Big Data Analytics com R

Árvores de decisão

Agenda

- Construir análise de classificação de dados e árvores de decisão utilizando R;
- Implementar visualizações das análises de classificação de dados e árvores de decisão utilizando R;
- Compreender os parâmetros opcionais para as classificação de dados e árvores de decisão utilizando R.

Árvores de decisão

Motivação

- Uma árvore de decisão chega a uma predição por meio de uma série de perguntas relacionadas a pertencer ou não à determinados grupos.
- Cada questão deve ter apenas duas respostas (sim ou não), o que leva à geração de uma árvore binária.
- Esta característica deve-se ao uso intensivo de probabilidades e o conceito de probabilidade complementar.

Motivação

- Iniciamos pela pergunta conhecida como nó raiz (*root node*) e vamos percorrendo a árvore por seus ramos de acordo com os fatores que levem a decidir por determinado grupo ou não até chegar em uma folha (*leaf node*).
- A proporção alcançada na folha indicará a probabilidade procurada.

Motivação

Motivação

- Imagine que estejamos interessados em saber a probabilidade de sobrevivência a um desastre.
- Certos grupos, como mulheres e crianças, têm prioridade em situações como esta, ganhando então uma chance maior de sobrevivência.

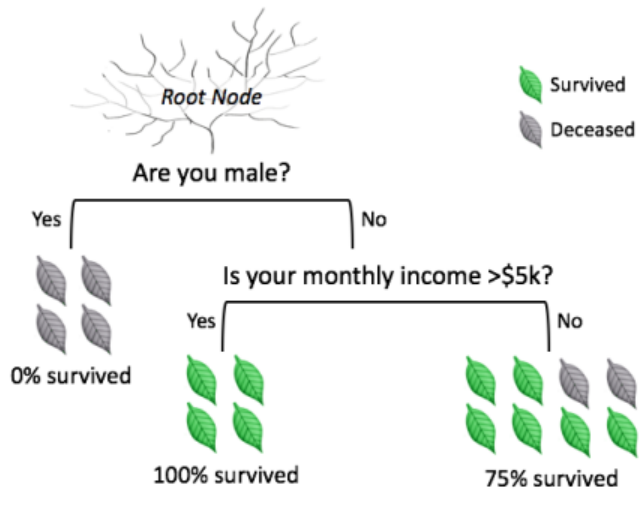
Motivação

- A categorização por estes grupos pode levar à probabilidade de salvamento.
- Para identificar os grupos que possuem maior probabilidade de salvamento, podemos utilizar árvores de decisão.

Motivação

- Apesar de estarmos buscando neste exemplo uma probabilidade de sobrevivência, as árvores de decisão têm uma ampla gama de aplicações.
- Em áreas de negócios, por exemplo, elas podem ser utilizadas para definir perfis de consumidores ou até previsões de quem poderá pedir demissão.
- Na área financeira é utilizada para precificar ativos.
- Em gestão de projetos, possui aplicações em análise de riscos.

Exemplo



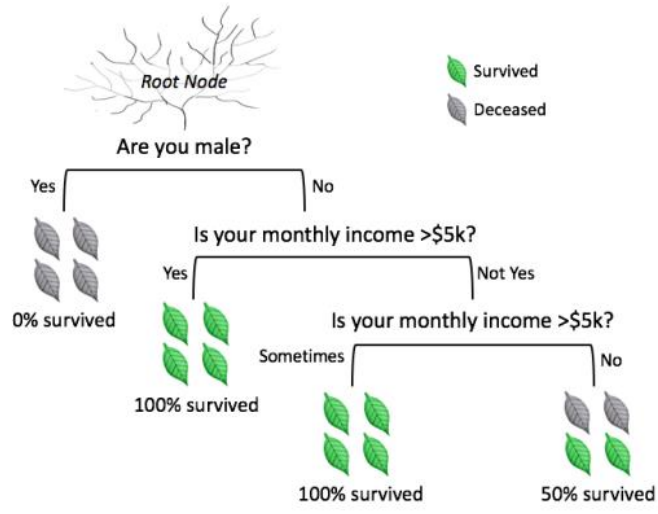
Árvores de decisão

- As árvores de decisão são versáteis pois podem responder questões sobre grupos e/ou categorias (ex.: homens e mulheres), assim como perguntas sobre variáveis contínuas (ex.: receita financeira).
- Caso a pergunta seja sobre uma variável contínua, esta pode ser dividida em grupos como por exemplo, comparar valores que sejam “acima da média” ou “abaixo da média”.

Árvores de decisão

- Como já vimos, árvores de decisão aceitam apenas duas respostas possíveis a cada pergunta (ex.: sim/não), configurando então uma árvore binária.
- Caso seja necessária três ou mais respostas (ex.: sim/não/às vezes), pode-se incluir mais ramos na árvore.

Árvores de decisão



Árvores de regressão vs. classificação

- As árvores chamadas de regressão funcionam de forma idêntica às árvores chamadas de classificação.
- A seguir vamos ver as diferenças entre cada uma delas.

Árvores de regressão vs. classificação

- As árvores de regressão são utilizadas quando a variável dependente é contínua. Já as árvores de classificação são utilizadas para variáveis dependentes categóricas.
- No caso das árvores de regressão, o valor obtido nas folhas é a resposta média das observações que estão naquela região. Logo, se um novo dado a ser avaliado cair na mesma região, a predição será feita por meio desta resposta média.

Árvores de regressão vs. classificação

- Já no caso das árvores de classificação, o valor (classe) obtido nas folhas é a moda das observações que estão naquela região. Logo, se um novo dado a ser avaliado cair na mesma região, a predição será feita seguindo a moda.

Árvores de regressão vs. classificação

- Ambas as árvores dividem o conjunto de variáveis independentes (também conhecido por espaço de predição) em regiões distintas, sem interseção.
- Para simplificar, pode-se considerar estas regiões independentes como caixas.

Árvores de regressão vs. classificação

- Ambas as árvores seguem uma abordagem *top-down* gulosa (*greedy*, em inglês) conhecida por divisão recursiva binária (*recursive binary splitting*).
- É "*top-down*" pois inicia pelo "topo" da árvore, onde todas as observações pertencem à mesma região e vai dividindo o conjunto de forma binária em quebras sucessivas.
- É chamado "gulosa" pois o algoritmo avalia uma quebra de cada vez, sem se preocupar com futuras quebras.

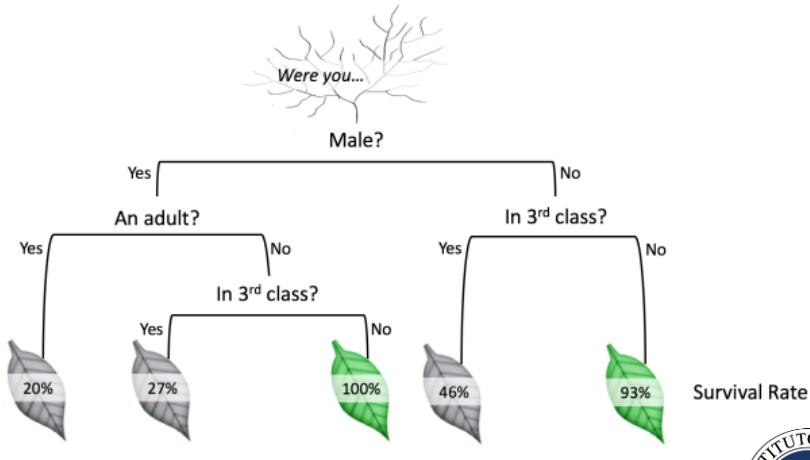
Exemplo

- Vamos ver um exemplo relacionado ao acidente ocorrido com o navio Titanic.
- O objetivo é saber quais grupos de passageiros teriam maior probabilidade de sobrevivência.

Exemplo

- Este conjunto de dados foi compilado originalmente pelo *British Board of Trade*, para investigar o acidente.
- O conjunto de dados utilizados na análise simplificada abaixo é um subconjunto dos dados originais, sendo amplamente disponível para estudos.

Exemplo



Exemplo

- Avaliando a árvore gerada, podemos ver que a probabilidade de sobrevivência seria maior se alguém pertencesse ao grupo de mulheres das cabines de primeira e segunda classes ou ainda ao grupo de crianças do sexo masculino, também das cabines de primeira e segunda classes.

Árvores de decisão

- Já percebemos o quão fácil é interpretar uma árvore de decisão.
- Vamos ver a seguir como elas são geradas.

Algoritmo

- A árvore de decisão começa pela divisão do grupo inicial em dois subgrupos, ambos com dados similares.
- A seguir, repete-se o procedimento de divisão binária em cada subgrupo.
- Desta forma, a cada divisão teremos uma menor quantidade de pontos, porém estes serão mais homogêneos.

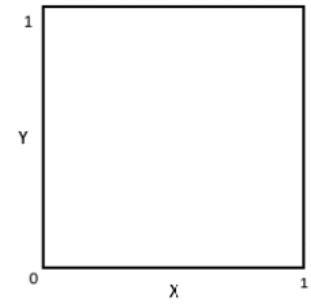
Algoritmo

- O princípio das árvores de decisão é baseado no fato que se isolarmos os diferentes grupos em ramos diferentes da árvore, todos que pertençam a estes grupos terão uma previsão similar.

Algoritmo

- O processo de particionar dados visando a obtenção de grupos homogêneos é chamado **partição recursiva** (*recursive partitioning*) e envolve apenas dois passos:
  1) Descobrir o fator binário que divida o conjunto de dados em dois grupos, da forma mais homogênea possível.
  2) Repetir o passo 1 em cada um dos subgrupos até que determinada condição de parada seja alcançada.

Algoritmo



For more tutorials: annalyzin.wordpress.com

Critérios de parada

- Critérios de parada podem ser definidos de várias formas, exemplos:
  - Parar quando os dados de uma folha pertencerem a uma determinada categoria/valor;
  - Parar quando uma folha tiver menos de cinco dados;
  - Parar quando novas divisões não melhorarem a homogeneidade do subgrupo.

Variáveis não significativas

- A partição recursiva faz uso apenas das melhores perguntas binárias para formar a árvore de decisão.
- Desta forma, a presença de variáveis não significantes não afeta o resultado final.
- Além disso, as questões binárias impõe uma divisão central dos dados, logo, as árvores de decisão são bem robustas quanto aos valores extremos (i.e. *outliers*).

Limitações

- Fazer uso das melhores questões binárias para divisão dos dados pode não levar às predições mais precisas.
- Utilizar divisões assimétricas no início do processo pode levar a predições melhores ao final.

Limitações

- Para resolver este problema, pode-se escolher diferentes combinações de questões binárias para iniciar múltiplas árvores e então agregar os resultados de predição destas árvores.
- Esta técnica é chamada de *random forest*.

Limitações

- Outra opção é selecionar estrategicamente as árvores (no lugar de uma seleção aleatória) de forma que a predição de cada uma das árvores geradas melhore gradativamente.
- Ao final, uma média ponderada das predições geradas por todas as árvores gerará o resultado final.
- Esta técnica é chamada *gradient boosting*.

Limitações

- Apesar de tanto a *random forest* quanto o *gradient boosting* gerarem predições mais precisas, sua complexidade gera dificuldades na visualização dos resultados.
- Assim, estas técnicas são conhecidas como “caixas-pretas”.

Referências

- https://algobeans.com/