \; P(toronja) > 0.5$$

Explicación

Es posible desarrollar un programa en donde se aplique la regresión logística para clasificar un conjunto de datos base.

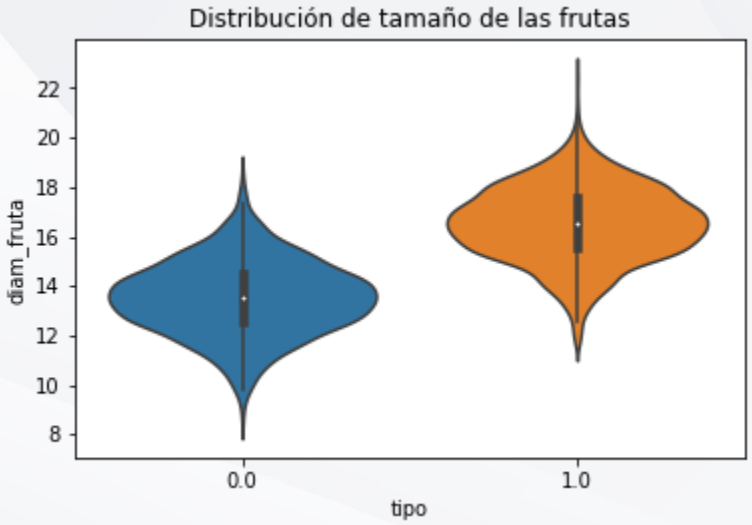
En este ejemplo se parte de la misma situación de las naranjas y las toronjas, pero utilizando un conjunto de datos clasificados con 2000 mediciones de los diámetros de las mismas (50% de cada tipo).

	diam_fruta	tipo
0	14.267035	0.0
1	12.202880	0.0
2	12.916410	0.0
3	11.492695	0.0
4	12.272310	0.0
...	...	...
1995	13.898548	1.0
1996	17.304767	1.0
1997	18.683588	1.0
1998	17.227021	1.0
1999	17.727017	1.0

2000 rows × 2 columns

Explicación

Como verificación de las condiciones, para aplicar la regresión logística se puede graficar la forma en la que están distribuidos los valores de los diámetros medidos para cada clase.

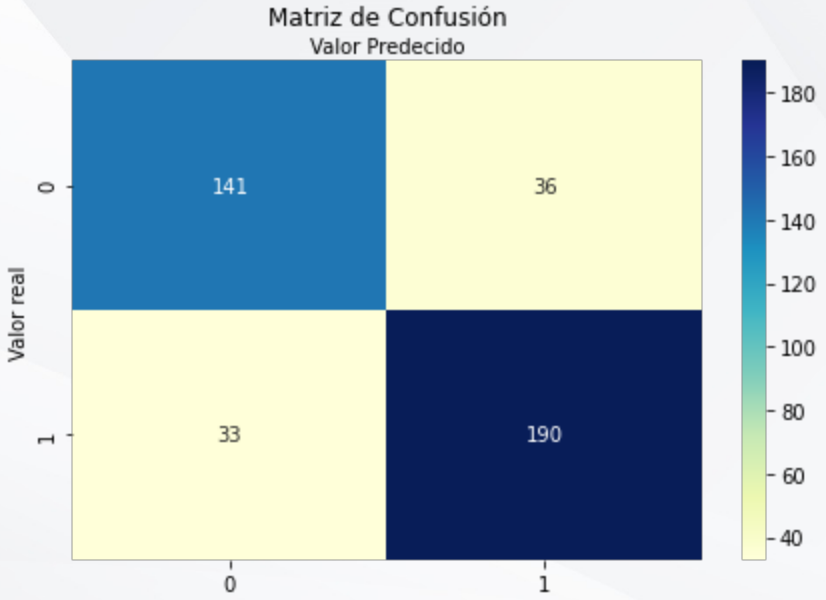


Se puede apreciar la existencia de dos clases principales, es decir, donde los valores correspondientes a las dimensiones de los diámetros medidos describen un comportamiento gaussiano.

Esta pequeña verificación permite tener claridad sobre las relaciones de los datos, por lo que asegura la presencia de algunas condiciones para utilizar la regresión logística como método de clasificación.

Explicación

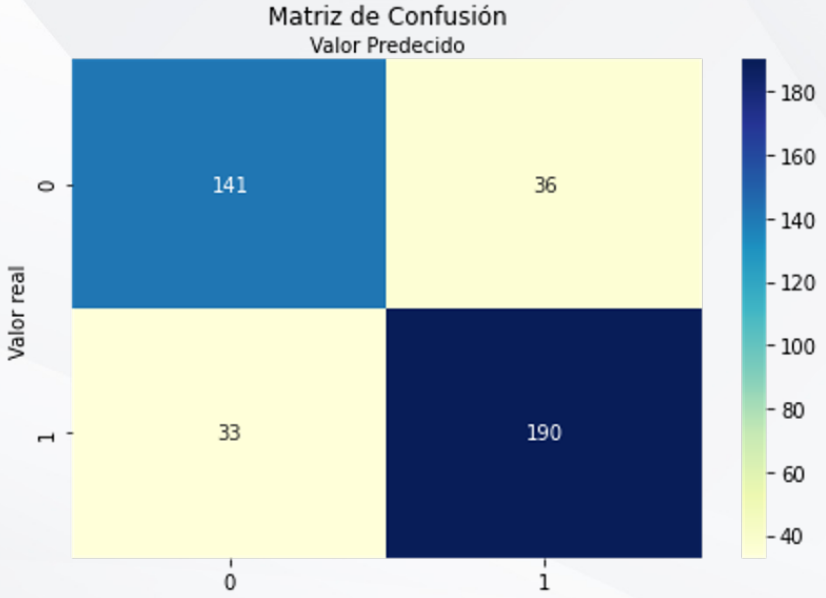
Después de aplicar la librería Scikit-learn de Python, se programó y entrenó a un modelo de regresión logística. Por ende, el siguiente paso es realizar pruebas con datos que no se utilizaron para el entrenamiento.



Una de las formas más comunes de comprobar la efectividad de un modelo de este tipo es utilizando la matriz de confusión. Este recurso muestra la relación entre la cantidad de muestras comprobadas de cada clase y sus respectivos aciertos o desaciertos en la predicción.

Explicación

En este caso, la matriz de confusión nos indica que para la muestra de prueba de los 141 elementos de la clase 0 (naranjas), 36 se clasificaron como toronjas; mientras que de los 190 elementos correspondientes a las toronjas (clase 1) solamente 33 se clasificaron como naranjas.



Explicación

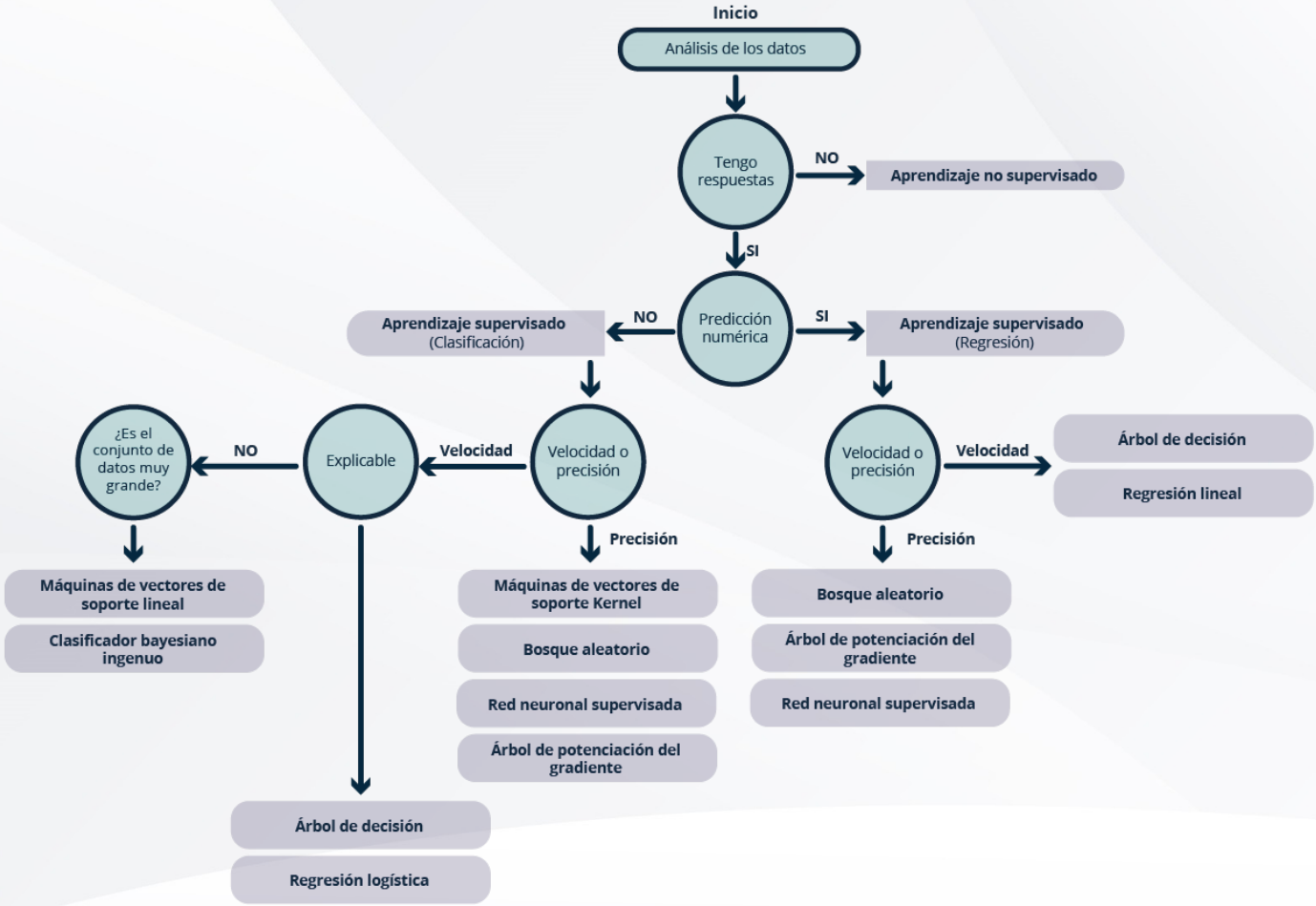
El aprendizaje supervisado se basa en la utilización de datos estructurados y en el conocimiento previo del comportamiento de un fenómeno. A partir de esta información, los problemas se pueden dividir en dos variantes principales:

- **Clasificación:** divide el conjunto de datos mediante etiquetas comunes y se encarga de analizar los patrones que identifican a cuál de dichas etiquetas pertenecen los nuevos datos que recibe el modelo. Algunos ejemplos de aplicaciones que implementan la clasificación pueden ser la detección de correos electrónicos basura, la posible decisión de compra de un cliente, la clasificación de una patología, la evaluación de un comportamiento anómalo, entre otros.

- **Regresión:** se ocupa de encontrar patrones continuos dentro de un conjunto infinito de posibles resultados. En un problema de regresión la respuesta obtenida es un valor numérico. Entre las aplicaciones típicas de la regresión se pueden mencionar las siguientes: la predicción del precio de venta de una propiedad inmobiliaria, la estimación de venta de un producto, la duración de una refracción cuando se utiliza bajo ciertas condiciones, el salario de un empleado a partir de sus resultados laborales, entre otras.

Explicación

La siguiente figura muestra una de las diversas formas de enfocar la toma de decisiones, en este caso sobre cuál algoritmo de aprendizaje automático supervisado es el más adecuado para solucionar una problemática específica:



Actividad

Después de haber estudiado el tema, aborda las siguientes cuestiones:

- ¿A qué otro tipo de conjunto de datos podrías aplicar el método de regresión logística?
- Busca ejemplos de aplicaciones que trabajen con modelos generalizados.

Cierre

En este tema aprendiste cómo los modelos lineales generalizados, por ejemplo, la regresión logística, expanden la capacidad de los modelos lineales al incorporar la capacidad de manejar datos no lineales. Para lograr este objetivo se utiliza una función base (como la función sigmoide), así como algunas transformaciones para obtener la forma lineal adecuada.

La regresión logística se utiliza principalmente para la resolución de problemas de clasificación binaria, por lo que el proceso de selección de los resultados aplica una predicción probabilística sobre el valor que estos obtienen, una vez que han sido evaluados por el modelo.

Los algoritmos de aprendizaje supervisado se enfocan en dos variantes de problemáticas principales: la regresión y la clasificación. Algunos de estos algoritmos permiten obtener resultados más precisos o rápidos, según la necesidad del problema, pero la diferencia principal entre cada una de estas perspectivas radica en la utilización dentro de los modelos (de funciones probabilísticas o no) para generar las correspondientes predicciones.

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Árboles de decisión

Introducción

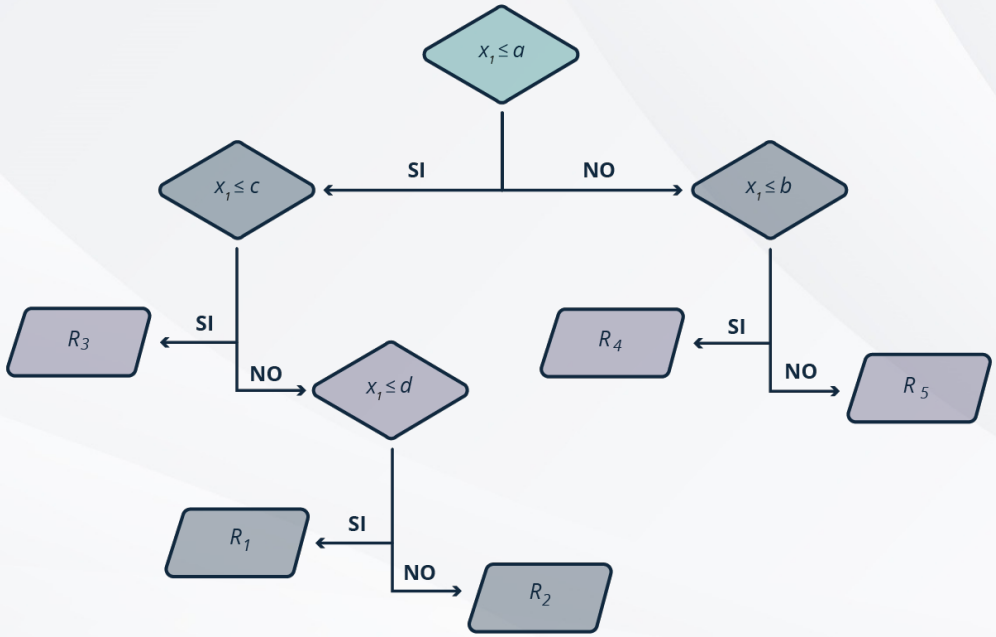
Los árboles de decisión representan conceptual y matemáticamente un enfoque diferente hacia el aprendizaje automático. Su uso presenta una gran flexibilidad porque no está restringido a datos estrictamente numéricos. Este elemento constituye una gran ventaja sobre varios de los otros métodos porque reduce en gran parte el arduo trabajo de preprocesamiento.

Por lo tanto, son estructuras heurísticas que pueden construirse mediante una secuencia de elecciones o comparaciones que se realizan en cierto orden. El motivo de la complejidad radica en determinar adecuadamente lo siguiente: ¿cómo elegir las preguntas y en qué orden? Siempre se puede comenzar a hacer preguntas aleatorias y, en última instancia, converger en la solución completa si los datos no tienen una gran dimensión. No obstante, este enfoque aleatorio o de fuerza bruta no es para nada práctico.

Explicación

Los árboles de decisión son estructuras heurísticas que pueden construirse mediante una secuencia de elecciones o comparaciones que se realizan en cierto orden.

La mayoría de los enfoques del aprendizaje automático toman como base la resolución de ecuaciones sobre las propiedades de los datos, mientras que en el caso de un árbol de decisión se parte del dibujo de una estructura de tipo árbol, de manera que en cada nodo se debe tomar una decisión.



Explicación

La utilización de los algoritmos de árbol de decisión como referencia para la solución de problemas incluye varias ventajas, entre las cuales se pueden mencionar las siguientes:

- Comportamiento más parecido al humano.
- Puede trabajar directamente con datos no numéricos, por ejemplo, categóricos.
- Puede trabajar directamente con datos faltantes, por lo que se puede omitir el paso de limpieza de datos.
- El árbol de decisión entrenado tiene una alta interpretabilidad, en comparación con la naturaleza abstracta de los modelos entrenados que utilizan otros algoritmos como redes neuronales o SVM, etc.
- Los algoritmos del árbol de decisión escalan fácilmente de datos lineales a datos no lineales sin ningún cambio en la lógica central.
- Los árboles de decisión se pueden utilizar como un modelo no paramétrico, por lo que el ajuste de hiperparámetros se vuelve innecesario.

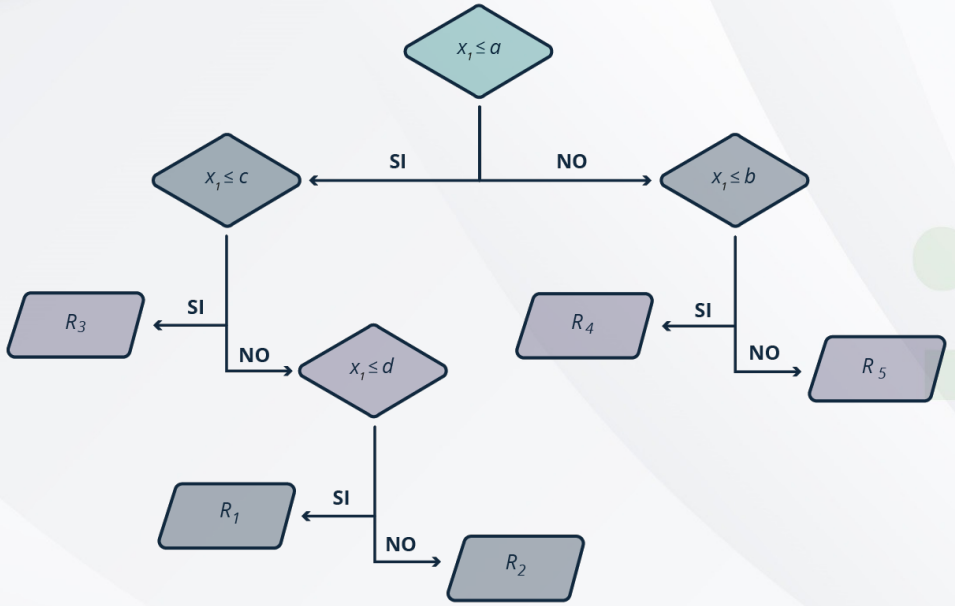
Explicación

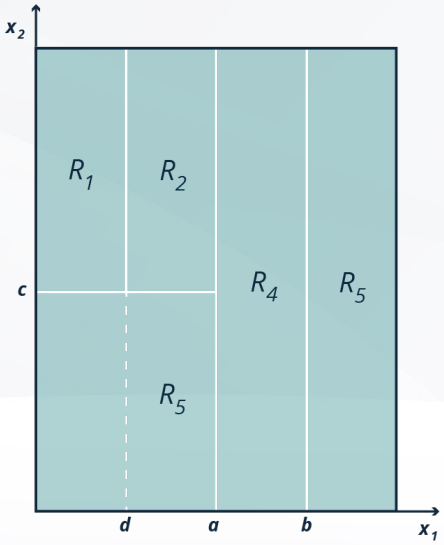
Los árboles de regresión están diseñados para predecir el valor de una función a partir de unas coordenadas dadas. Considerando un conjunto de datos de entrada n-dimensionales {xi, i = 1,..., p, donde xi ⊂ Rn}, las salidas correspondientes se definirían como {yi, i = 1,... , p, donde yi ⊂ R}. En el caso de los árboles de regresión, se requiere que los datos de entrada y salida sean numéricos y no categóricos.

Por ejemplo, con el apoyo de ambas figuras, considerando que los datos de entrada son bidimensionales y que las clases son las regiones R1, R2, R3, R4 y R5, la respuesta deseada del árbol de decisión se define como:

$$t(x) = r_k \forall x_i \in R_k$$

Donde rk ∈ R es un valor constante de la región de salida Rk.





Explicación

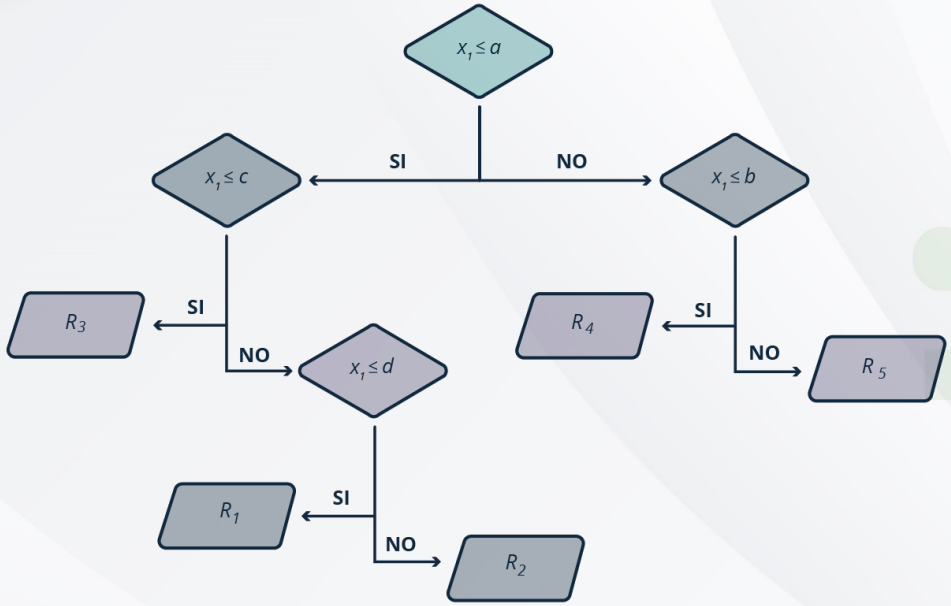
Si definimos el problema de optimización, en función de minimizar el error cuadrático medio:

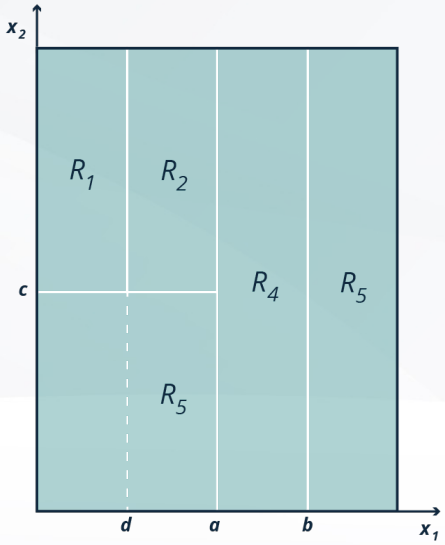
$$\sum_{i=1}^{i=p} (y_i - t(x_i)^2)$$

Entonces, el cálculo que mostraría el valor de la estimación dada por r_k viene dado por:

$$r_k = prom(y_i | x_i \in R_k)$$

Resolver el problema para encontrar las regiones óptimas globales y minimizar el error cuadrático medio es un problema muy complejo, por tanto, no se puede resolver de manera general con los métodos de tiempo finito.

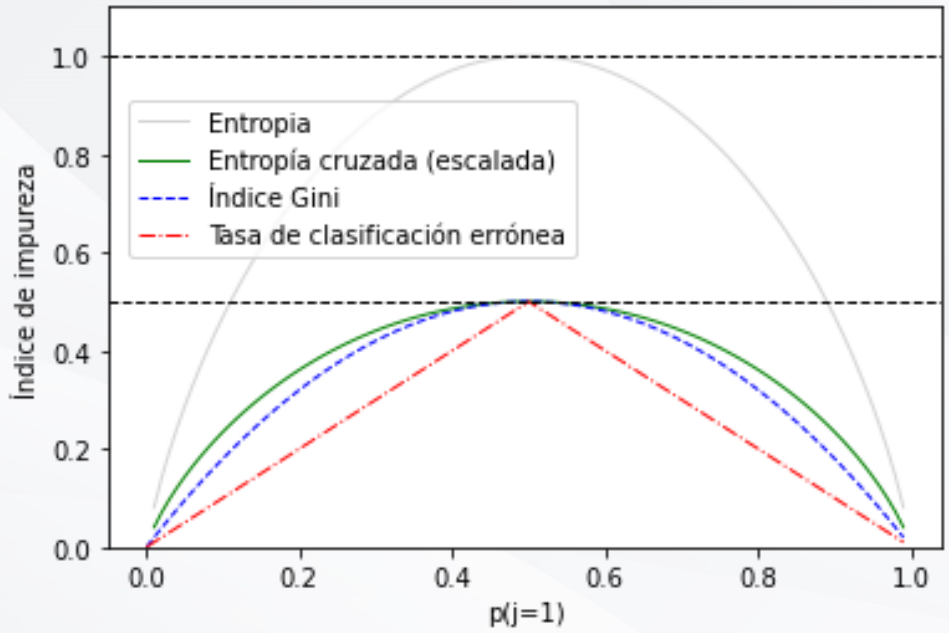




Explicación

En caso de la clasificación, la salida no es un valor numérico continuo, sino una etiqueta de clase discreta. El desarrollo del árbol sigue los mismos pasos que se mencionaron para su homólogo de regresión, pero en este caso los métodos de poda deben ser diferentes, ya que el método del error cuadrático medio no es adecuado para la clasificación. Existen tres tipos de métricas populares para realizar este proceso:

- Tasa de clasificación errónea.
- Índice de Gini.
- Entropía cruzada o desviación.

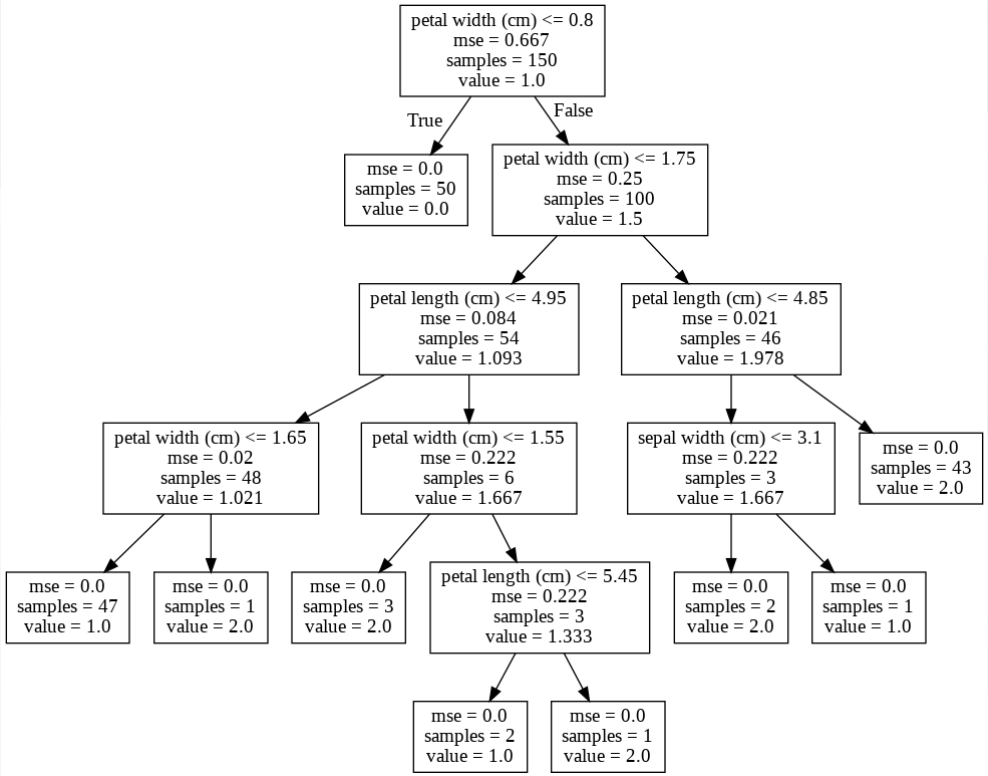


Explicación

A continuación, se verán dos ejemplos muy sencillos, pero bastante ilustrativos, realizados con Python y la librería Scikit-learn, que describen el proceso de construcción de un árbol de decisión con los dos enfoques particulares (regresión o clasificación).

Scikit-learn incluye varios conjuntos de datos para realizar pruebas y experimentos. Uno de los más populares por su versatilidad y claridad es el referente a las flores de iris, en el cual, a partir de las mediciones del ancho y largo de los pétalos y sépalos de la flor, define tres categorías de estas (iris setosa, iris versicolor e iris virginica).

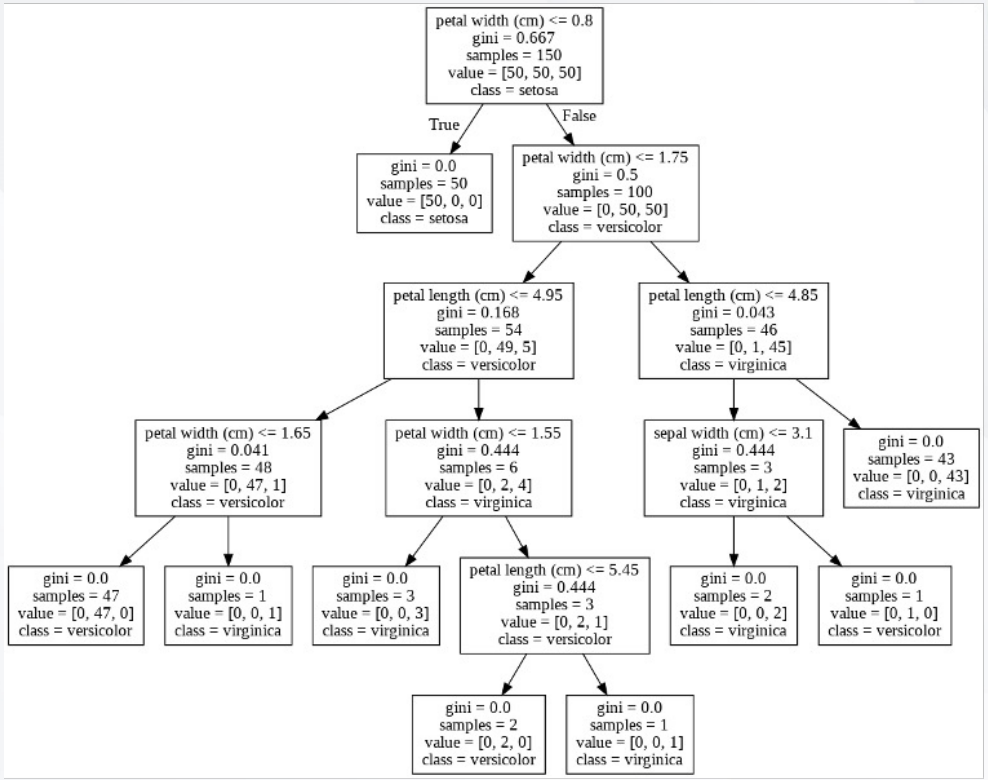
La figura representa el resultado de aplicar el algoritmo de un árbol de regresión.



Explicación

Se puede construir un árbol de clasificación utilizando como base el mismo conjunto de datos de la flor. En este caso, el modelo importado desde Scikit-learn corresponde a *DecisionTreeClassifier,* el cual utiliza por defecto la métrica Gini para efectuar la toma de decisión en cada nodo correspondiente. Si quisiéramos evaluar una nueva muestra, solamente se tendría que invocar el método *predict*() del modelo elaborado antes.

Utilizando un código similar al mostrado en el ejemplo del árbol de regresión, se puede generar el gráfico correspondiente al árbol de clasificación.



Explicación

En muchas situaciones, operar con un solo árbol funciona muy bien, pero es posible extraer más rendimiento de esta arquitectura si se agrupan varios de estos árboles y se combinan en una sola solución. Estas técnicas se denominan métodos de conjunto, por lo que ofrecen un rendimiento superior a costa de la complejidad del cálculo y del algoritmo. Existen tres tipos principales de conjuntos:

- **Empaquetado (*bagging*):** también conocida como agregación de bootstrap, es uno de los métodos que más se utilizan para combinar árboles de decisión.
- **Bosque aleatorio:** el proceso de empaquetado mejora la resiliencia de los árboles de decisión respecto a los valores atípicos.
- **Árboles de conjunto potenciados:** la diferencia fundamental entre el conjunto potenciado y el empaquetado (o bosque aleatorio para el mismo caso) es el entrenamiento secuencial de los árboles, en vez del entrenamiento paralelo que realiza este último.

Actividad

Piensa en otro problema en donde puedas aplicar el método de árbol de regresión y clasificación.

Revisa los algoritmos e idea un plan para aplicarlo al problema que escogiste. Por tanto, reemplaza y adapta las variables necesarias para buscar una solución.

Cierre

En este tema estudiaste el concepto de los árboles de decisión y sus diferentes variaciones. Este tipo de modelos está directamente motivado por el proceso jerárquico de toma de decisiones, muy similar al comportamiento humano al abordar problemas de la vida real, por lo que son más intuitivos que los otros métodos. Con ellos se pueden resolver problemas de regresión, es decir, cuando un valor de salida requiere una evaluación numérica y la entrada es una función continua, pero también se pueden utilizar para realizar clasificaciones a partir de entradas discretas, entregando resultados categóricos.

El apoyo de un complemento visual no solamente muestra la forma de encontrar una solución, sino que también comparte una manera lógica de comprender un fenómeno cuando la información que lo describe es demasiado compleja.

Como punto final, conociste los métodos de conjunto, los cuales utilizan el agregado de múltiples árboles de decisión para optimizar el rendimiento general para que los modelos sean más robustos y genéricos.

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
