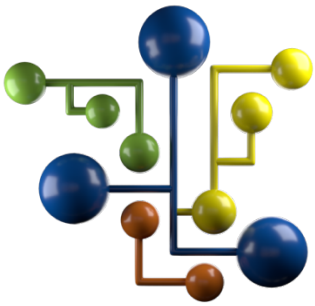

Data Science Academy

www.datascienceacademy.com.br

Machine Learning

Compreendendo a Heurística ID3

As árvores de decisão são conhecidas como uma das mais poderosas e amplamente utilizadas técnicas de modelagem de aprendizado de máquina. As árvores de decisão naturalmente podem induzir regras que podem ser usadas para classificação de dados ou para realizar previsões.

O que são Heurísticas?

Heurísticas são soluções baseadas em algum tipo de conhecimento prévio sobre as propriedades dos dados, na procura de uma boa solução (mas não necessariamente a melhor). A pesquisa por heurísticas é uma pesquisa realizada por meio da quantificação de proximidade a um determinado objetivo.

Diz-se que se tem uma boa (ou alta) heurística se o objeto de avaliação está muito próximo do objetivo; dizemos que a heurística é ruim (ou baixa) se o objeto avaliado estiver muito longe do objetivo. A palavra heurística significa descobrir (e que deu origem também ao termo Eureca!, que vc já deve ter ouvido muitas e muitas vezes). Heurísticas de construção, tais como o método guloso (que vimos nos vídeos anteriores), são aquelas onde uma ou mais soluções são construídas elemento a elemento, seguindo algum critério heurístico de otimização, até que se tenha uma solução viável; O algoritmo ID3 possui uma heurística de árvore de decisão.

Este algoritmo segue os seguintes passos:

1- Começa com todos os exemplos de treino.

2- Escolhe o teste (atributo) que melhor divide os exemplos, ou seja, agrupa exemplos da mesma classe ou exemplos semelhantes. Nesta etapa, para achar o melhor atributo é necessário encontrar a entropia para cada atributo possível naquele nó. Isso é feito pelo algoritmo. Atributo com maior ganho de informação é selecionado para ser raiz da árvore de decisão.

3- Para o atributo escolhido, é criado um nó filho para cada valor possível do atributo. O próximo passo na heurística ID3 é calcular o ganho de informação para cada atributo que pode ser selecionado como nó na árvore. Essencialmente é apenas calcular a entropia de todo o conjunto de dados e diminuir este da entropia do sub-conjunto particionado para tal atributo (aquela fórmula que mostramos a você nos vídeos anteriores). Este processo é feito para cada atributo do conjunto de dados, e o atributo com o maior ganho de informação será o selecionado para o próximo nó da árvore.

4- Transporta os exemplos para cada filho considerando o valor do filho.

5- Repete o procedimento para cada filho não "puro". Um filho é puro quando cada atributo X tem o mesmo valor em todos os exemplos.

E como o algoritmo sabe o melhor atributo a escolher? Através do Ganho de Informação e Entropia!! A heurística ID3 usa o conceito de entropia para formular o ganho de informação na escolha de um atributo particular para ser o próximo nó na árvore de decisão.

A Heurística ID3

A física usa o termo entropia para descrever a quantidade de desordem associada a um sistema. Na teoria da informação, este termo tem um significado semelhante -- ele mede o grau de desordem de um conjunto de dados. A heurística ID3 usa este conceito para encontrar o próximo melhor atributo de um dado para ser utilizado como nó de uma árvore de decisão.

Logo, a ideia por trás do algoritmo ID3 é achar um atributo que reduza em maior valor a entropia de um conjunto de dados, assim reduzindo a aleatoriedade - dificuldade de previsão - da variável que define classes. Seguindo esta heurística, você estará essencialmente encontrando o melhor atributo para classificar os registros (de acordo com a redução da quantia de informação necessária para descrever a partição dos dados que foram divididos) a fim de que os mesmos tenham utilidade máxima (exemplos são da mesma classe).

Espaço de Hipóteses do ID3

O ID3 busca no espaço de hipóteses alguma que seja adequada para representar o conjunto de treinamento. Esse espaço de hipóteses é formado por um conjunto de todas possíveis árvores de decisão. O ID3 começa com árvore vazia e progressivamente elabora hipóteses até chegar em uma árvore de decisão.

A busca por hipóteses é guiada pelo Ganho de Informação dos atributos. Como ID3 não mantém todas hipóteses consistentes com o conjunto de treinamento, o ID3 não tem a habilidade de determinar quantas árvores de decisão alternativas são consistentes com os dados de treinamento. O ID3 também não realiza backtracking na busca, ou seja, uma vez que tenha selecionado um atributo, não reavalia a árvore de decisão formada. O ID3 emprega todos os dados de treinamento em cada passo da busca por hipóteses e toma decisões estatísticas em cada passo. A abordagem do ID3 privilegia árvores mais curtas em relação às mais longas observadas e atributos de maior Ganho de Informação mais próximos do topo ou raiz da árvore.

O ID3 deu origem a outros algoritmos semelhantes, os quais descrevemos abaixo e que serão estudados mais adiante.

ID3 (Iterative Dichotomizer 3) - O ID3 é um algoritmo usado para gerar árvores de decisão. Os atributos do conjunto de dados devem ser obrigatoriamente categóricos.

C4.5 - Um algoritmo para geração de árvores de decisão, sucessor do algoritmo ID3. O algoritmo C4.5 considera atributos numéricos e categóricos.

C5.0 - O C5.0 foi inicialmente uma versão comercial, que aliás ainda é comercializado (link na seção de links úteis), mas que teve o código aberto para a comunidade open-source. Também uma evolução do ID3.

CART (Classification and Regression Trees) - Técnica não-paramétrica que produz árvores de classificação ou regressão, dependendo se as variáveis são categóricas ou numéricas, respectivamente.