1 Árvores de decisão

Árvores de decisão são métodos de aprendizado de máquina utilizados em problemas de classificação e regressão. Para computação, árvores são estruturas de dados, ou seja, armazenam informações. No caso de uma árvore de decisão, essa estrutura de dados segue regras através de seus nós até os nós folhas(nós que não possuem filhos) que indicam a decisão a ser tomada pelo algoritmo.
Em problemas de classificação, a construção de uma árvore de decisão consiste em particionar as classes relacionadas com seus atributos no espaço, recursivamente, de modo a atingir a pureza nas sub-regiões criadas. Nesse caso, a pureza indica a homogeneidade da classe na região. Então, a partir dessas sub-regiões são criadas as regras a serem seguidas para tomada de decisão da árvore.

A fim de encontrar os melhores pontos para particionar espaço, temos como interesse medidas de impureza, tais como, Gini Index, Entropia e ganho de informação.

2 Entropia

A entropia está relacionada com a ordem ou desordem dos dados, evidenciando como os dados estão divididos. O valor da entropia varia entre 0 e 1. Sendo que quanto maior a entropia maior a desordem dos dados e quanto menor, maior a ordem dos dados.
A entropia é calculada através da fórmula:

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Onde p é a probabilidade das classes.

3 Ganho de Informação

O ganho de informação é uma medida utilizada para determinar qual atributo nos retorna a maior quantidade de informação sobre uma classe. Essa medida é baseada no cálculo da entropia, a redução da entropia indica um ganho informacional.

4 Gini Index

O Gini Index é uma medida de impureza que mede a probabilidade de uma determinada variável ser classificada erroneamente quando escolhida de forma aleatória. Essa medida nos ajuda a escolher o melhor ponto de "corte" para a

criação das sub-regiões citadas anteriormente, assim, também auxilia na construção da árvore, pois definindo os pontos de "corte" os atributos mais importantes para a tomada de decisão também são definidos. O Gini Index é calculado através da fórmula:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Onde p é a probabilidade de um objeto ser classificado em determinada classe.