GEX1082 - Tópicos Especiais em Computação XXXIII
Deep Learning

Introdução ao Machine Learning

1100/1101 - CIÊNCIA DA COMPUTAÇÃO

Prof. Dr. Giancarlo D. Salton

O que é *Machine Learning*?

Como o *machine learning* funciona?

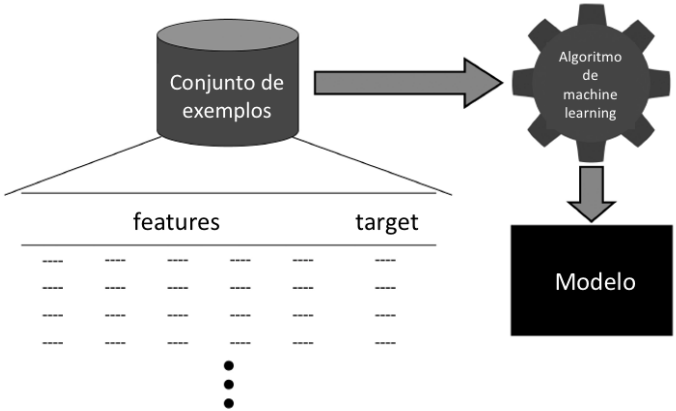
O que pode dar errado no *Machine Learning*?

O ciclo de vida do projeto de *machine learning*: Crisp-DM

Resumo

O que é *Machine Learning*?

- As técnicas (supervisionadas) de aprendizado de máquina aprendem de forma automática um modelo do relacionamento entre um conjunto de variáveis descritivas (***descriptive features***) e uma variável alvo (***target feature***) a partir de um conjunto de exemplos históricos (***training dataset***).



Utilizando *machine learning* para induzir um modelo preditivo de um conjunto de exemplos históricos.

Por quê precisamos dos exemplos?

X	Y	Z
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5

- Multiplicação?
- Divisão?
- Exponencial?

Por quê precisamos dos exemplos?

X	Y	Z
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
2	2	4
5	3	15

- Multiplicação!

Por quê precisamos dos exemplos?

X	Y	Z
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
2	2	4
5	3	15
2	2	5
5	3	72

- wtf?????

Utilizando o modelo aprendido para fazer predições sobre novas instâncias que não possuem a resposta (*target*) definido.

ID	Profissão	Idade	Proporção Salário-Empréstimo	Classe
1	indústria	34	2.96	pago
2	autônomo	41	4.64	atraso
3	autônomo	36	3.22	atraso
4	autônomo	41	3.11	atraso
5	indústria	48	3.80	atraso
6	indústria	61	2.52	pago
7	autônomo	37	1.50	pago
8	autônomo	40	1.93	pago
9	indústria	33	5.25	atraso
10	indústria	32	4.15	atraso

- Qual a relação entre as **variáveis descritivas** Profissão, Idade, Proporção Salário-Empréstimo e a **variável alvo** Classe no *dataset* acima?

if Proporção Salário-Empréstimo > 3 **then**
  Classe=*'atraso'*
else
  Classe=*'pago'*
end if

if Proporção Salário-Empréstimo > 3 **then**
  Classe=*'atraso'*
else
  Classe=*'pago'*
end if

- Este é um exemplo de um **modelo preditivo**
- Este é também um exemplo de um modelo preditivo **consistente**
- Perceba que este modelo não utiliza todas as variáveis e a variável utilizada é uma "variável derivada" (neste caso, uma proporção): **design de *features*** e **seleção de *features*** são dois temas importantes e voltaremos a eles várias vezes.

- Qual a relação entre as colunas contendo **variáveis descritivas** e a coluna Classe (***target feature***)?

ID	Valor	Renda Anual	Razão Renda-Empr.	Idade	Profissão	Propriedade	Classe
1	245,100	66,400	3.69	44	indústria	fazenda	pago
2	90,600	75,300	1.2	41	indústria	fazenda	pago
3	195,600	52,100	3.75	37	indústria	fazenda	atraso
4	157,800	67,600	2.33	44	indústria	apto.	pago
5	150,800	35,800	4.21	39	autônomo	apto.	atraso
6	133,000	45,300	2.94	29	indústria	fazenda	atraso
7	193,100	73,200	2.64	38	autônomo	casa	pago
8	215,000	77,600	2.77	17	autônomo	fazenda	pago
9	83,000	62,500	1.33	30	autônomo	casa	pago
10	186,100	49,200	3.78	30	indústria	casa	atraso
11	161,500	53,300	3.03	28	autônomo	apto.	pago
12	157,400	63,900	2.46	30	autônomo	fazenda	pago
13	210,000	54,200	3.87	43	autônomo	apto.	pago
14	209,700	53,000	3.96	39	indústria	fazenda	atraso
15	143,200	65,300	2.19	32	indústria	apto.	atraso
16	203,000	64,400	3.15	44	indústria	fazenda	pago
...	...	...	...	...	...	...	...

if Proporção Salário-Empréstimo < 1.5 **then**
  Classe=*'pago'*
else if Proporção Salário-Empréstimo > 4 **then**
  Classe=*'atraso'*
else if Idade < 40 **and** Profissão $=$*'industria'* **then**
  Classe=*'atraso'*
else
  Classe=*'pago'*
end if

if Proporção Salário-Empréstimo < 1.5 **then**
  Classe=*'pago'*
else if Proporção Salário-Empréstimo > 4 **then**
  Classe=*'atraso'*
else if Idade < 40 **and** Profissão $=$*'industria'* **then**
  Classe=*'atraso'*
else
  Classe=*'pago'*
end if

- O valor real do *machine learning* se torna aparente em situações como essa quando queremos criar modelos preditivos a partir de grandes conjuntos de dados com muitas *features*.

Como o *machine learning* funciona?

- Os algoritmos de *machine learning* funcionam pesquisando dentre um conjunto de possíveis modelos de previsão aquele modelo que melhor captura a relação entre as *features* e o *target*.
- Um critério óbvio de pesquisa é procurar modelos que sejam consistente com os dados.
- Contudo, devido ao fato de que um *dataset* é sempre uma amostra, *machine learning* é considerado um problema mal-posto.

Consistência (1)

ID	Filhos	Álcool	Orgânicos	Grupo
1	não	não	não	?
2	não	não	sim	?
3	não	sim	não	?
4	não	sim	sim	?
5	sim	não	não	?
6	sim	não	sim	?
7	sim	sim	não	?
8	sim	sim	sim	?

$$grupo = \{ \text{casal, família, solteiro} \}$$

Consistência (2)

Flh	Álc	Org	Grp	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
não	não	não	?	casal	casal	solteiro	casal	casal		casal
não	não	sim	?	solteiro	casal	solteiro	casal	casal		solteiro
não	sim	não	?	família	família	solteiro	solteiro	solteiro		família
não	sim	sim	?	solteiro	solteiro	solteiro	solteiro	solteiro		casal
sim	não	não	?	casal	casal	família	família	família	...	família
sim	não	sim	?	casal	família	família	família	família		casal
sim	sim	não	?	solteiro	família	família	família	família		solteiro
sim	sim	sim	?	solteiro	solteiro	família	família	casal		família

- 6561 soluções possíveis!

Consistência (3)

ID	Filhos	Álcool	Orgânicos	Grupo
1	não	não	não	casal
2	sim	não	sim	família
3	sim	sim	não	família
4	não	não	sim	casal
5	não	sim	sim	solteiro

Consistência (4)

Flh	Álc	Org	Grp	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
não	não	não	?	casal	casal	solteiro	casal	casal		casal
não	não	sim	?	solteiro	casal	solteiro	casal	casal		solteiro
não	sim	não	?	família	família	solteiro	solteiro	solteiro		família
não	sim	sim	?	solteiro	solteiro	solteiro	solteiro	solteiro		casal
sim	não	não	?	casal	casal	família	família	família	...	família
sim	não	sim	?	casal	família	família	família	família		casal
sim	sim	não	?	solteiro	família	família	família	família		solteiro
sim	sim	sim	?	solteiro	solteiro	família	família	casal		família

- Só neste slide, 3 soluções consistentes estão visíveis!

Consistência (5)

Flh	Álc	Org	Grp	$\mathbb{M}_2$	$\mathbb{M}_4$	$\mathbb{M}_5$	...
não	não	não	?	casal	casal	casal	
não	não	sim	?	casal	casal	casal	
não	sim	não	?	família	solteiro	solteiro	
não	sim	sim	?	solteiro	solteiro	solteiro	
sim	não	não	?	casal	família	família	
sim	não	sim	?	família	família	família	
sim	sim	não	?	família	família	família	
sim	sim	sim	?	solteiro	família	casal	

- Só neste slide, 3 soluções consistentes estão visíveis!

- Consistência $\approx$ memorização
- Consistência não é desejável especialmente quando há erros nos dados
- Objetivo: modelo que seja "genérico" e funcione além do *dataset* usado para aprender o modelo
- Então, quais critérios devem ser usados para escolher um dentre vários modelos?

- Bias indutivo é o conjunto de suposições que definem os critérios de seleção de modelos de um algoritmo de *machine learning*.
- São dois os tipos de bias que podem ser usados:
  1. bias de restrição
  2. bias de preferência
- Bias indutivo é necessário para que se consiga a generalização para além do *dataset* de treino.

Como o *Machine Learning* funciona (resumo)

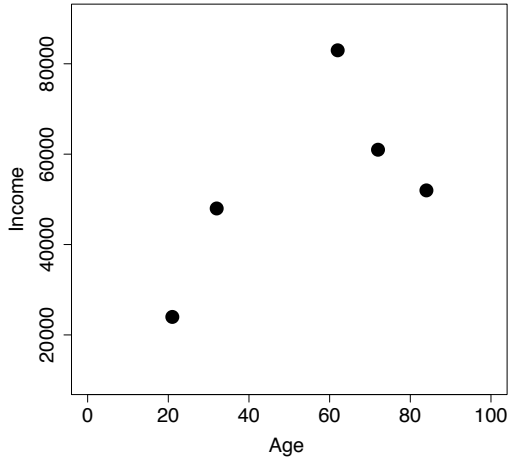
- Algoritmo de *Machine Learning* funcionam procurando por em um conjunto de modelos potenciais.
- Duas fontes de informação guiam a busca:
  1. o *dataset* de treino,
  2. o bias indutivo do algoritmo

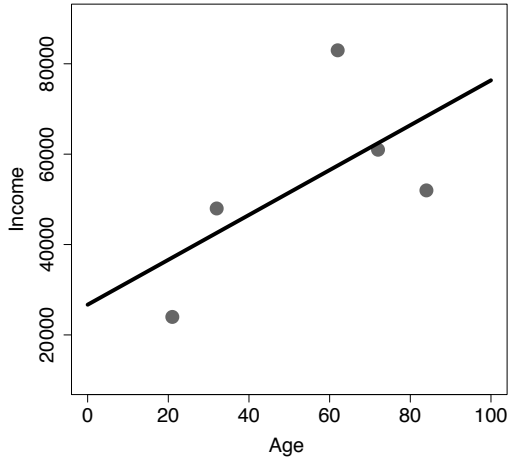
O que pode dar errado no *Machine Learning*?

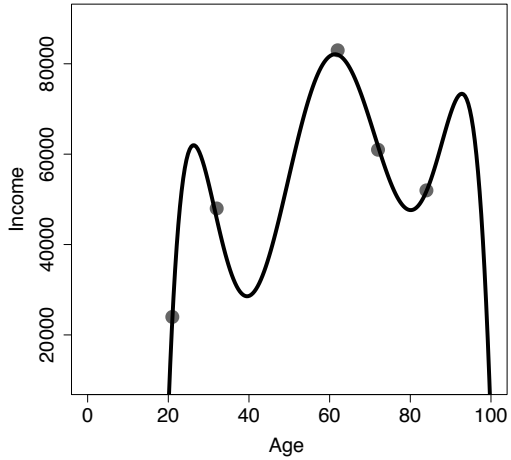
- Não existe bala de prata!
- O que acontece quando se escolhe o bias indutivo inadequado:
  1. **underfitting**
  2. **overfitting**

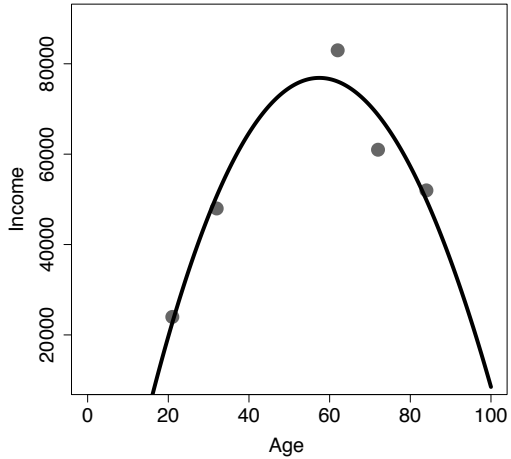
Tabela: O *dataset age-income*.

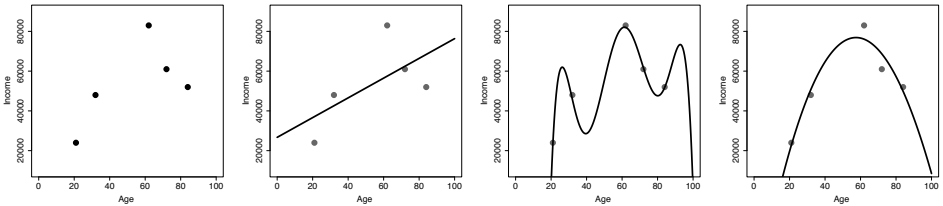
ID	Age	Income
1	21	24.000
2	32	48.000
3	62	83.000
4	72	61.000
5	84	52.000


Income
20000
40000
60000
80000
0
20
40
60
80
100
Age


Income
20000
40000
60000
80000
0
20
40
60
80
100
Age


Income
20000
40000
60000
80000
0
20
40
60
80
100
Age


Income
20000
40000
60000
80000
0
20
40
60
80
100
Age



(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

Encontrando o equilíbrio entre complexidade e simplicidade do modelo (*i.e.*, entre underfitting e overfitting) ao tentar predizer o *target* income a partir da *feature* age.

- Existem vários tipos de algoritmos de *machine learning*.
- Nós vamos nos concentrar em quatro famílias de algoritmos:
  - ✓ **baseados em informação**
  - ✓ **baseados em similaridade**
  - ✓ **baseados em probabilidade**
  - ✓ **baseados em erro**

O ciclo de vida do projeto de *machine learning*: Crisp-DM

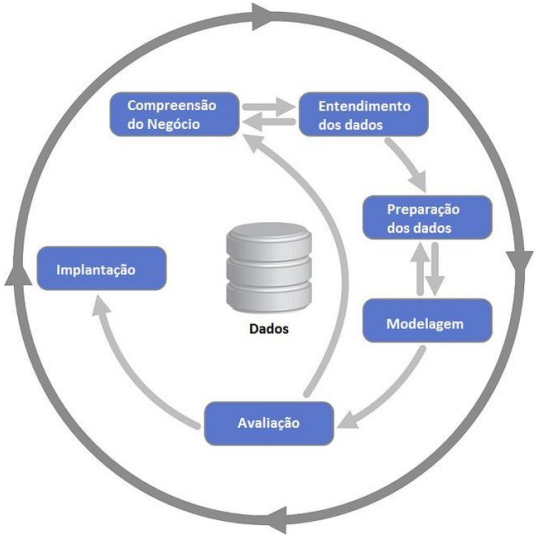


Diagrama do CRISP-DM demonstra as seis fases e indica as relações entre elas.

Resumo

- Técnicas de *machine learning* aprendem automaticamente as relações entre ***features* descritivas** e um ***target*** a partir de um conjunto de exemplos históricos.
- *Machine Learning* é um problema **mal-posto**:
  1. **generalização**,
  2. **bias indutivo**,
  3. ***underfitting***,
  4. ***overfitting***.
- Encontrar o equilíbrio entre complexidade e simplicidade do modelo (*i.e.*, entre *underfitting* e *overfitting*) é a parte mais difícil do *machine learning*.

O que é *Machine Learning*?

Como o *machine learning* funciona?

O que pode dar errado no *Machine Learning*?

O ciclo de vida do projeto de *machine learning*: Crisp-DM

Resumo