

Importancia de características

Aprendizaje Automático

Juan David Martínez

jdmartinev@eafit.edu.co

Agenda

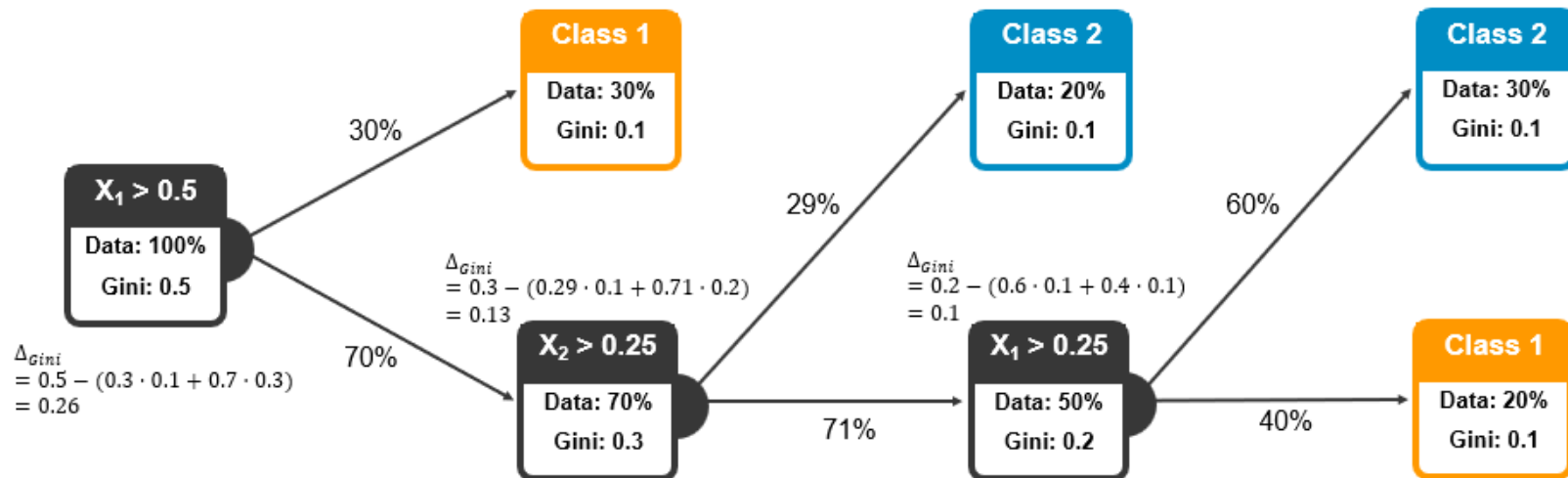
- Importancia de características en árboles
- Importancia de características en ensambles

Importancia de características

- Los métodos basados en árboles tienen un beneficio:
Evalúan explícitamente la importancia de cada una de las variables en el proceso de toma de decisiones.
- Esto se puede utilizar para seleccionar las características más importantes para entrenar un modelo o para interpretar el modelo

Importancia de características

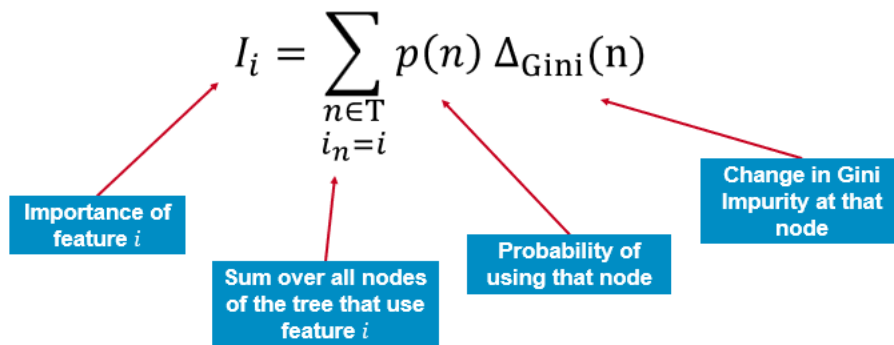
- Importancia de características en un árbol



Importancia de características

- A medida que se iba construyendo el árbol, se seleccionaban las particiones que maximizaban la ganancia de información
- Podemos asociar a cada característica una ganancia de información:

Para cada nodo asociado a esa característica, sume la fracción de los datos que pasa por dicho nodo ponderado por la ganancia de información

$$I_i = \sum_{\substack{n \in T \\ i_n = i}} p(n) \Delta_{\text{Gini}}(n)$$


Importance of feature i

Sum over all nodes of the tree that use feature i

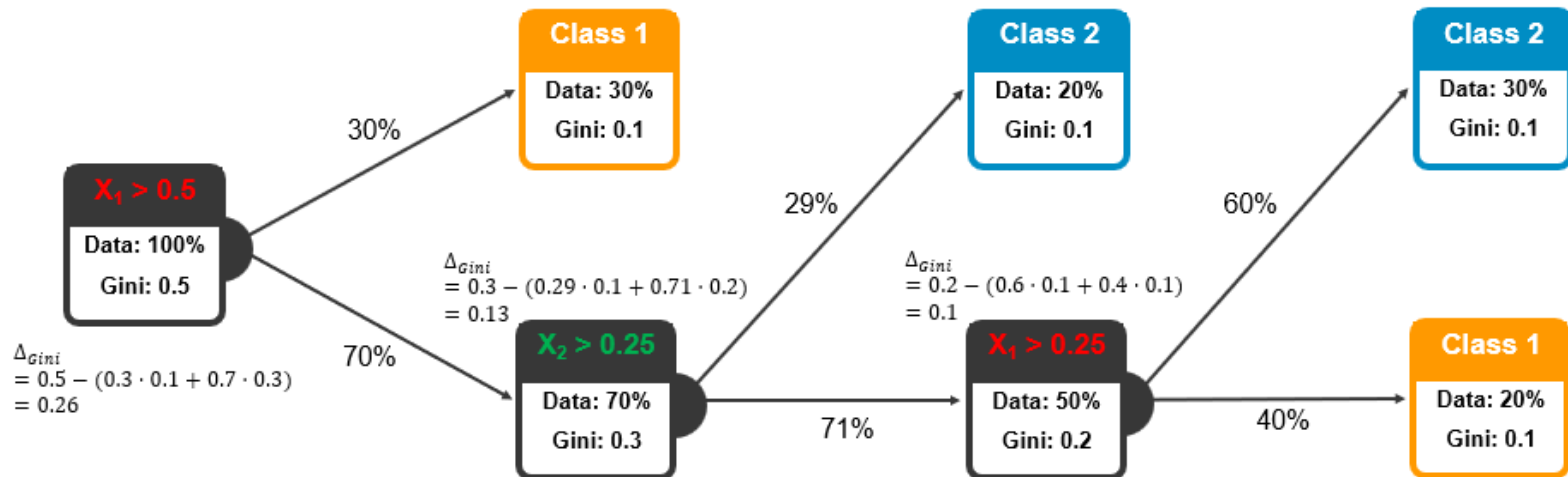
Probability of using that node

Change in Gini Impurity at that node

Importancia de características

$$I_1 = 1.0 \cdot 0.26 + 0.5 \cdot 0.1 = 0.31$$

$$I_2 = 0.7 \cdot 0.13 = 0.09$$



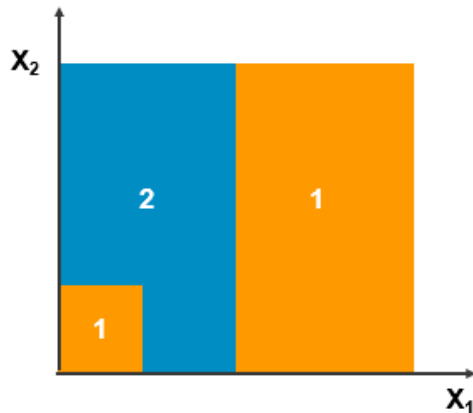
Importancia de características

- Estas importancias nos dicen la importancia relativa de las dos características

Importancia de $x_1 = 0.31$

Importancia de $x_2 = 0.09$

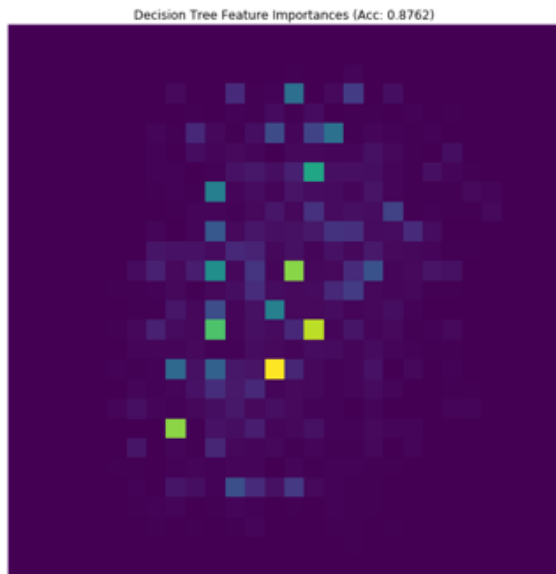
- Si graficamos el árbol nos damos cuenta de que este resultado tiene sentido



x_1 has a much larger
impact on the final class
than x_2

Limitaciones

- Depende del árbol y no de las predicciones (dos árboles que hagan las mismas predicciones pueden generar diferentes importancias)
- Puede ser confuso si tenemos variables correlacionadas



Importancia de características RF

- Promediamos sobre todos los árboles la importancia de la característica en cada árbol
- Ayuda a resolver los problemas mencionados

$$I_i = \frac{1}{|B|} \sum_{T \in B} I_i(T)$$

Limitaciones

- Las variables continuas a menudo reportan mayores valores de importancia que las discretas (mayor número de particiones, mayor número de apariciones en el árbol)
- Las variables discretas de más clases reportan mayores valores de importancia que variables discretas con menos clases. Este [artículo](#) y este [blog](#) hablan de esto.
- Algunas veces, a variable aleatoria continua se ve más relevante que una variable discreta realmente importante.
- Colinealidad: Las columnas correlacionadas compartirán relevancia.