Minicurso Introdutório de Machine Learning

Pyladies Campinas

O que é *Machine Learning*?

"É a ciência que faz com que os computadores exerçam seu papel de forma natural, sem que pareçam explicitamente programados para tal."

Quando usar *Machine Learning*?

- Quando não se tem um conjunto de passos bem definidos
- Quando temos dados
  - Problemas complexos
  - Big data!

Como computadores funcionam?

- Operações lógicas e aritméticas (matemática)
- Memória
- Passos bem definidos (Algoritmos/Heurísticas)

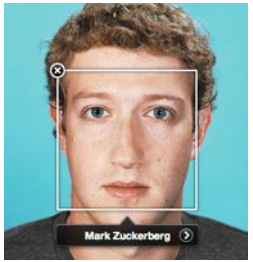
Para computeiras e computeiros:

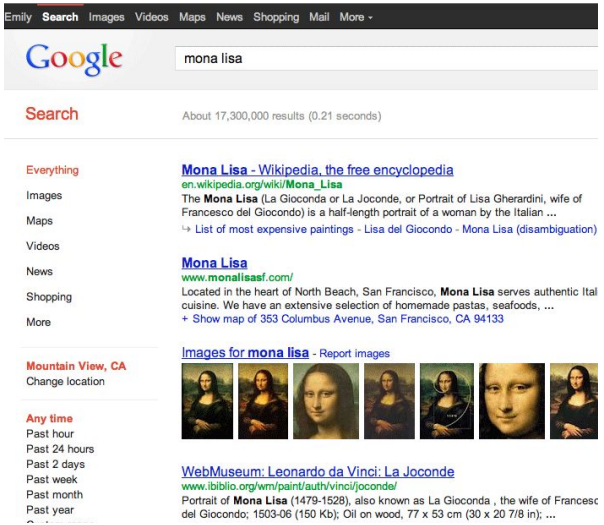
- Como detectar se a foto é num parque nacional?
- Como detectar se a foto tem um pássaro?

Figura extraída do blog xkcd.com

Algumas aplicações

- Detecção de spams em e-mails
- Reconhecimento facial (e.g. Facebook)
- Pesquisa personalizada (e.g. Google, Bing)
- Recomendações (e.g. Netflix, Amazon)





Por que aprender *Machine Learning*?

- Uma das carreiras que mais cresce na computação
  - Aplicações em diversas áreas (biologia, medicina, economia...)

- **Pode ser usada contra você!**
  - Grandes empresas
  - Governos

Caso de estudo: NSA's SKYNET (denunciado por Edward Snowden)

SKYNET (surveillance program)

From Wikipedia, the free encyclopedia

SKYNET is a program by the U.S. National Security Agency that performs machine learning analysis on communications data to extract information about possible terror suspects. The tool is used to identify targets, such as al-Qaeda couriers, who move between GSM cellular networks. These couriers often swap SIM cards within phones that have the same ESN, MEID or IMEI number.[1] The tool uses classification techniques like random forest analysis. Because the data set includes a very large proportion of true negatives and a small training set, there is a risk of overfitting. Bruce Schneier argues that a false positive rate of 0.008% would be low for commercial applications where "if Google makes a mistake, people see an ad for a car they don't want to buy" but "if the government makes a mistake, they kill innocents."[1]

Caso de estudo: NSA's SKYNET (II)

- Usa padrões como:
  - trocas de número (chip) de celular
  - o quanto você viaja
  - você morar no paquistão

- Para determinar se você **é um membro da Al-Qaeda!**
- Usa *(dentre várias coisas)* um algoritmo chamado ***random forests***

E o que são “random forests”?

- Um algoritmo de aprendizado de máquina
- Muito usado para **classificação** e **reconhecimento de padrões**

- Vamos falar mais dele daqui a pouco....

Classificações de Algoritmos de M.L.

- **Aprendizado supervisionado**
  - Classificação
  - Regressão
- **Aprendizado não supervisionado**
  - *“Clustering”*
- Aprendizado semi supervisionado
  - “Misturas”
- Aprendizado por reforço
  - Penaliza erros

Aprendizado supervisionado

- Conhecemos **algumas** respostas e queremos um modelo

Amostra	x1	x2	x3	y1
1	0,24	0,19	0,14	A
2	1,97	1,37	0,30	B
3	0,81	0,29	0,02	A
4	3,86	0,55	0,25	B
5	2,55	0,68	0,31	?
6	3,71	0,46	0,39	?
7	2,17	1,79	0,22	?
8	6,78	4,23	3,68	?

Conjunto de treinamento

Conjunto de testes

Classificação x Regressão

➢ **Classificação** (discreto)

x1	x2	x3	y1
0,24	0,19	0,14	A
1,97	1,37	0,30	B
0,81	0,29	0,02	A
3,86	0,55	0,25	B
2,55	0,68	0,31	?
3,71	0,46	0,39	?
2,17	1,79	0,22	?
6,78	4,23	3,68	?

➢ **Regressão** (contínuo)

x1	x2	x3	y1	y2
0,07	0,02	0,02	0,13	0,52
0,49	0,31	0,05	1,15	1,82
1,19	0,23	0,08	1,78	2,18
2,04	1,26	0,00	3,41	3,36
4,49	0,46	0,36	?	?
4,05	1,63	1,24	?	?
5,72	0,68	0,10	?	?
6,62	3,63	0,01	?	?

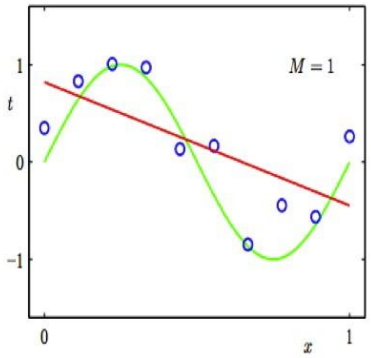
Exemplos de algoritmos

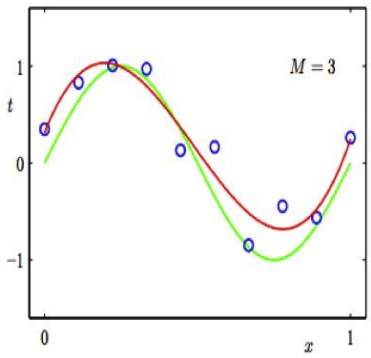
- É preciso entender detalhes dos algoritmos?
  - Depende, mas vamos deixar isso de lado por enquanto

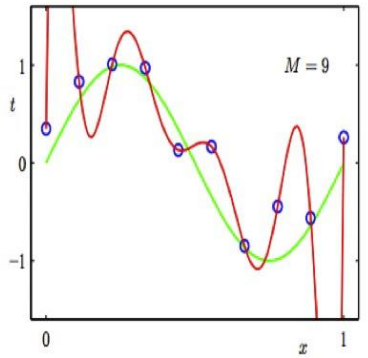
- Exemplo 1 no Ipython notebook
  - SVMs
  - *Random Forests*
  - *Multi-layer Perceptron*

Problema: por que ficou pior?

Regression:



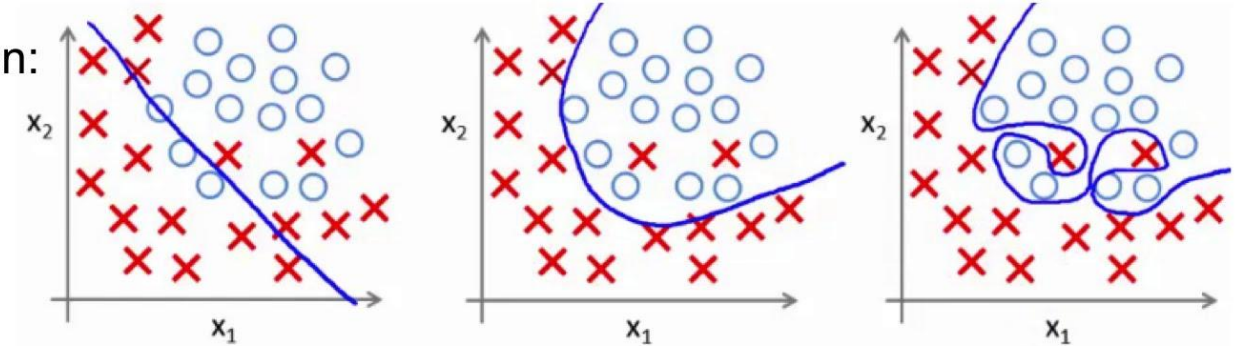




predictor too inflexible:
cannot capture pattern

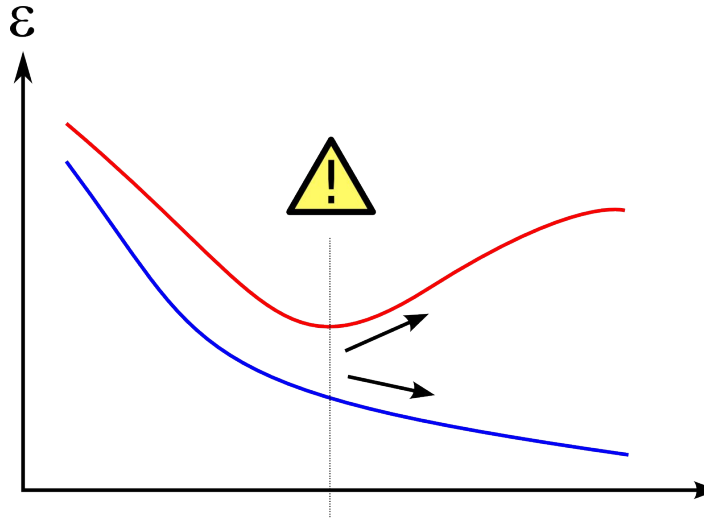
predictor too flexible:
fits noise in the data

Classification:



Overfitting: soluções

- Parar de treinar na hora certa
  - limitar o número de iterações
- Penalizar modelos complexos demais
  - parâmetros de penalização



Qual algoritmo escolher?

- Depende
  - Essa pergunta vale milhões

- Dois tipos básicos
  - Linearmente separáveis
  - Não linearmente separáveis

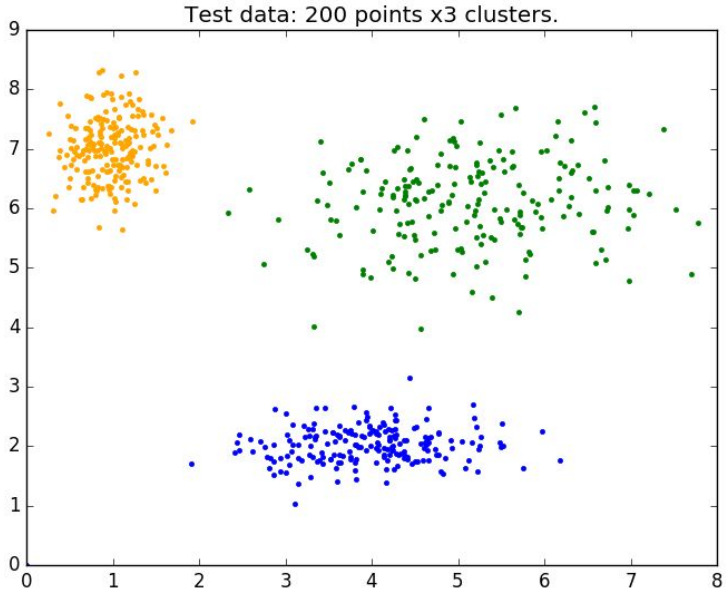
- Algumas dicas

Machine Learning é programar ?

- Sim e não!
- Programamos
  - os algoritmos de *machine learning*
  - métodos de pré-processamento
- O resultado não é programável
  - Uma estrutura matemática complexa gerada pelos dados
  - A programadora ou programador não tem controle

Aprendizado não supervisionado

- ***Clustering*:**
  - Não conhecemos nenhum exemplo
  - Queremos agrupar coisas



Exemplos de algoritmos

- Exemplo 2 no Ipython notebook
  - K-means
  - Métodos semi-supervisionados

Obrigada!

Pyladies Campinas