INSTITUTO FEDERAL
Ceará

Aprendizado Supervisionado

Classificação por Árvores de Decisão

Prof. Raphael Carvalho

Árvore de Decisão

- É uma representação gráfica de regras de classificação
- Estas regras demonstram visualmente as condições para categorizar dados por meio de uma estrutura que contém nó raiz, nós folha e nós finais.
- É possível atravessar a árvore de decisão partindo do nó raiz até cada folha por meio de diversas regras de decisão – lembrando que o destino final (nó folha) contém sempre uma das classes pré-definidas.

Árvore de Decisão

- Uma árvore de decisão pode ser utilizada tanto para classificação quanto para regressão
- Na classificação, as classes são categóricas e finitas
- Na regressão, são contínuas e infinitas

Árvores de Decisão

- Cada nó interno denota um teste de um atributo
- Cada ramificação denota o resultado de um teste
- Cada nó folha apresenta um rótulo de uma classe definida de antemão por um supervisor

Árvores de Decisão

- O objetivo dessa técnica é criar uma árvore que verifica cada um dos testes até chegar a uma folha, que representa a categoria, classe ou rótulo do item avaliado

O que isso tem a ver com aprendizado de máquina?

- A árvore de decisão, por si só, não tem a ver com aprendizado de máquina...

- No entanto, seu processo de construção automático e recursivo a partir de um conjunto de dados pode ser considerado um algoritmo de aprendizado de máquina.
- O processo de construção do modelo de uma árvore de decisão se chama indução e busca fazer diversas divisões ou particionamentos dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos.

Como assim?

Imagine o seguinte....

- Temos uma tabela com diversas linhas e colunas, em que as linhas representam pessoas que desejam obter um cartão de crédito e as linhas representam atributos ou variáveis dessas pessoas.
- Para que uma operadora de cartão decida se vai disponibilizar um cartão de crédito para uma pessoa ou não, ela deve avaliar qual é o risco de tomar um calote dessa pessoa.
- Logo, a nossa árvore de decisão buscará analisar variáveis para classificar o risco de calote de uma pessoa

Como isso é feito?

- Sabemos que as árvores de decisão são uma das ferramentas do algoritmo de classificação e que algoritmos de classificação são supervisionados.
- Podemos concluir que a árvore de decisão necessita de um supervisor externo para treinar o algoritmo e indicar quais serão as categorias que ele deve classificar uma pessoa.
- Para o nosso exemplo, vamos assumir que as categorias/classes sejam:
  - Risco Baixo, Risco Médio ou Risco Alto

Como isso é feito?

- O algoritmo de aprendizado de máquina vai analisar um conjunto de variáveis ou atributos de diversas pessoas e categorizá-las em uma dessas três classes possíveis.

E como o algoritmo vai descobrir quais são as variáveis mais importantes?

- Não é uma tarefa simples –ainda mais se existirem muitas variáveis

Como isso é feito?

- A idade é mais relevante para definir o risco de calote de uma pessoa do que seu salário anual?
- O estado civil é mais relevante para definir o risco de calote de uma pessoa do que seu saldo em conta? ● Em que ordem devemos avaliar cada variável? ● Para humanos, essas perguntas são extremamente difíceis de responder, mas é aí que entre em cena o aprendizado de máquina!

- Existe uma lógica interna de construção da árvore de decisão que automaticamente pondera a contribuição de cada uma das variáveis com base em dados históricos
- Podemos fazer a máquina aprender oferecendo para ela uma lista que contenha os valores dessas variáveis referentes a diversos clientes antigos e uma coluna extra que indique se esses clientes deram calote na operadora ou não

Aprendizado de Máquina

- O algoritmo da árvore de decisão analisará cada uma dessas variáveis (e seus possíveis pontos de corte) a fim de descobrir quais são as melhores para realizar o particionamento dos dados de modo que se formem dois subgrupos mais homogêneos possíveis

Exemplo

- O algoritmo fará diversos testes e poderá começar analisando a variável de estado civil
- Separa as pessoas em dois subgrupos (casados e não-casados) e verifica qual é a porcentagem de casados caloteiros e não-casados caloteiros
- Depois ele pode analisar a idade: separa em dois subgrupos (< 25 anos e >= 25 anos) e verifica qual é a porcentagem de < 25 anos caloteiros e >= 25 anos caloteiros.

Exemplo

- Quando ele dividiu a primeira variável em casados caloteiros e não-casados caloteiros, ele pode ter descoberto que havia muito mais não-casados caloteiros do que casados caloteiros.
- Logo, o fato de uma pessoa ser casada tende a reduzir seu risco de calote

- O algoritmo vai fazer milhares de testes com cada uma das variáveis, vai testar diversos pontos de corte diferentes e diversas sequências de análise de variáveis diferentes.
- De modo que vamos sair de um grupo muito misturado (menos homogêneo) para dois subgrupos menos misturados (mais homogêneos)
- Ganho de Informação ou Redução de Entropia

O que é entropia?

O que é entropia?

- Entropia é a uma medida que nos diz o quanto um conjunto de dados está desorganizado ou misturado.
- Toda vez que particionamos os dados em subgrupos, obtemos dados mais homogêneos e organizados, logo reduzimos a entropia.
- O que a árvore de decisão busca fazer é pegar um conjunto de dados e encontrar um conjunto de regras sobre variáveis ou pontos de corte que permite separar esses dados em grupos mais homogêneos

E qual é o resultado?

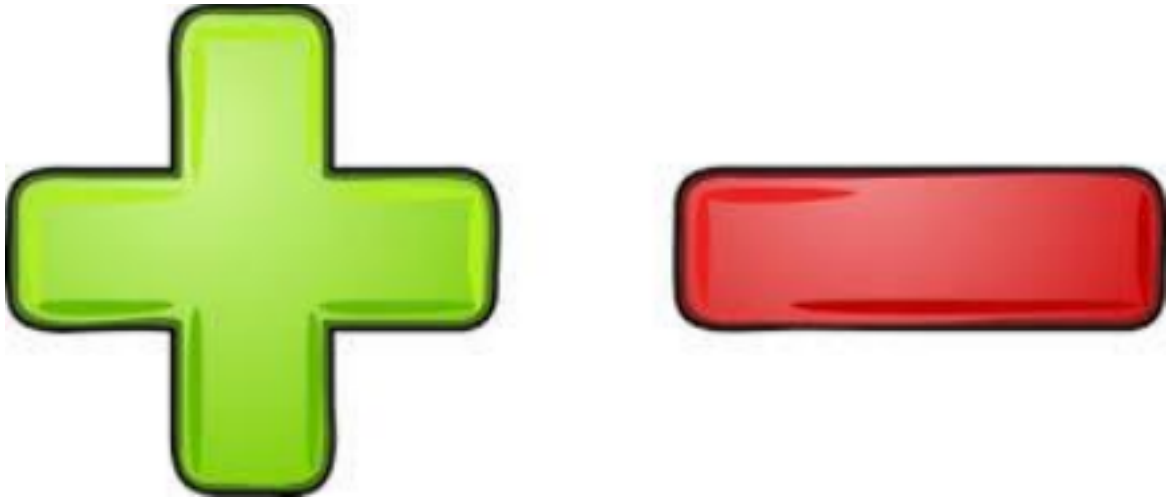
Entendendo os resultados

- O processo de construção do modelo de uma árvore de decisão busca fazer diversas divisões dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos
- E se tivermos muitas variáveis?
- Quando o algoritmo vai parar de dividir?

Entendendo os resultados

- Em tese, ele pode ir dividindo, dividindo, dividindo indefinidamente até que – no pior caso – tenhamos um único dado para cada nó folha
- O nome desse fenômeno é *overfitting*
- Deve ser evitado porque pode tornar o modelo de árvore de decisão completamente inútil.
- Para evitar, é necessário estabelecer limitar as divisões.
    - Definir uma altura/profundidade máxima da árvore

Vantagens x Desvantagens



Vantagens

- As árvores de decisão podem gerar regras

compreensíveis e executam a classificação sem exigir muitos cálculos, sendo capazes de lidar com variáveis contínuas e categóricas.

- As árvores de decisão fornecem uma indicação clara de quais campos são mais importantes para predição ou classificação.

Vantagens

- As árvores de decisão são menos apropriadas para

tarefas de estimativa em que o objetivo é prever o valor de um atributo contínuo.

- As árvores de decisão estão sujeitas a erros em problemas de classificação com muitas classes e um número relativamente pequeno de exemplos de treinamento.

Desvantagens

- Árvores de decisão são bastante propensas ao

overfitting dos dados de treinamento

- Uma única árvore de decisão normalmente não faz grandes previsões, portanto várias árvores são frequentemente combinadas em forma de florestas chamadas *Random Forests*.
- Somente se as informações forem precisas e exatas, a árvore de decisão fornecerá resultados promissores. 25

Desvantagens

- Mesmo se houver uma pequena alteração nos dados de entrada, isso pode causar grandes alterações na árvore. ●

Se o conjunto de dados é enorme, com muitas colunas e linhas, é uma tarefa muito complexa projetar uma árvore de decisão com muitos ramos.

- Se uma das regras do modelo estiver incorreta, isso gerará divisões equivocadas da árvore, fazendo com que o erro se propague por todo o resto da árvore.