PUC Minas

DIRETORIA DE
EDUCAÇÃO CONTINUADA

Pós Graduação *Lato Sensu*

Ciência de dados e Big Data

Técnicas Estatísticas de Predição

Programa

Calendário	Conteúdo	Referencial Teórico
05/02/2019	Regressão Linear (Simples e Multipla)	An Introduction to Statistical Learning with Applications in R, 2013
12/02/2019	Regressão Logística (Simples e Multipla)	
19/02/2019	**Aula cancelada**	
26/02/2019	**Aula cancelada**	
12/03/2019	Modelo Multinomial	Categorical Data Analysis, 2013
19/03/2019	Árvores de decisão e Florestas Aleatórias	An Introduction to Statistical Learning with Applications in R, 2013
26/03/2019	KNN, PCA e K Means Clustering	
08/04/2019	Avaliação	-

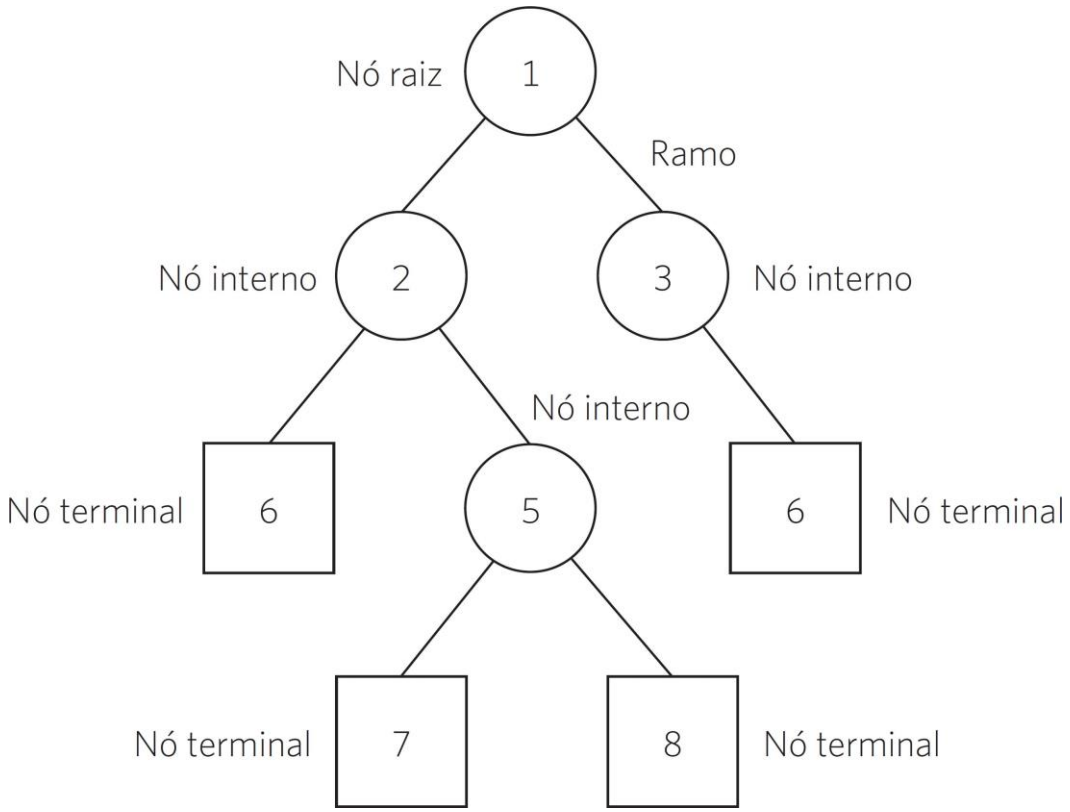
Árvores de decisão e Florestas Aleatórias

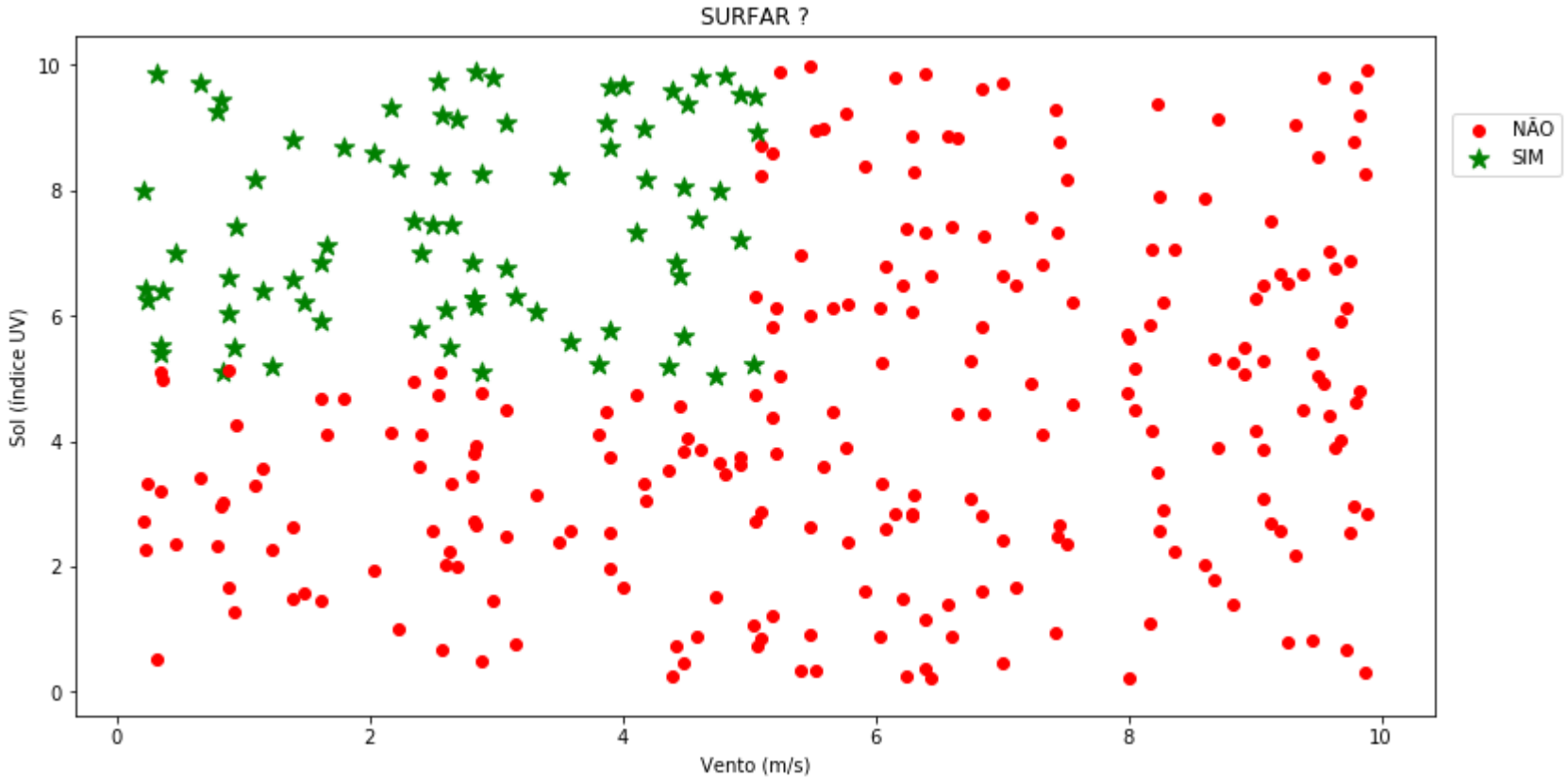
Livro texto: ***An Introduction to Statistical Learning***

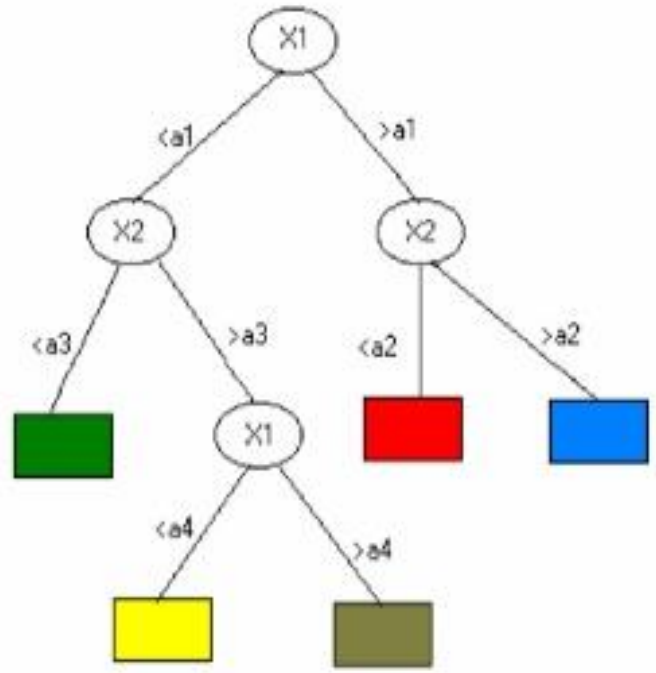
Cap. 8 – Tree-Based MethodsModels

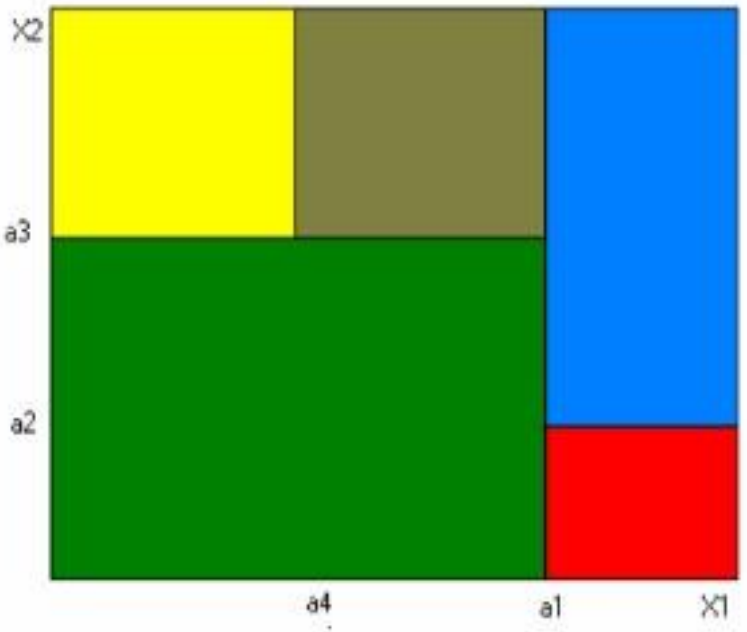
8.1 - The Basics of Decision Trees

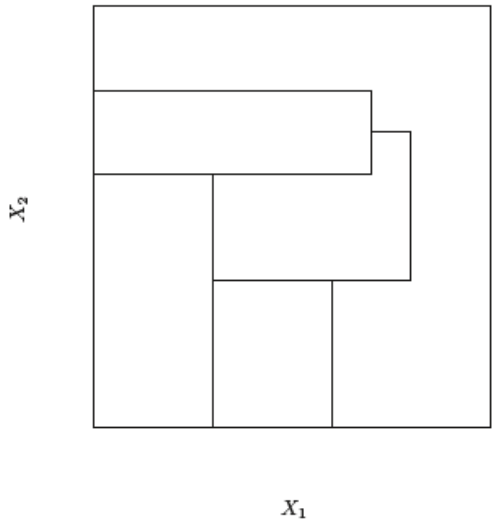
8.2 - Bagging, Random Forests, Boosting

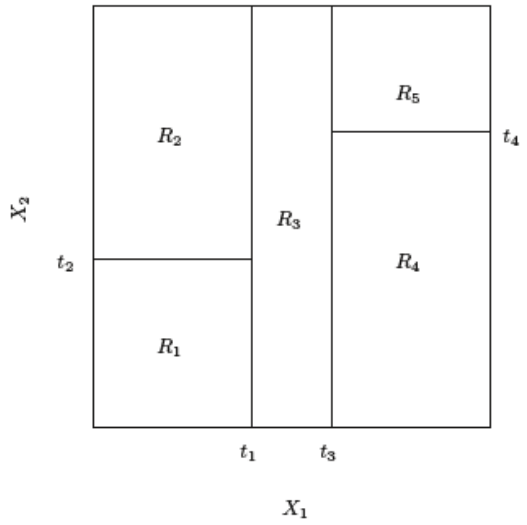

Nó raiz
1
Ramo
Nó interno
2
3
Nó interno
Nó interno
Nó terminal
6
5
6
Nó terminal
Nó terminal
7
8
Nó terminal

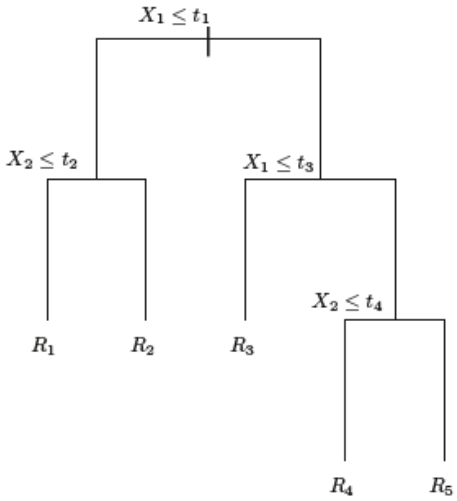

SURFAR ?
Sol (índice UV)
Vento (m/s)
NÃO
SIM
0
2
4
6
8
10

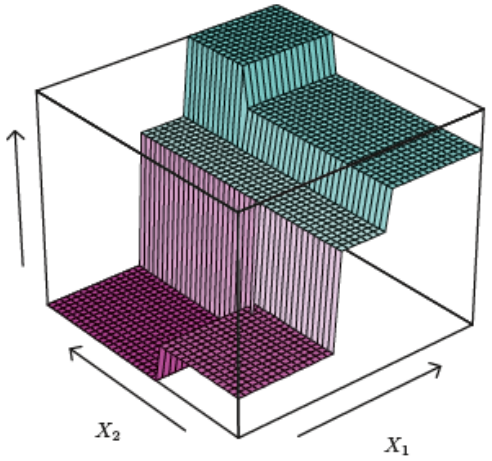

X1
<a1
>a1
X2
X2
<a3
>a3
<a2
>a2
X1
<a4
>a4


X2
a3
a2
a4
a1
X1


X_2
X_1


R_1
R_2
R_3
R_4
R_5
t_1
t_2
t_3
t_4
X_2
X_1


$X_1 \leq t_1$
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3
R_4
R_5


X_2
X_1

Vantagens

1. Fácil de entender: A visualização de uma árvore de decisão torna o problema fácil de compreender, mesmo para pessoas que não tenham perfil analítico. Não requer nenhum conhecimento estatístico para ler e interpretar. Sua representação gráfica é muito intuitiva e permite relacionar as hipóteses também facilmente.

2. Útil em exploração de dados: A árvore de decisão é uma das formas mais rápidas de identificar as variáveis mais significativas e a relação entre duas ou mais variáveis. Com a ajuda de árvores de decisão, podemos criar novas variáveis/características que tenham melhores condições de predizer a variável alvo.

3. Menor necessidade de limpar dados: Requer menos limpeza de dados em comparação com outras técnicas de modelagem. Até um certo nível, não é influenciado por pontos fora da curva "outliers" nem por valores faltantes ("missing values").

4. Não é restrito por tipos de dados: Pode manipular variáveis numéricas e categóricas.

5. Método não paramétrico: A árvore de decisão é considerada um método não-paramétrico. Isto significa que as árvores de decisão não pressupõe a distribuição do espaço nem a estrutura do classificador.

Desvantagens

1. Sobreajuste ("Over fitting"): Sobreajuste é uma das maiores dificuldades para os modelos de árvores de decisão. Este problema é resolvido através da definição de restrições sobre os parâmetros do modelo e da poda (esse fato vai variar muito de um problema para outro).

2. Não adequado para variáveis contínuas: ao trabalhar com variáveis numéricas contínuas, a árvore de decisão perde informações quando categoriza variáveis em diferentes categorias.

Exercício

Queremos saber!

Como é realizada a divisão dos nós de uma árvore?

Dica:

- Índice de Gini
- Qui-Quadrado
- Entropia
- Redução na Variância

Qual a principal diferença entre arvore de decisão e arvore de regressão?

META

- Participação em uma competição de ***Machine Learning*** no *Kaggle*
- Aplicar uma das técnicas de predição utilizadas no curso
- Escrever um artigo sobre modelagem preditiva no *Linkedin*
  - *Causalidade*
  - *Risco*
  - *Identificação*
  - *etc*

Exercício

Jupyter Notebook