MAGISTER EN ESTADÍSTICA

ÁRBOLES DE DECISIÓN

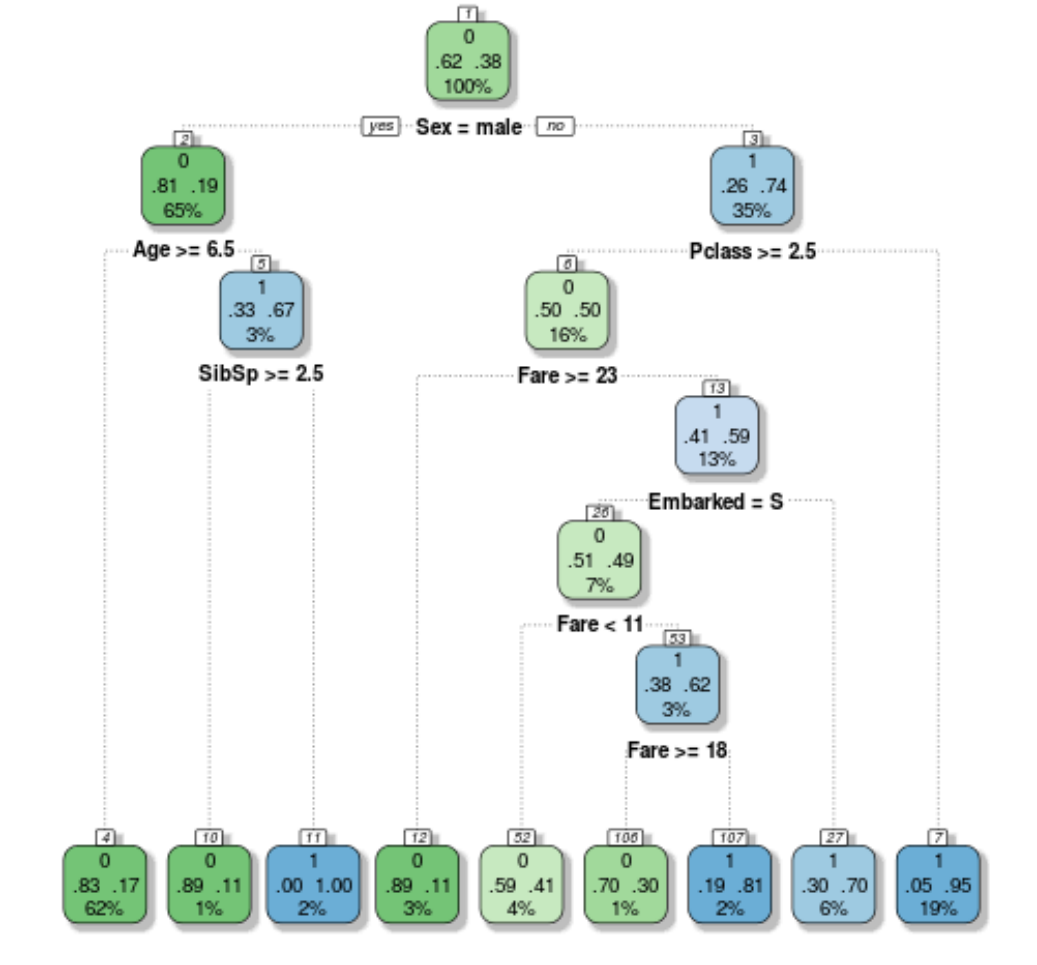
Agenda

1. Introducción
2. Tipos de árboles
3. Árboles de regresión
   - Construcción del árbol
   - Predicción
4. Árboles de clasificación
   - Construcción del árbol
   - Poda del árbol
   - Post-poda

Introducción

Un árbol de decisión es un conjunto de condiciones en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Introducción



Introducción

Ventajas

- La arquitectura de los árboles es simple de explicar, dado que es un conjunto de reglas predictivas.
- En los árboles se pueden incluir las variables categoricas sin problema de realizar una mala clasificación.
- Los datos outliers y los datos faltantes no afectan considerablemente al modelo.
- Los árboles pueden manejar tanto predictores cuantitativos como cualitativos sin tener que crear variables dummy.
- Al tratarse de métodos no paramétricos, no es necesario que se cumpla ningún tipo de distribución específica.

Introducción

Desventajas

- En lo árboles se corre el riesgo de sobreajustar el modelo.
- No son robustos, esto quiere decir que si se cambian los datos de entrenamiento (*train*) el conjunto de reglas puede alterarse sustancialmente.
- Cuando tratan con variables continuas, pierden parte de su información al categorizarlas en el momento de la división de los nodos. Por esta razón, suelen ser modelos que consiguen mejores resultados en clasificación que en regresión.

Tipos de árboles

Existen dos métodos de árboles, la diferencias entre ellos viene dada por la variable respuesta.

- **Árboles de regresión**: son el subtipo de árboles de predicción que trabaja con variables respuesta continuas.
- **Árboles de clasificación**: se asemejan mucho a los árboles de regresión, con la diferencia de que predicen variables respuesta cualitativas en lugar de continuas.

Árboles de regresión

Construcción del árbol

El proceso de construcción de un árbol de predicción (regresión o clasificación) se divide en dos etapas:

- División sucesiva del espacio de los predictores generando regiones no solapantes (nodos terminales) $R_1, R_2, R_3, \ldots, R_j$. Aunque, desde el punto de vista teórico las regiones podrían tener cualquier forma, si se limitan a regiones rectangulares (de múltiples dimensiones), se simplifica en gran medida el proceso de construcción y se facilita la interpretación.
- Predicción de la variable respuesta en cada región.

Árboles de regresión

Construcción del árbol

el criterio más frecuentemente empleado para identificar las divisiones es el Residual Sum of Squares (RSS). El objetivo es encontrar las J regiones $(R_1, \ldots, R_j)$ que minimizan el Residual Sum of Squares (RSS) total:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})^2$$

donde $\hat{y}_{Rj}$ es la media de la variable respuesta en la región R_j.

Árboles de regresión

División binaria recursiva

es encontrar en cada iteración el predictor X_j y el punto de corte (umbral) s tal que, si se distribuyen las observaciones en las regiones $\{X|X_j < s\}$ y $\{X|X_j \geq s\}$, se consigue la mayor reducción posible en el RSS. El algoritmo seguido es:

- El proceso se inicia en lo más alto del árbol, donde todas las observaciones pertenecen a la misma región.
- Se identifican todos los posibles puntos de corte (umbrales) s para cada uno de los predictores $(X_1, X_2, \ldots, X_p)$. En el caso de predictores cualitativos, los posibles puntos de corte son cada uno de sus niveles. Para predictores continuos, se ordenan de menor a mayor sus valores, el punto intermedio entre cada par de valores se emplea como punto de corte.

Árboles de regresión

División binaria recursiva

- Se calcula el RSS total que se consigue con cada posible división identificada en el paso 2.

$$RSS_{total} = RSS_1 + RSS_2$$

- Se selecciona el predictor X_j y el punto de corte S que resulta en el menor RSS total, es decir, que da lugar a las divisiones más homogéneas posibles. Si existen dos o más divisiones que consiguen la misma mejora, la elección entre ellas es aleatoria.
- Se repiten de forma iterativa los pasos 1 a 4 para cada una de las regiones que se han creado en la iteración anterior hasta que se alcanza alguna norma de stop. Algunas de las más empleadas son: que ninguna región contenga un mínimo de n observaciones, que el árbol tenga un máximo de nodos terminales o que la incorporación del nodo reduzca el error en al menos un % mínimo.

Árboles de regresión

Predicción: Se recorre el árbol en función de los valores que tienen sus predictores hasta llegar a uno de los nodos terminales. En el caso de regresión, el valor predicho suele ser la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo.

Árboles de regresión

Construcción del árbol

La construcción de los árboles de clasificación poseen una construcción a los árboles de regresión, con la diferencia que no es posible utilizar el método de RSS dado que la variable objetivo en este caso es binaria. Es por esto, que las métricas de selección de particiones vienen dada por:

- **Tasa de error de clasificación**: Se define como la proporción de observaciones que no pertenecen a la clase más común en el nodo.

$$E_m = 1 - max_k(\hat{p}_{mk}),$$

- **Índice de Gini**: Es una medida de la varianza total en el conjunto de las K clases del nodo m. Se considera una medida de pureza del nodo.

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Cuando $\hat{p}_{mk}$es cercano a 0 o a 1 (el nodo contiene mayoritariamente observaciones de una clase), el término $\hat{p}_{mk}(1-\hat{p}_{mk})$ es muy pequeño. Como consecuencia, cuanto mayor sea la pureza del nodo, menor el valor del índice Gini G.

Árboles de clasificación

Construcción del árbol

- **Entropía cruzada**: es otra forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se corresponde con la impureza. Si un nodo es puro, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo de 1.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \cdot log(\hat{p}_{mk})$$

Árboles de clasificación

Construcción del árbol

- **Entropía cruzada**: es otra forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se corresponde con la impureza. Si un nodo es puro, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo de 1.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \cdot log(\hat{p}_{mk})$$

Árboles de clasificación

Construcción del árbol

- **Ji-cuadrado**: Esta aproximación consiste en identificar si existe una diferencia significativa entre los nodos hijos y el nodo parental, es decir, si hay evidencias de que la división consigue una mejora. Para ello, se aplica un test estadístico ji-cuadrado de bondad de ajuste empleando como distribución esperada H_0 la frecuencia de cada clase en el nodo parental. Cuanto mayor el estadístico χ^2, mayor la evidencia estadística de que existe una diferencia.

$$\chi^2 = \sum_k \frac{(\text{observado}_k - \text{esperado}_k)^2}{\text{esperado}_k}$$

Árboles de clasificación

Independientemente de la medida empleada como criterio de selección de las divisiones, el proceso siempre es el mismo:

- Para cada posible división se calcula el valor de la medida en cada uno de los dos nodos resultantes.
- Se suman los dos valores ponderando cada uno por la fracción de observaciones que contiene cada nodo. Este paso es muy importante, ya que no es lo mismo dos nodos puros con 2 observaciones, que dos nodos puros con 100 observaciones.

$$\frac{\text{n observaciones nodo A}}{\text{n observaciones totales}} \times \text{pureza A} + \frac{\text{n observaciones nodo B}}{\text{n observaciones totales}} \times \text{pureza B}$$

- La división con menor o mayor valor (dependiendo de la medida empleada) se selecciona como división óptima.

Poda del árbol

El proceso de construcción de árboles tiende a reducir el error de entrenamiento, por lo que el modelo se ajusta bien a las observaciones definidas en la data de entrenamiento. Como consecuencia, se genera un sobre-ajuste que reduce la capacidad predictiva al momento de testear con nuevas observaciones (data test). Para esto existen dos formas:

- Controlar el tamaño del árbol.
- La poda del árbol.

Poda del árbol

Controlar el tamaño del árbol

El tamaño final que adquiere un árbol puede controlarse mediante reglas de parada que detengan la división de los nodos dependiendo de si se cumplen o no determinadas condiciones.

- **Observaciones mínimas para división**: define el número mínimo de observaciones que debe tener un nodo para poder ser dividido. Cuanto mayor el valor, menos flexible es el modelo.
- **Observaciones mínimas de nodo terminal**: define el número mínimo de observaciones que deben tener los nodos terminales. Su efecto es muy similar al de observaciones mínimas para división.

Poda del árbol

Controlar el tamaño del árbol

- **Profundidad máxima del árbol**: define la profundidad máxima del árbol, entendiendo por profundidad máxima el número de divisiones de la rama más larga (en sentido descendente) del árbol.
- **Número máximo de nodos terminales**: define el número máximo de nodos terminales que puede tener el árbol. Una vez alcanzado el límite, se detienen las divisiones. Su efecto es similar al de controlar la profundidad máxima del árbol.
- **Reducción mínima de error**: define la reducción mínima de error que tiene que conseguir una división para que se lleve a cabo.

Poda del árbol

Poda

Consiste en generar árboles grandes, sin condiciones de parada más allá de las necesarias por las limitaciones computacionales, para después podarlos, manteniendo únicamente la estructura robusta que consigue un test error bajo. La selección del sub-árbol óptimo puede hacerse mediante validación cruzada, sin embargo, dado que los árboles se crecen lo máximo posible (tienen muchos nodos terminales) no suele ser viable estimar el test error de todas las posibles sub-estructuras que se pueden generar.

Poda del árbol

Poda de complejidad de costos

Es un método de penalización de tipo perdida + penalización, similar al empleado en ridge regression o lasso. En este caso, se busca el sub-árbol T que minimiza la ecuación:

$$\sum_{j=1}^{|T|} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T|$$

donde |T| es el número de nodos terminales del árbol. A medida que va aumentando el valor de α la penalización es mayor. Por el contrario si α es cercano a cero es casi nula la penalización en la poda.

Construcción de árbol con poda

1. Se emplea recursive binary splitting para crear un árbol grande y complejo (T_o) empleando los datos de entrenamiento y reduciendo al máximo posible las condiciones de parada. Normalmente se emplea como única condición de parada el número mínimo de observaciones por nodo terminal.
2. Se aplica el poda de complejidad de costos de árbol T_o para obtener el mejor sub-árbol en función de α. Es decir, se obtiene el mejor sub-árbol para un rango de valores de α.
3. Identificación del valor óptimo de α mediante k-cross-validation. Se divide el training data set en K grupos. Para $k = 1, \ldots, k = K$:
   - Repetir pasos 1 y 2 empleando todas las observaciones excepto las del grupo k_i.
   - Evaluar el mean squared error para el rango de valores de α empleando el grupo k_i.
   - Obtener el promedio de los K mean squared error calculados para cada valor α.
4. Seleccionar el sub-árbol del paso 2 que se corresponde con el valor α que ha conseguido el menor validación cruzada mean squared error en el paso 3.