Web Data mining com R: aprendizagem de máquina

Fabrício Jailson Barth

Faculdade BandTec e VAGAS Tecnologia

Junho de 2013

Sumário

- O que é Aprendizagem de Máquina?
- Hierarquia de aprendizado.
- Exemplos de aprendizagem supervisionada (modelos preditivos).
- Exemplos de aprendizagem não supervisionada (modelos descritivos).
- Referências e exercícios.

Contexto e exemplos

- **Data Mining**: grandes bases de dados têm crescido com a automatização de alguns processos e com o advento da Web, por exemplo: web click data, registros médicos, dados biológicos, dados capturados a partir de sensores.
- **Aplicações que não podem ser programadas "na mão"**: por exemplo, helicopteros autônomos, reconhecedor de escrita, processadores de linguagem natural e sistemas de visão computacional.
- **Aplicações personalizáveis**: Amazon, Netflix.
- Compreensão do aprendizado humano.

O que é Aprendizagem de Máquina?

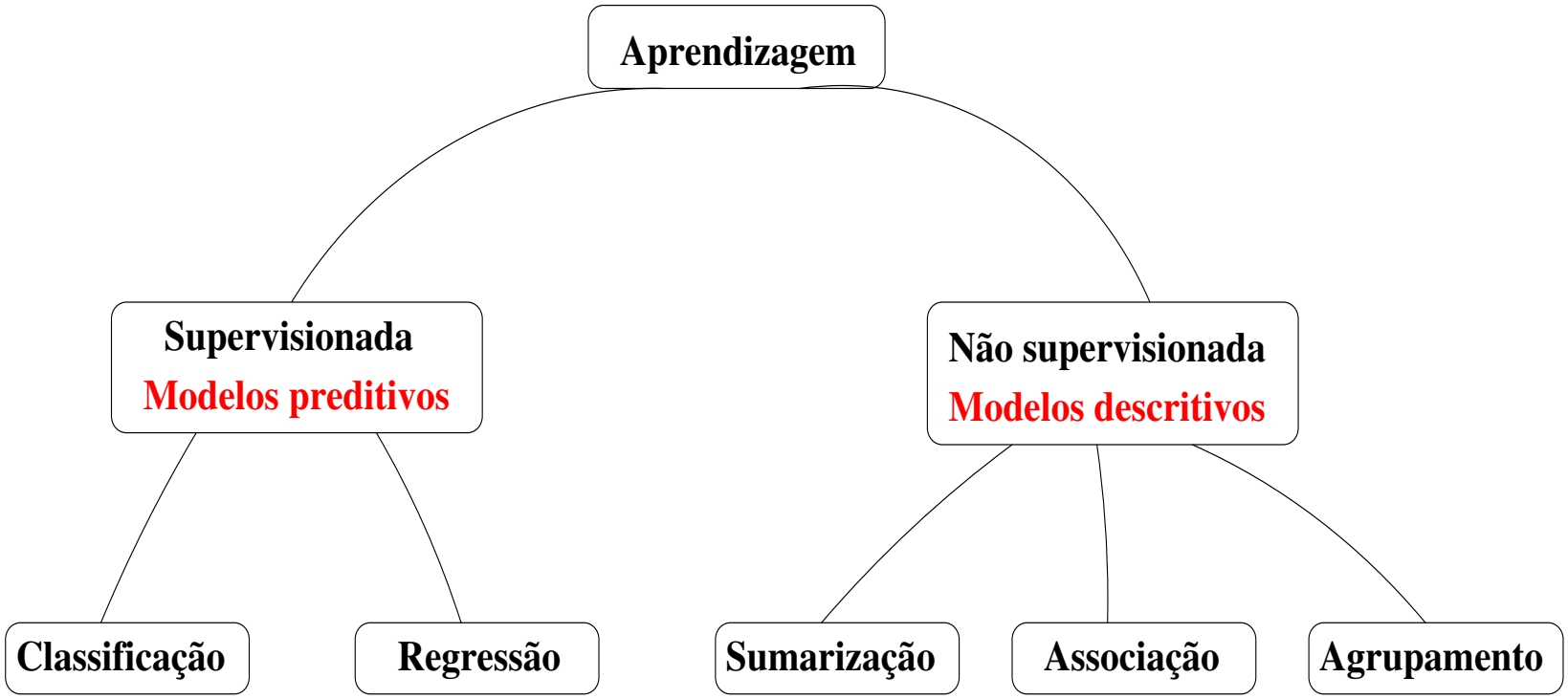
- Área de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados [Arthur Samuel (1959)].
- **Definição bem formada**: *A computer program is said to learn from experience A with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E* [Tom Mitchell (1998)].

Exercício

O que cada uma das sentenças abaixo descreve segundo a definição do Tom Mitchell?

- Classificar e-mails como spam ou não spam.
- Verificar quais e-mails o usuário classifica como spam.
- O número (ou fração) de e-mails corretamente classificados como spam ou não spam.

Hierarquia de aprendizado



Exemplos de aprendizagem supervisionada

- Estimar o preço de uma casa.
  - ⋆ atributos: tamanho, posição geográfica, material.
  - ⋆ classe: preço (regressão).
- Determinar se uma pessoa tem câncer benigno ou maligno.
  - ⋆ atributos: tamanho do tumor, formato do tumor, idade do paciente.
  - ⋆ classe: tumor benigno ou tumor maligno (classificação).

- Determinar se é um texto publicado em uma rede social é inadequado ou não.
  - ⋆ atributos: quantidade de palavras encontradas no texto, quantidade de palavras proibidas encontradas no texto, quantidade de textos já criados pelo usuário, idade do usuário no sistema, quantidade de textos criados pelo usuário e moderados, ...
  - ⋆ classe: texto adequado ou não (classificação).
  - ⋆ **classes**: texto adequado, texto inadequado, texto com propaganda (classificação com múltiplas classes).

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**

Exercícios

Que problema deve ser tratado como problema de regressão e que problema deve ser tratado como problema de classificação?

- A sua empresa possui 1.000 itens idênticos em estoque. Você quer predizer quantos destes itens serão vendidos nos próximos três meses.
- Você quer examinar clientes seus e para cada um decidir se ele irá pagar todo o financiamento ou não.

Exemplos de aprendizagem não supervisionada

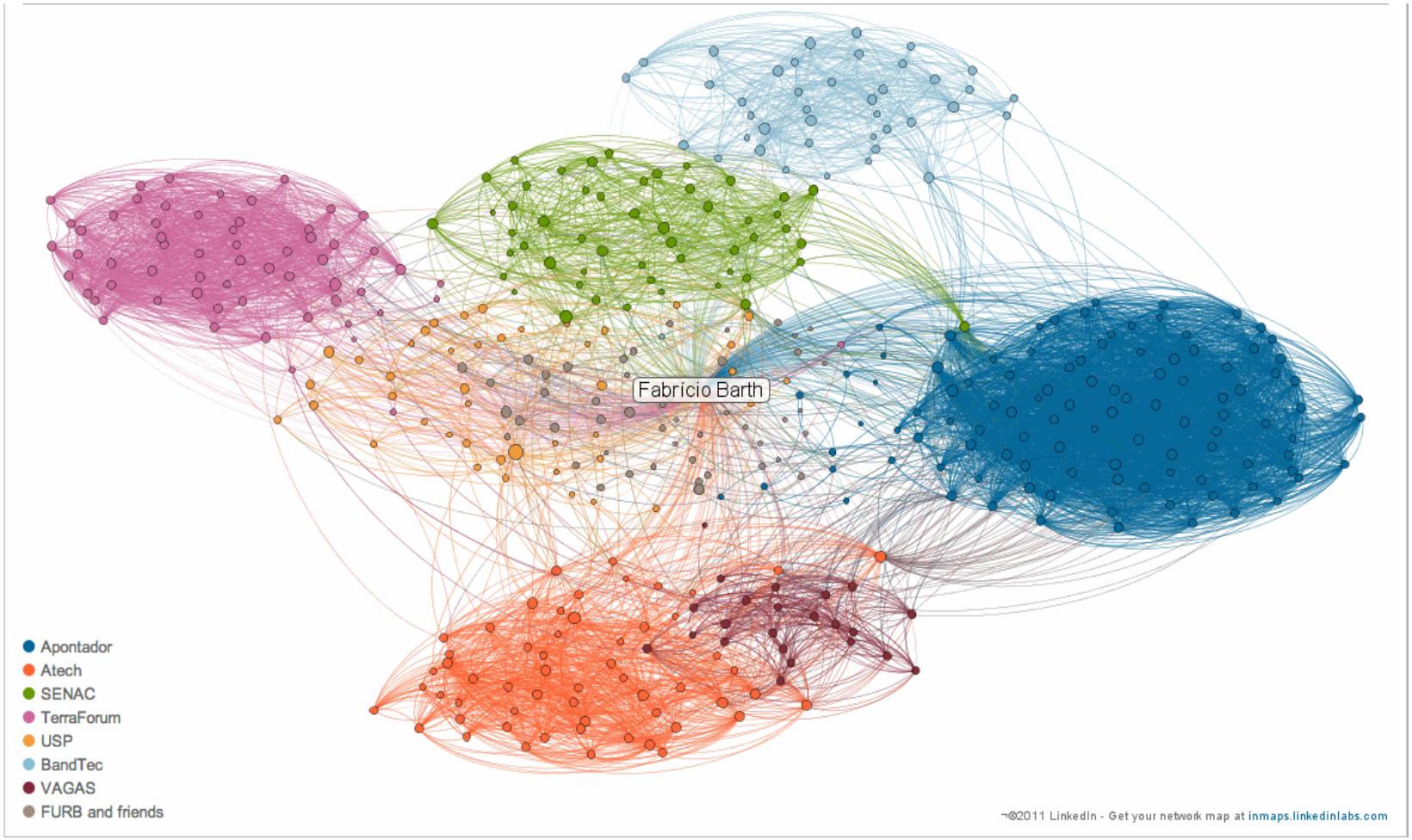
- Dado conjuntos de itens adquiridos na mesma compra, identificar padrões de compra.
- Identificar padrões de navegação em sites.
- Agrupar notícias semelhantes publicadas por várias fontes de informação.
- Numa rede social, identificar sub-grupo de pessoas.

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



Exercícios

Quais dos problemas abaixo você iria resolver com uma abordagem não supervisionada de aprendizagem?

- Dado e-mails rotulados como spam e não spam, desenvolver um filtro de spam.
- Dado um conjunto de notícias encontradas na Internet, agrupá-las em conjunto de notícias que tratam do mesmo assunto.

- Dado uma base de clientes, descobrir segmentos de clientes.
- Dado uma base de pacientes diagnosticados com diabetes ou não, aprender a classificar novos pacientes com diabetes ou não.

Material de **consulta**

- Tom Mitchell. Machine Learning, 1997.
- Iah H. Witteh and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), 2011.
- Andrew Ng. http://www.ml-class.org
- *Faceli, Lorena, Gama, Carvalho. Inteligência Artificial: uma abordagem de aprendizado de máquina, 2011.*