Introdução a machine learning com aplicações em R

Aula 1 - Conceitos básicos de machine learning

Felipe Barletta

Departamento de Estatística

29 outubro, 2020

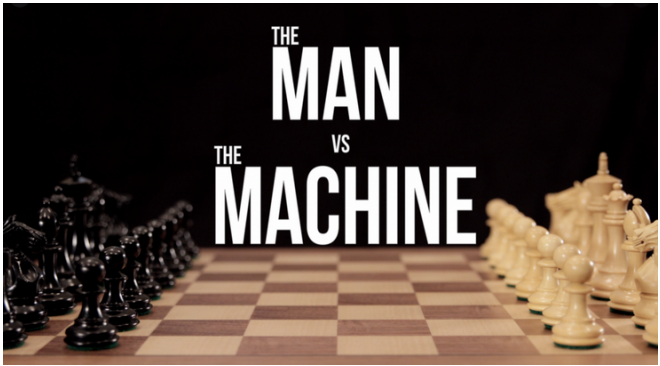
Sumário

A origem do machine learning

- **Machine learning** - Aprendizado de máquina (em português).
- Automatização e registros sistemáticos de informações. (Constante crescimento).
- Conjunto de ferramentas que os computadores usam para transformar dados em conhecimento.
- O aprendizado de máquina é mais parecido com treinar um funcionário do que criar um filho. (Lantz, B., **Machine Learning with R**).
- Representação de Inteligência Artificial.

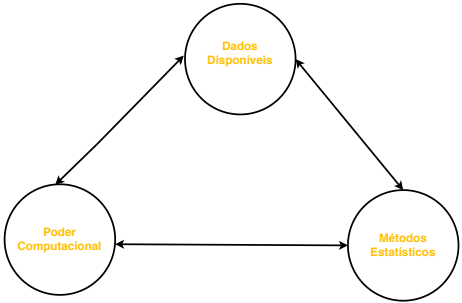
Exemplo - *Match Garry Kasparov vs Deep Blue*

- O primeiro match foi realizado em 1996 na Filadélfia, EUA - Kasparov venceu por 4-2.
- Segundo macth em 1997 - Kasparov perdeu por 3-2.



Introdução

- Algoritmos de computador para transformar dados em ações inteligentes
- Relação entre machine learning e data mining (mineração de dados em português)



Introdução

- Modelos sob uma ótica preditivista (assim como uma revisão de modelos tradicionais sob esta mesma ótica.)

$$Y \in \mathbb{R} \dashrightarrow x = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$$

- Estatística + Ciência da computação = *Machine Learning*
- Métodos estatísticos aplicados a problemas da ciência da computação

Predição versus Inferência

- **Cultura da modelagem dos dados**
  - Domina a comunidade estatística.
  - Principal objetivo é a interpretação dos parâmetros.
  - Testar suposições é fundamental.
- **Cultura da modelagem por algoritmos**
  - Domina a comunidade de Machine Learning.
  - O modelo é utilizado para criar bons algoritmos preditivos.
  - Interpretamos os resultados, mas esse - em geral - não é o foco.

Exemplos de aplicações de machine learning

- Identificar e filtrar mensagens de spam de e-mail
- Prever atividades criminosas
- Automatizar os sinais de trânsito de acordo com as condições da estrada
- Produzir estimativas financeiras
- Prever tempestades e desastres naturais
- Examinar a rotatividade de clientes
- Risco na concessão de crédito
- Criar aviões e carros com direção automática
- Identificar indivíduos com capacidade de compra
- Direcionar publicidade para tipos específicos de consumidores
- Risco de evolução de uma doença em pacientes hospitalizados

Definição do objetivo

- Primeiro passo de um projeto de *machine learning*
  1. Qual é o problema?
  2. Por que o problema precisa ser resolvido?
  3. Como eu resolvo o problema?

Qual o problema?

Sepse no Brasil - Fonte:Instituto Latino Americano de Sepse (ILAS)

1. Atinge cerca de 14% dos pacientes internados.
2. Responsável por 25% da ocupação em UTI's.
3. 50% dos casos vão a óbito.
4. Maior causa de morte e readmissão em hospitais.
5. Apenas 27% dos médicos sabem diagnosticar a sepse.

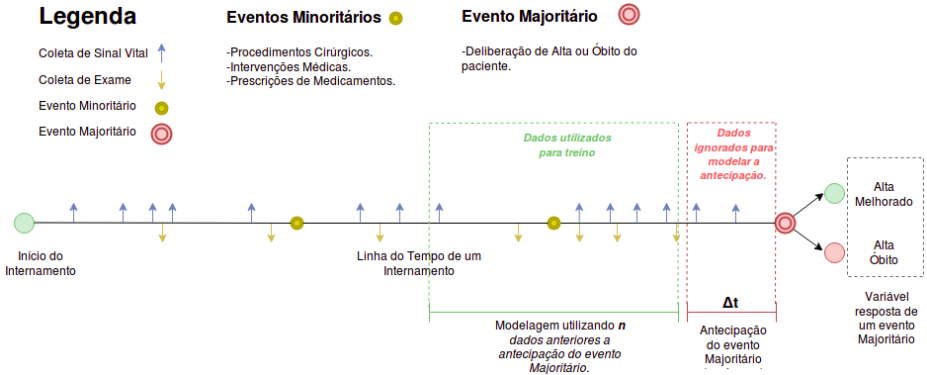
Qual o problema?

Sepse no Brasil - Fonte:Instituto Latino Americano de Sepse (ILAS)

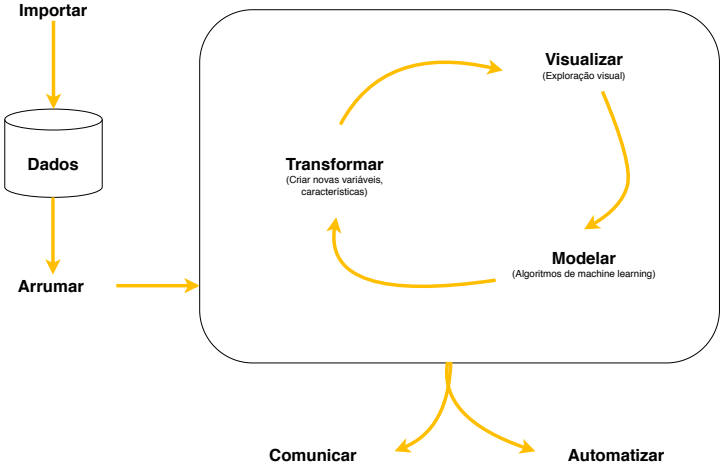
1. ~~Atinge cerca de 14% dos pacientes internados.~~
2. ~~Responsável por 25% da ocupação em UTI's.~~
3. 50% **dos casos vão a óbito.**
4. ~~Maior causa de morte e readmissão em hospitais.~~
5. ~~Apenas 27% dos médicos sabem diagnosticar a sepse.~~

Entenda o problema, depois pense como resolvê-lo

Objetivo

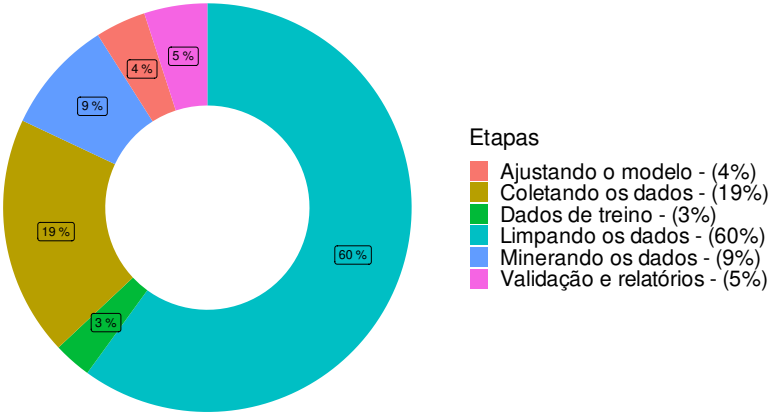

Legenda
Coleta de Sinal Vital
Coleta de Exame
Evento Minoritário
Evento Majoritário
Eventos Minoritários
-Procedimentos Cirúrgicos.
-Intervenções Médicas.
-Prescrições de Medicamentos.
Evento Majoritário
-Deliberação de Alta ou Óbito do paciente.
Dados utilizados para treino
Dados ignorados para modelar a antecipação.
Alta Melhorado
Alta Óbito
Início do Internamento
Linha do Tempo de um Internamento
Δt
Variável resposta de um evento Majoritário
Modelagem utilizando n dados anteriores a antecipação do evento Majoritário.
Antecipação do evento Majoritário

Como atingir o objetivo



Etapas em um ajuste de algoritmos de machine learning

Tempo dedicado a cada etapa



Resumo da prática

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise. Devemos saber onde queremos chegar!
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Base de dados

Os 3 pilares da ciência de dados

1. Objetivos
2. **Base de dados**
   - 2.1. **Coleta dos dados**
   - 2.2. **Exploração e preparação dos dados**
3. Modelos.

Base de dados - perguntas iniciais

1. Quais os tipos de dados você tem disponível?
2. Quais dados desejáveis não estão disponíveis?
3. Quais dados você não precisa para resolver o problema?

Base de dados - Fonte dos dados

- Banco de dados relacional (SQL)
  - somente dados estruturados

- Banco de dados não relacional (NoSQL)
  - dados estruturados e não estruturados

- Base de dados em nuvem (Cloud)
  - relacional e não relacional

Coleta de dados - Exemplo1

Coleta de dados - Exemplo2

Referências

1. Hastie, T., Tibshirani, R. e Friedman, J., **The Elements of Statistical Learning**, 2009.
2. James, G., Witten, D., Hastie, T. e Tibshirani, **An Introduction to Statistical Learning**, 2013.
3. L. Breiman. Statistical modeling: **The two cultures. Statistical Science**, 16(3):199-231, 2001.
4. Lantz, B., **Machine Learning with R**, Packt Publishing, 2013.
5. Tan, Steinbach, and Kumar, **Introduction to Data Mining**, Addison-Wesley, 2005.
6. Mitchell, T. M. (1997). **Machine Learning**. McGraw-Hill.
7. Wickham, H.; Grolemund, G. **R for data science: import, tidy, transform, visualize, and model data**. " O'Reilly Media, Inc.", 2016.