Web Data Mining com R

Fabrício J. Barth
fabricio.barth@gmail.com

VAGAS Tecnologia e Faculdade BandTec

Maio de 2014

Objetivo

O objetivo desta palestra é apresentar conceitos sobre Web Data Mining, fluxo de trabalho e exemplos de tarefas de Web Data Mining utilizando o R.

Sumário

- Conceitos: web data mining, aprendizagem de máquina e a linguagem de programação R.
- Análise de mensagens do twitter usando algoritmos de agrupamento.
- Desenvolvimento de algoritmos anti-spam.
- Considerações finais.
- Referências.

Conceitos

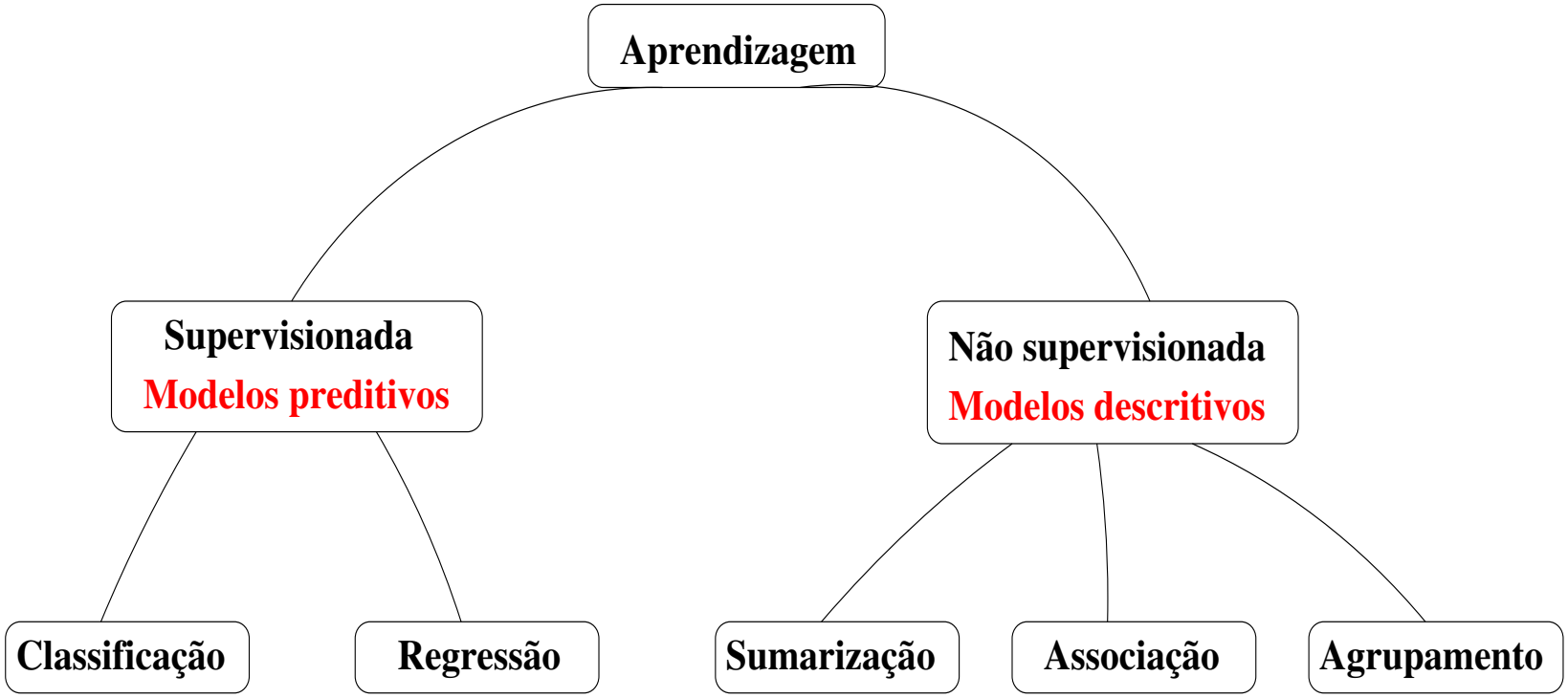
Web Data Mining

A área de Web Data Mining tem como objetivo **descobrir conhecimento útil** a partir da estrutura dos hyperlinks da Web, conteúdo das páginas e log de utilização dos sites.

Descobrir conhecimento útil:

- Sintetizar informação:
  - ⋆ a partir de logs de servidores web, identificar qual é o caminho mais frequente de navegação dos usuários no site.
  - ⋆ a partir de notícias publicadas em veículos web, sumarizar os principais eventos do dia.
- Prescrever ações:
  - ⋆ a partir do histórico de candidaturas em vagas de um candidato, recomendar novas vagas para o mesmo.
  - ⋆ a partir de conteúdo previamente moderado, construir uma aplicação capaz de moderar conteúdo automaticamente.

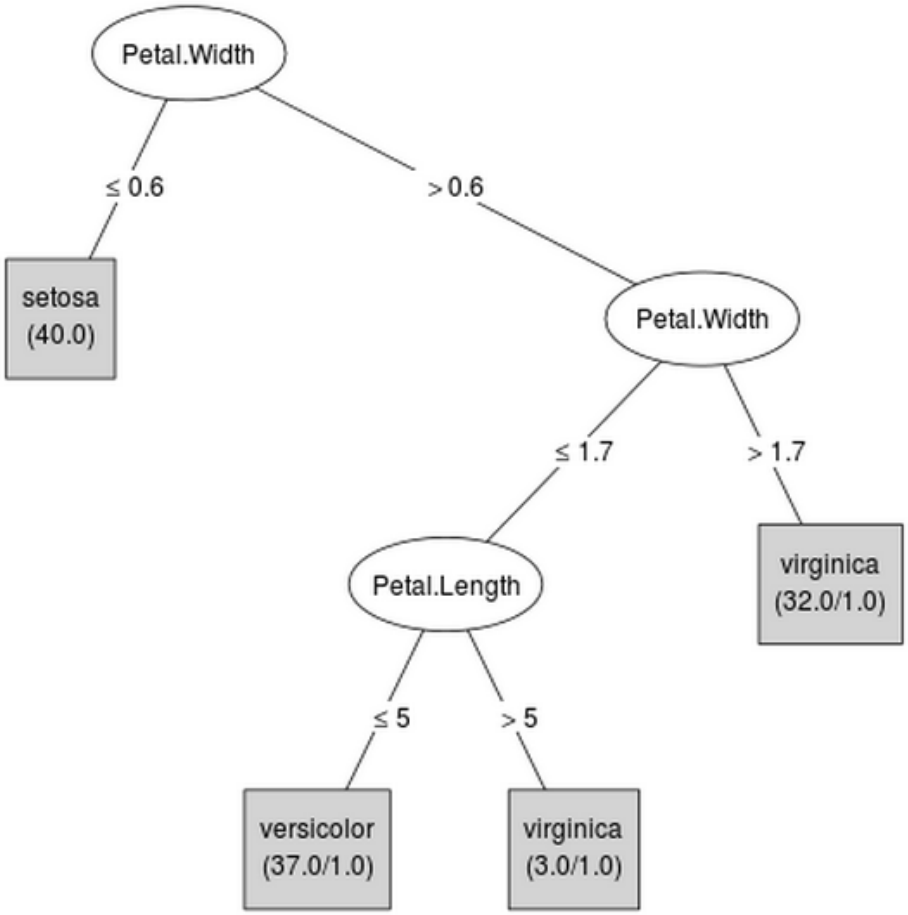
Aprendizagem de máquina



Exemplo de dataset com **classe**

Idade	Miopia	Astigmat.	Lacrimej.	**Lentes**
jovem	míope	não	reduzido	**nenhuma**
jovem	míope	não	normal	**fraca**
jovem	míope	sim	reduzido	**nenhuma**
jovem	míope	sim	normal	**forte**
...	...	...	...	...
adulto	míope	não	reduzido	**nenhuma**

Exemplo de modelo preditivo

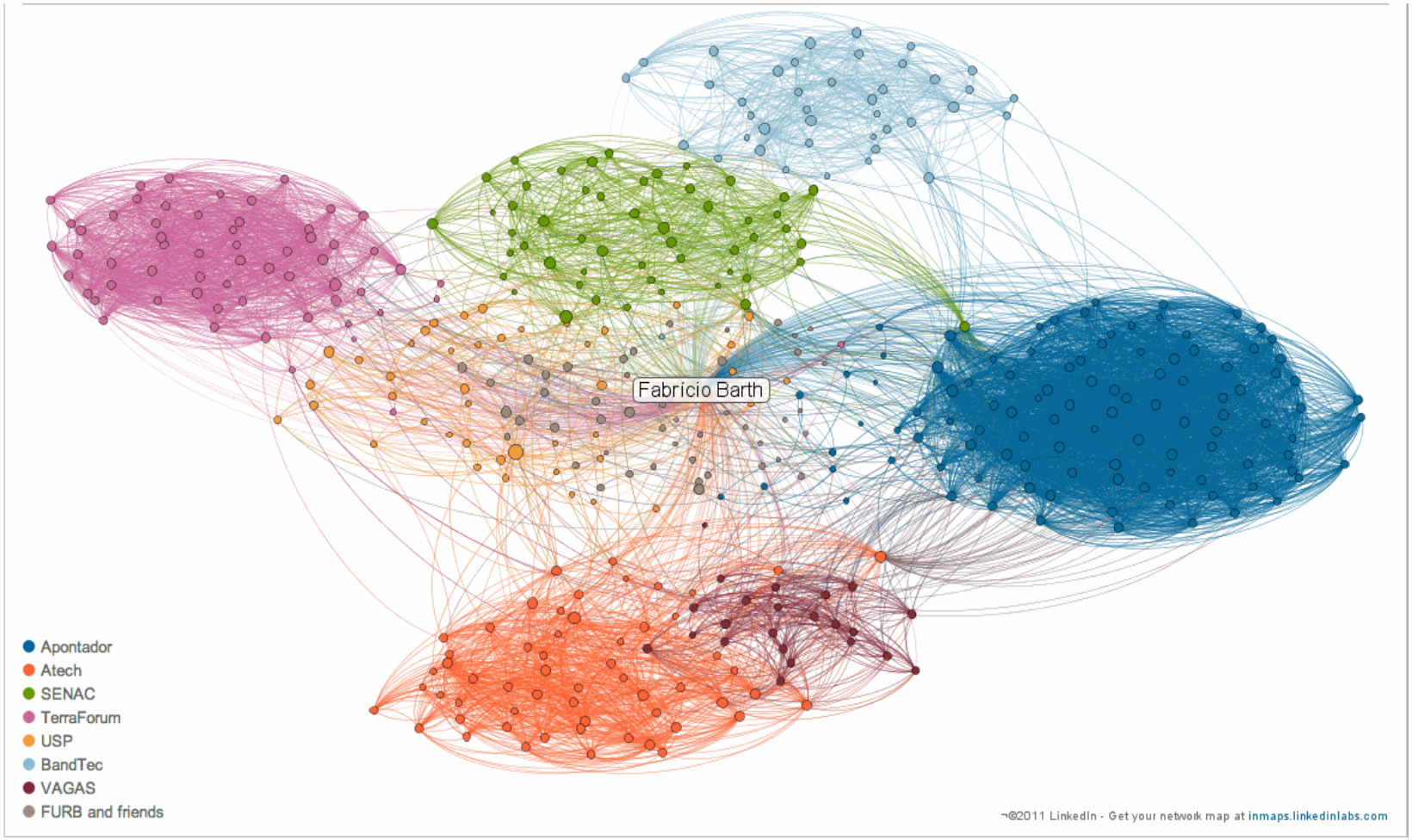


Exemplos de aprendizagem não supervisionada

Table 1: Exemplo de tabela com as transações dos usuários

usuário	$categoria_1$	$categoria_2$	$categoria_3$	$\cdots$	$categoria_m$
$user_1$	0	2	0	$\cdots$	1
$user_2$	1	1	0	$\cdots$	0
$user_3$	2	0	1	$\cdots$	0
$user_4$	0	1	0	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$user_n$	1	1	0	$\cdots$	1

Exemplo de identificação de grupos em redes sociais



Projeto R

- http://www.r-project.org/
- R Studio - http://www.rstudio.com/
- É free
- É a linguagem de programação mais popular para análise de dados
- Script é melhor que clicar e arastar:
  - ⋆ É mais fácil de comunicar.
  - ⋆ Reproduzível.
  - ⋆ É necessário pensar mais sobre o problema.
- Existe uma quantia grande de pacotes disponíveis

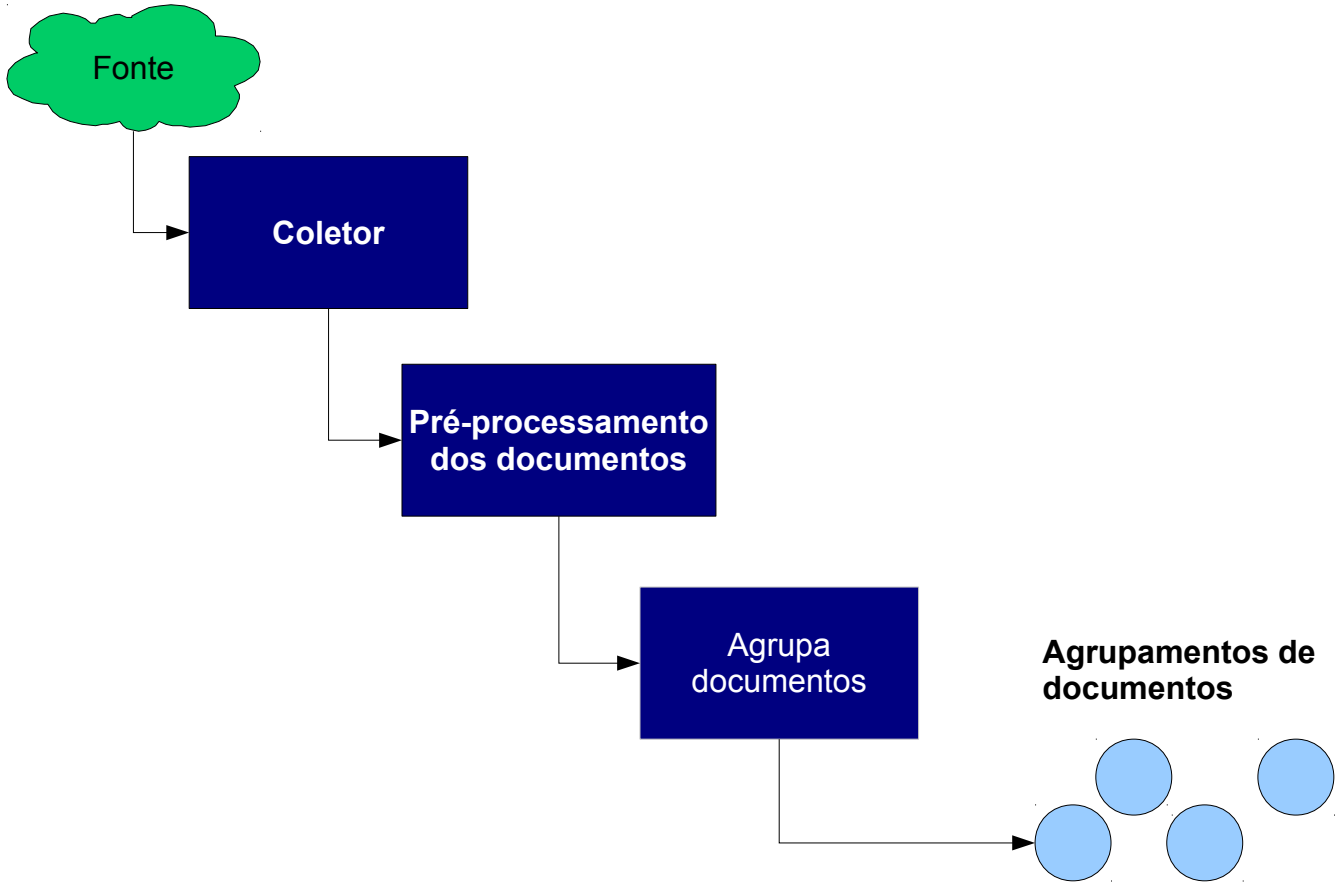
Web Data Mining e *dados não estruturados*

A área de Web Data Mining tem como objetivo descobrir conhecimento útil a partir da estrutura dos **hyperlinks da Web**, **conteúdo das páginas** e **log de utilização dos sites**.

- São todos dados não estruturados.
- Estes dados precisam ser pré-processados e convertidos em dados estruturados.

Análise de mensagens do twitter usando algoritmos de agrupamento

Componentes para uma solução...



Coletando dados do twitter com o R

```
library(twitteR)
cred <- OAuthFactory$new(
                consumerKey="XXXX",
                consumerSecret="XXXX",
                requestURL="https://api.twitter.com/oauth/request_token",
                accessURL="https://api.twitter.com/oauth/access_token",
                authURL="http://api.twitter.com/oauth/authorize")

cred$handshake()
registerTwitterOAuth(cred)

dados <- searchTwitter('economist_brasil', n=250)
df <- twListToDF(dados)
save(df, file="../data/20140424_economist_brasil.rda")
```

Formato de um documento

... Esta disciplina tem como objetivo apresentar os principais conceitos da área de Inteligência Artificial, caracterizar as principais técnicas e métodos, e implementar alguns problemas clássicos desta área sob um ponto de vista introdutório.
A estratégia de trabalho, o conteúdo ministrado e a forma dependerão dos projetos selecionados pelos alunos.
Inicialmente, os alunos deverão trazer os seus Projetos de Conclusão de Curso, identificar intersecções entre o projeto e a disciplina, e propor atividades para a disciplina. ...

Conjunto de Exemplos - Atributo/Valor

Doc.	**apresent**	**form**	**tecnic**	**caracteriz**	$\cdots$
d_1	0.33	0.33	0.33	0.33	$\cdots$
d_2	0	0.5	0.2	0.33	$\cdots$
d_3	1	0.6	0	0	$\cdots$
d_4	0.4	0.3	0.33	0.4	$\cdots$
d_5	1	0.4	0.1	0.1	$\cdots$
d_n	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Atributo/Valor usando vetores

Como representar os documentos?

$$\overrightarrow{d_i} = (p_{i1}, p_{i2}, \cdots, pin) \tag{1}$$

- Os atributos são as palavras que aparecem nos documentos.
- As palavras do texto precisam ser normalizadas: caixa baixa, remover acentuação, remover stop-words, aplicar algoritmos de steamming.

Remover stop-words

- Em todos os idiomas existem átomos (palavras) que não significam muito. **Stop-words**

Esta disciplina **tem como** objetivo apresentar **os** principais conceitos **da** área **de** Inteligência Artificial**,** caracterizar **as** principais técnicas **e** métodos**, e** implementar alguns problemas clássicos **desta** área **sob um** ponto **de** vista introdutório**.**

. . .

Algoritmos de steamming

- Algumas palavras podem aparecer no texto de diversas maneiras: **técnica**, **técnicas**, **implementar**, **implementação**...
- **Stemming** - encontrar o radical da palavra e usar apenas o radical.

Atributo/Valor usando vetores

- Já conhecemos os atributos.
- E os valores?
  - ⋆ **Booleana** - se a palavra aparece ou não no documento (1 ou 0)
  - ⋆ **Por freqüência do termo** - a freqüência com que a palavra aparece no documento (normalizada ou não)
  - ⋆ **Ponderação tf-idf** - o peso é proporcional ao número de ocorrências do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece.

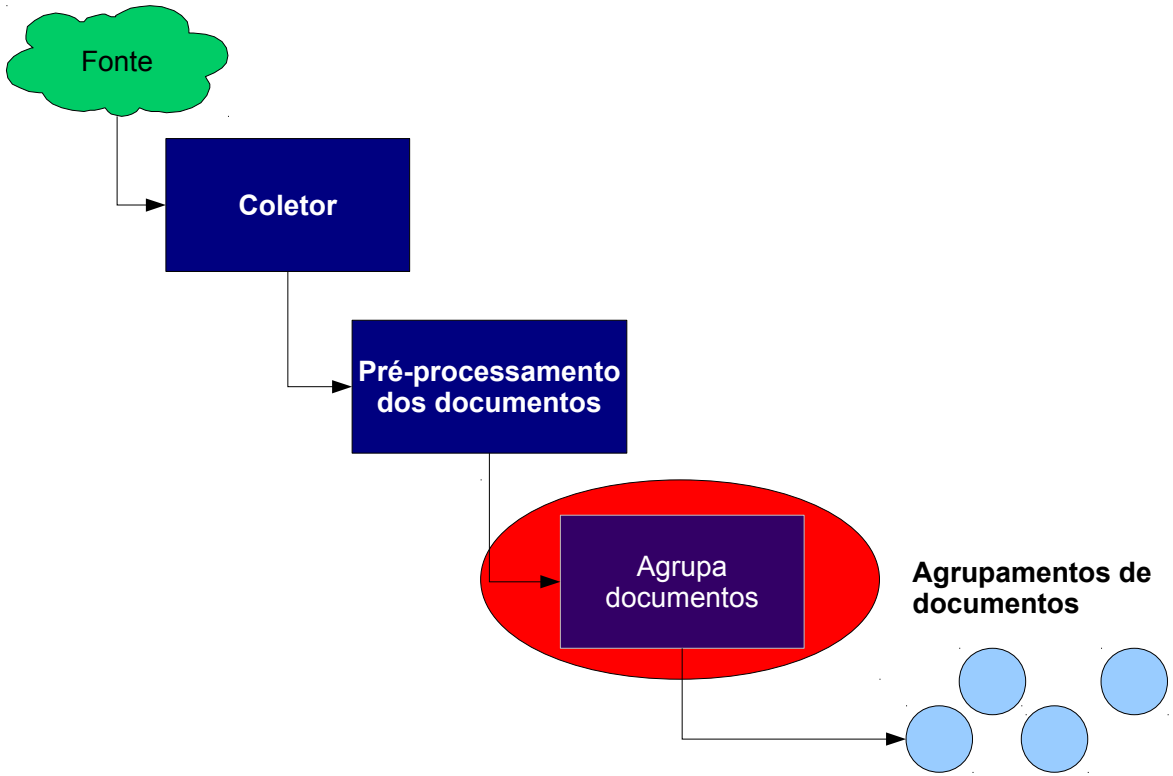
Por freqüência do termo

(apresent,0.33) (form,0.33) (tecnic,0.33) (caracteriz,0.33)
(projet,1.0) (introdutori,0.33) (objet,0.33) (inteligente,0.33)
(conclusa,0.33) (selecion,0.33) (intersecco,0.33) (classic,0.33)
(identific,0.33) (conceit,0.33) (trabalh,0.33) (disciplin,1.0)
(traz,0.33)

Executando esta etapa no R

http://rpubs.com/fbarth/agrupamentoTwitterConalytics

Componentes para uma solução...



Algoritmos para Agrupamento

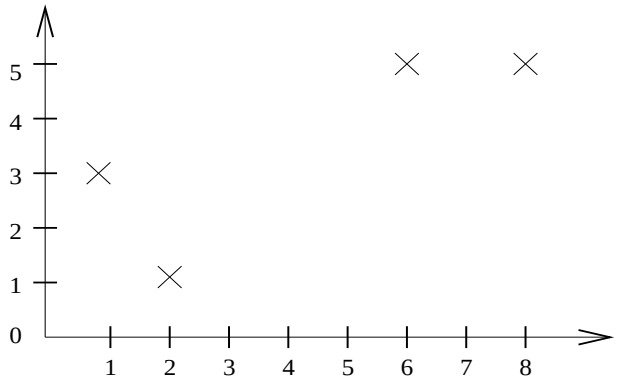
Definições de Algoritmos de Agrupamento

- O objetivo dos algoritmos de agrupamento é colocar os objetos **similares** em um **mesmo grupo** e objetos **não similares** em **grupos diferentes**.
- Normalmente, objetos são descritos e agrupados usando um conjunto de **atributos e valores**.
- Não existe nenhuma informação sobre a classe ou categoria dos objetos.

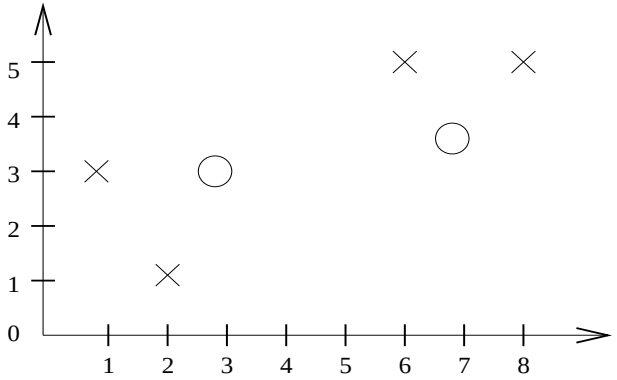
Algoritmos para Agrupamento - *K-means*

- **K** significa o número de agrupamentos (que deve ser informado à priori).
- Sequência de ações **iterativas**.
- A parada é baseada em algum critério de qualidade dos agrupamentos (por exemplo, similaridade média).

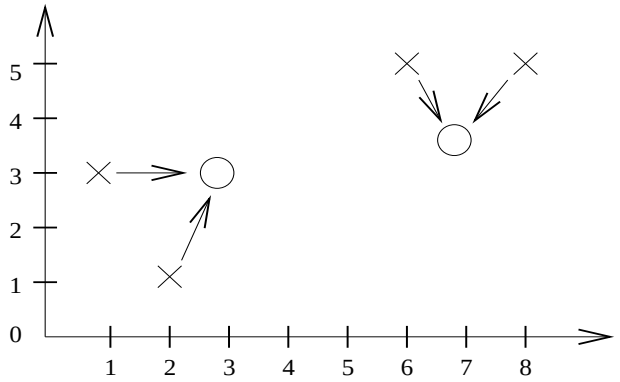
Algoritmo para Agrupamento - *K-means*



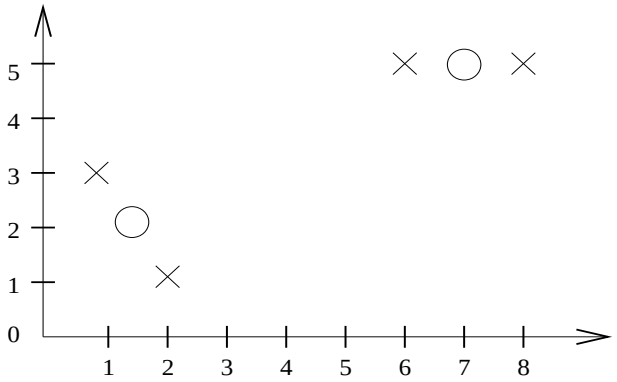
(1) Objetos que devem ser agrupados



(2) Sorteio dos pontos centrais dos agrupamentos



(3) Atribuição dos objetos aos agrupamentos



(4) Definição do centro do agrupamento

Algoritmo **K-means**

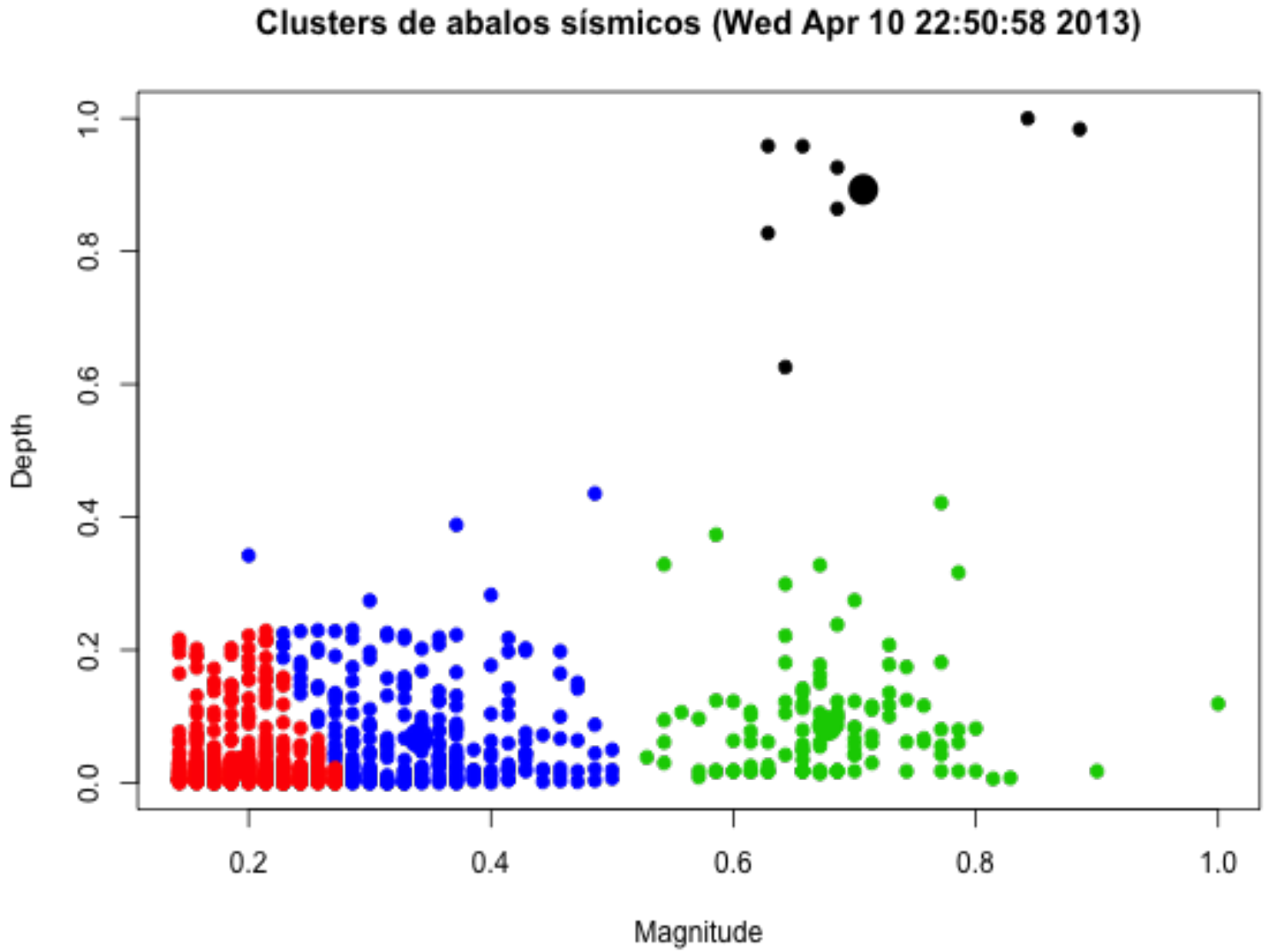
- A medida de distância pode ser a distância Euclidiana:

$$| \vec{x} - \vec{y} | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

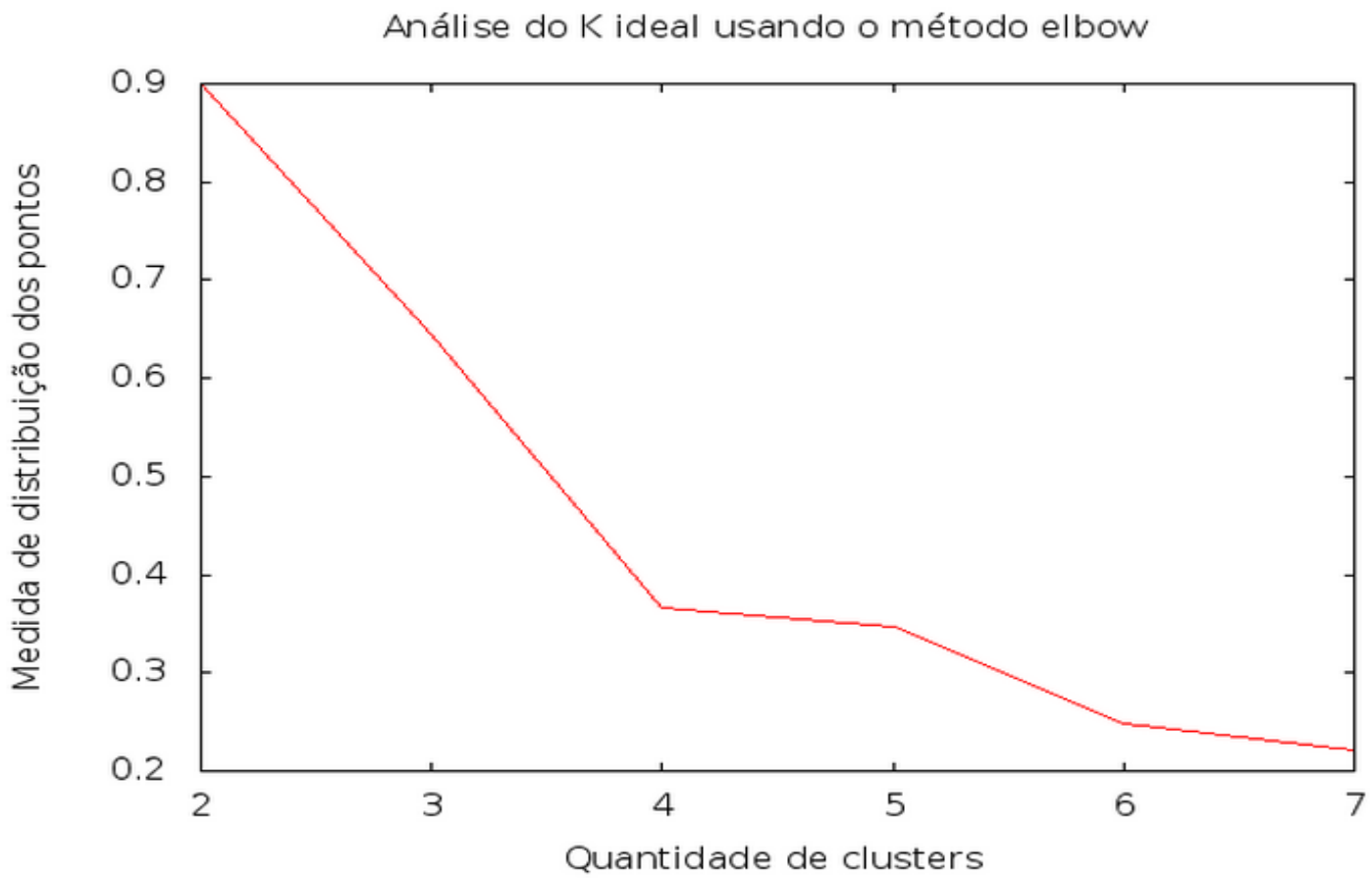
- a função para computar o ponto central pode ser:

$$\vec{\mu} = \frac{1}{M} \sum_{\vec{x} \in C} \vec{x} \tag{3}$$

onde M é igual ao número de pontos no agrupamento C.


Clusters de abalos sísmicos (Wed Apr 10 22:50:58 2013)
Depth
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
Magnitude

Como determinar o melhor k?

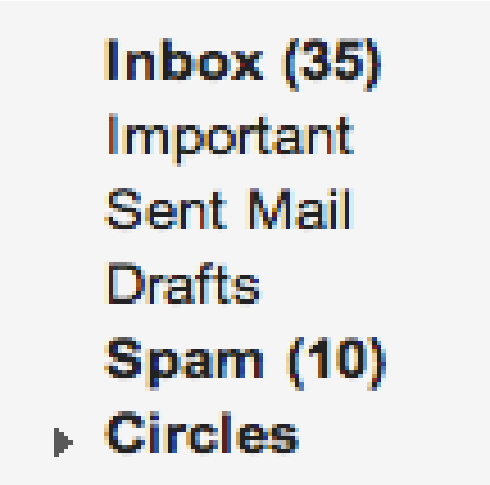


A medida de distribuição dos pontos normalmente empregada é *sum of squared errors*.

Agrupamento de mensagens do twitter com o R

http://rpubs.com/fbarth/agrupamentoTwitterConalytics

Desenvolvimento de algoritmos anti-spam



[redacted]
Membro desde 26/05/2011

6449 avaliações | 2193 seguidores | 3 locais | 53 fotos

15/01/2014 via Apontador Android

excelente churrascaria,tudo muito gostoso

Essa avaliação me ajudou *(0)*

Reportar abuso

[redacted]
Membro desde 11/07/2013

12 avaliações | 1 seguidor | 0 local | 0 foto

11/07/2013 via Apontador

A comida é boa, porém o ambiente é um pouco cheio e causa demora no atendimento.

Essa avaliação me ajudou *(2)*

Reportar abuso

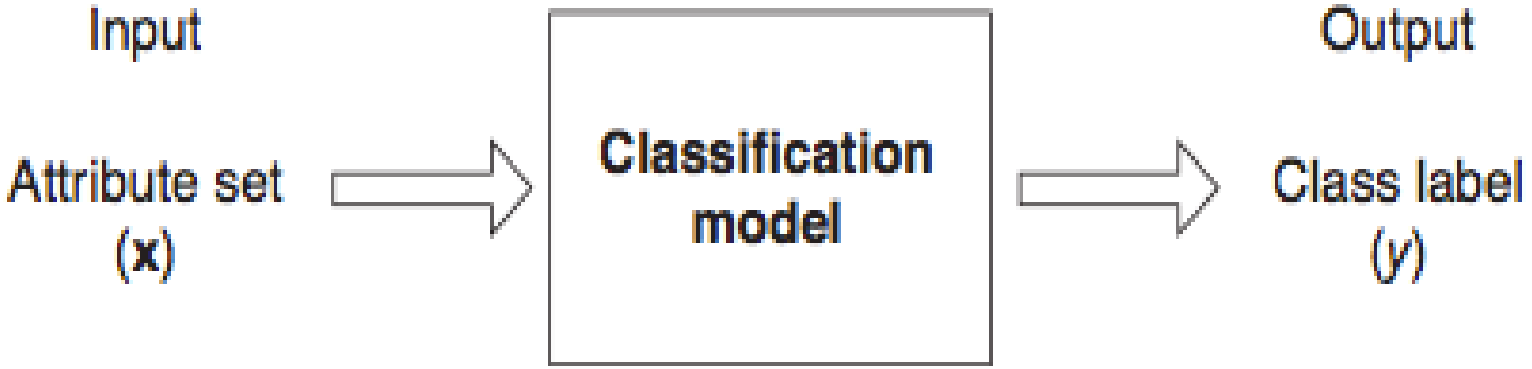
[redacted]
Membro desde 26/02/2011

646 avaliações | 493 seguidores | 126 locais | 300 fotos

24/06/2013 via Apontador

uma boa comida vc entra aqui, recomendo a todos!!

Modelos preditivos para classificação



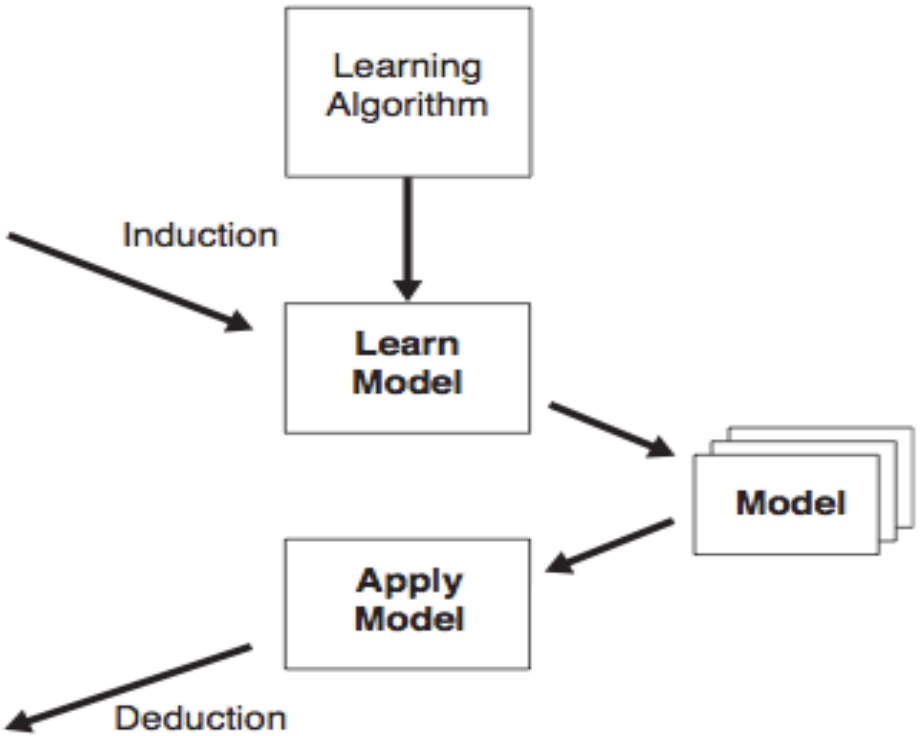
Desenvolvimento de modelos preditivos para classificação

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

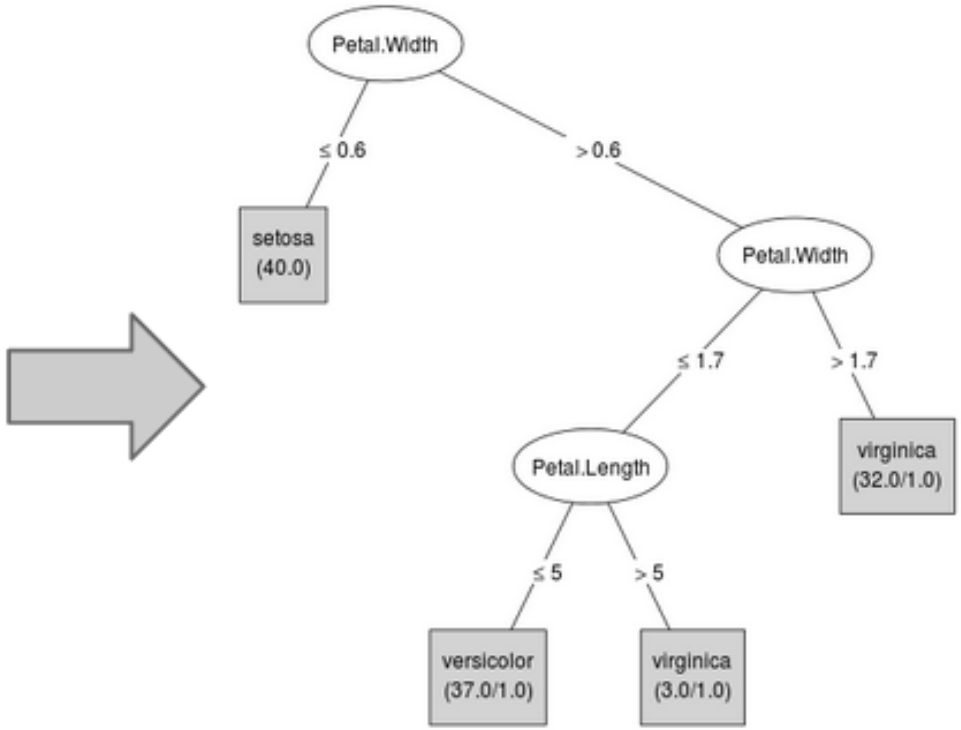
Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



Aprendizado de árvores de decisão

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
19	5.7	3.8	1.7	0.3	setosa
30	4.7	3.2	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
50	5.0	3.3	1.4	0.2	setosa
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
81	5.5	2.4	3.8	1.1	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
104	6.3	2.9	5.6	1.8	virginica
112	6.4	2.7	5.3	1.9	virginica
122	5.6	2.8	4.9	2.0	virginica
126	7.2	3.2	6.0	1.8	virginica
146	6.7	3.0	5.2	2.3	virginica
148	6.5	3.0	5.2	2.0	virginica



Florestas de árvores de decisão

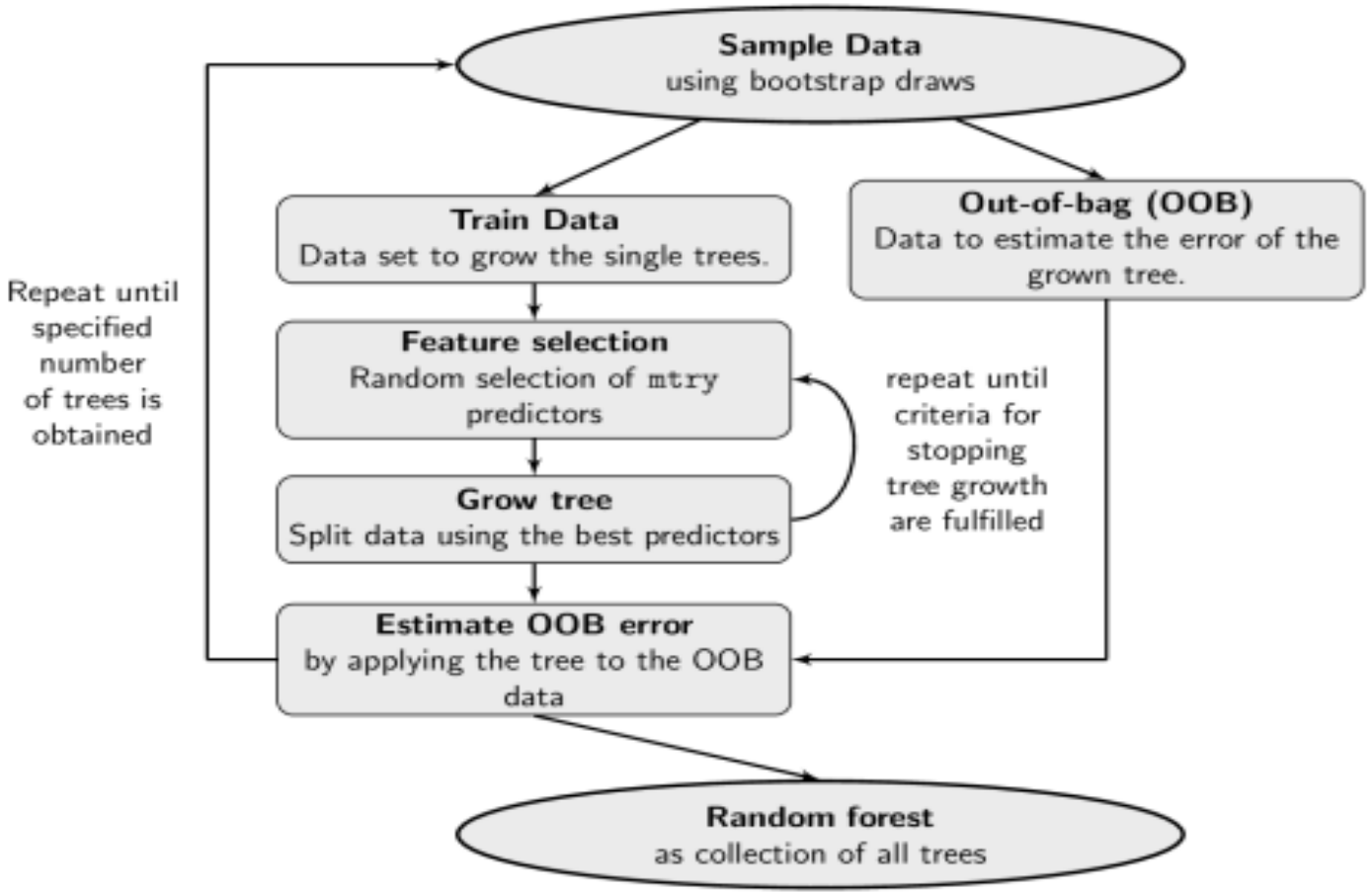


Figure 1: Random Forest Algorithm

Exemplo de classificação de Spam usando RandomForest

http://rpubs.com/fbarth/classificacaoSpamRandomForest

Considerações finais

- Análise de mensagens do twitter
  - ⋆ Transformação de informação não-estruturada em estruturada.
  - ⋆ Uso do algoritmo k-means
  - ⋆ Este mesmo processo pode ser aplicado para outros problemas similares: análise de notícias, análise de patentes e artigos científicos.

- Desenvolvimento de algoritmos anti-spam
  - ⋆ Uso do algoritmo random forest.
  - ⋆ Como desenvolver e avaliar um modelo preditivo.
  - ⋆ Este mesmo processo pode ser aplicado para outros problemas similares, inclusive problemas de recomendação de itens.

Material de **consulta**

- fbarth.net.br/materiais/docs/webMiningRconalytics.pdf: link para os slides.
- http://fbarth.net.br/materiais/webMiningR.html: tutorial apresentado no Mozilla Tech Day 2013.
- http://rpubs.com/fbarth/: scripts em R para problemas de Aprendizagem de Máquina.
- fabricio.barth@gmail.com

Referências

- Bing Liu. Web Data Mining: exploring hyperlinks, contents, and usage data, 2008.
- Tom Mitchell. Machine Learning, 1997.
- Iah H. Witteh and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), 2011.
- Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining, 2006.
- Andrew Ng. http://www.ml-class.org

- Andy and Matthew. Classification and regression by randomForest. R News, vol. 3, number 3, pages 18-22, 2002.
- Costa, H.; Merschmann, L. H. C.; Barth, F.; Benevenuto, F. Pollution, Bad-mouthing, and Local Marketing: The Underground of Location-based Social Networks. Information Sciences, 2014.
- RDataMining.com: Text Mining. http://www.rdatamining.com/examples/text-mining. Acessado em 14 de junho de 2013.
- Ingo Feinerer. Introduction to the tm Package: Text Mining in R. http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Acessado em 14 de junho de 2013.

- Barth, F. J. Ferramentas para a detecção de grupos em Wikis. In: VII Simpósio Brasileiro de Sistemas Colaborativos, 2010, Belo Horizonte. Anais do VII Simpósio Brasileiro de Sistemas Colaborativos. IEEE Computer Society, 2010. v.II. p.8 - 11.
- Barth, F. J. ; Belderrain, M. C. R. ; Quadros, N. L. P. ; Ferreira, L. L. ; Timoszczuk, A. P. . Recuperação e mineração de informações para a área criminal. In: VI Encontro Nacional de Inteligência Artificial, 2007, Rio de Janeiro. Anais do XXVII Congresso da SBC, 2007.