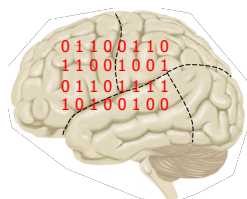


Inteligência Artificial

Mineração de Dados e Aprendizado de Máquina



Docente: Ricardo Cerri
Contém slides gentilmente cedidos pelos
Professores André C. P. L. F. de Carvalho e
Heloisa de Arruda Carmargo

1

Demanda

Want to improve Apple's music recommendation and playlisting services, and have a chance to influence the next generation of Apple products? We'd like to hear from strong scientific engineers who'd be interested in joining us in London. You'll need to have a good knowledge of machine learning (but hopefully that's why you're reading this list) and experience of working at scale. And to be a serious music lover.

Contact me directly if this sounds like you, or check out <http://www.apple.com/jobs> if you're interested in other opportunities with Apple.

Mark Levy
Applied Researcher, Apple

2

Oportunidades

- "Data Scientist: The Sexiest Job of the 21st Century"
 - *Harvard Business Review*, Outubro de 2012
- Ajuda tomadores de decisão a mudar análise subjetiva para análise baseada em dados



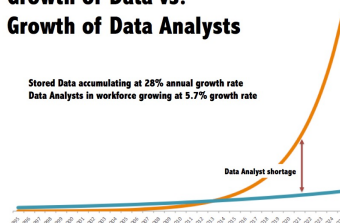
3

3

Falta de cientistas de dados

Growth of Data vs. Growth of Data Analysts

Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate



Fonte: www.deloitteanalytics.net

4

4




Tópicos

- Introdução
- Big Data
- Ciência de Dados
- Mineração de Dados
- Aprendizado de Máquina
- Métodos Preditivos
- Métodos Descritivos

5

5




Introdução

- Sem perceber, as pessoas geram dados a todo momento
 - Aplica para um cartão de fidelidade
 - Empresa aérea, supermercado, ...
 - Faz uma compra com cartão de débito ou crédito
 - Navega na internet
 - Vai ao médico
- Esses dados são armazenados em computadores (pessoais ou em nuvem)

6

6




Ciência de dados

- Big data
 - Lidar com grandes volumes de dados heterogêneos gerados a uma grande velocidade
 - Inclui BD e ciência de dados (*data science*)
- O que é ciência de dados?
 - Conjunto de princípios que dão suporte e guiam a extração de conhecimento de dados
 - Mineração de dados (MD) cria e utiliza técnicas que incorporam esses princípios

7

7



Introdução

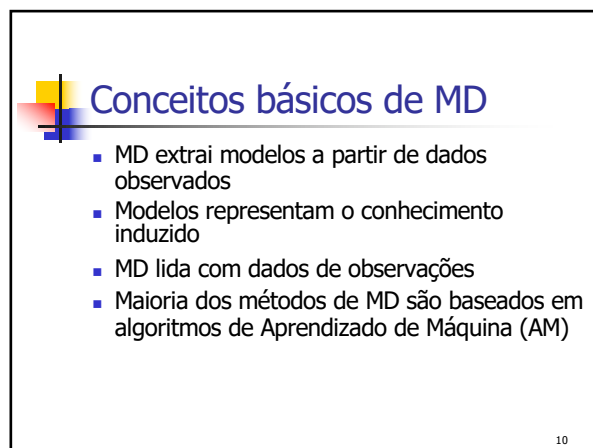
- Definições
 - *MD é a busca por informação valiosa em grandes volumes de dados*
(S. M. Weiss and N. Indurkha)
 - *MD é a análise de conjuntos de dados observacionais (geralmente grandes) para encontrar relacionamentos desconhecidos em novas formas que são ambos compreensíveis e úteis para o proprietário dos dados*
(D. Hand, H. Mannila and P. Smyth)

8

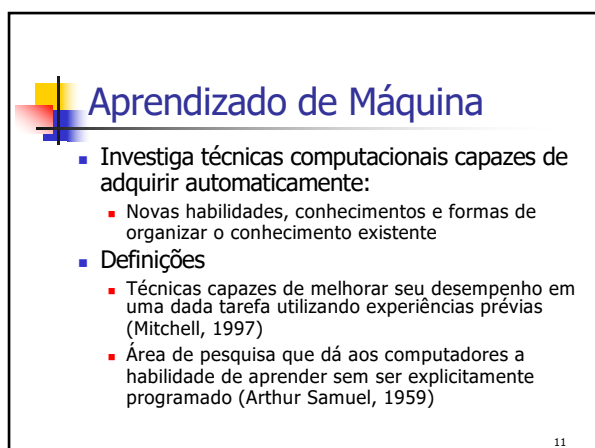
8



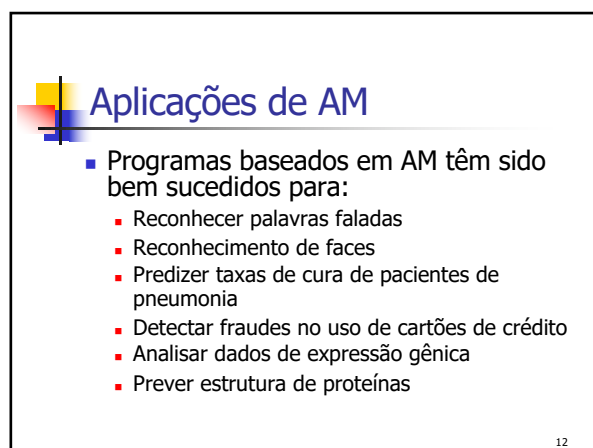
9



10



11



12

Aplicações clássicas de AM

- Aprender a reconhecer palavras faladas
 - SPHINX (Lee 1989)
- Aprender a conduzir um automóvel
 - ALVINN (Pomerleau 1989)
- Aprender a classificar objetos celestiais
 - (Fayyad et al 1995)
- Aprender a jogar gamão
 - TD-GAMMON (Tesauro 1992)

13

13

ALVINN



Dean Pomerleau
CMU

14

14

ALVINN

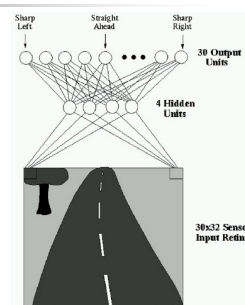
- Sistema automático de navegação para automóveis
 - Baseado em uma câmera montada no veículo
 - Dirigiu a 70 M/h (110 Km/h) em uma rodovia pública americana
 - De costa a costa em 1989 por 2850 milhas (com exceção de 50 milhas)

15

15

ALVINN

- Utiliza uma Rede Neural
 - 960 entradas
 - Matriz 30x32 derivada dos pixels de uma imagem
 - 4 unidades intermediárias
 - 30 unidades de saída
 - Cada uma representando um comando para a direção



16

16

Carros da Google

- Stanford Artificial Intelligence Laboratory
 - Sebastian Thrun
- Comunicação por sensor (topo do carro)
 - Recebe informação do Google street view
 - Atua no volante de direção e nos pneus
 - 175,000 milhas sem acidentes
- Califórnia, Flórida, Nevada e Michigan têm lei permitindo *driverless cars* (2013)

17

17

Carros da Google



<http://www.omg-facts.com/Technology/Google-has-developed-a-driverless-car/51099>

18

18

Carros da Google



<http://www.wired.com/autopia/2013/04/sergev-brins-mother-in-law/>

19

19

Carros da Tesla



<https://www.youtube.com/watch?v=UgNhYGAmZp>

20

20

Curiosidade

- Robô Mars
- NASA e Jet Propulsion laboratory
- Mais de 1 tonelada



21

21

Algoritmos de AM

- Grande número
 - Agrupamento de dados (K-médias)
 - Algoritmos de indução de Árvores de Decisão
 - K-NN
 - Máquinas de Vetores de Suporte
 - Naive Bayes
 - Raciocínio Baseado em Casos
 - Redes Neurais Artificiais
 - Sistemas Inteligentes Híbridos

22

22

Algoritmos de AM

- Podem ser agrupados por diferentes critérios
 - Baseados em distâncias
 - K-NN
 - Baseadas em otimização
 - RNs
 - Baseados em probabilidade
 - NB
 - Baseadas em procura (lógicos)
 - Indução de ADs
 - Evolutivo – teoria da evolução de Darwin
 - Algoritmos Genéticos

23

23

Aprendizado Indutivo

- Independente do paradigma utilizado, a grande maioria das estratégias de aprendizado de máquina realiza o que é chamado de **aprendizado indutivo**
- **Indução**: forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos.
- Um **conceito** é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados.
- **Hipóteses** geradas podem ou não preservar a verdade

24

24

Aprendizado Indutivo

- Aprender fazendo generalizações sobre casos específicos;
- Indução é uma forma de inferência lógica.

Exemplo 1
 Exemplo 2
 Exemplo 3
 ...
 Exemplo n

Indução

Hipótese
 ou
 conceito

25

25

Viés indutivo

- Indução de hipóteses
 - Aprender a partir de um conjunto de exemplos
 - Induzir modelo ou hipótese
 - Aplicado depois a novos dados (dedução)
 - Todo algoritmo de AM indutivo tem um viés
 - Tendência a privilegiar uma dada hipótese ou um dado conjunto de hipóteses

26

26

Viés indutivo

- Pode ser:
 - Viés de preferência ou busca
 - Como as hipóteses são pesquisadas no espaço de hipóteses
 - Preferência de algumas hipóteses sobre outras
 - Ex.: preferência por hipóteses simples (curtas)
 - Viés de representação ou linguagem
 - Define o espaço de busca ou de hipóteses
 - Restrição das hipóteses que podem ser geradas
 - Ex.: hipóteses podem conter apenas regras conjuntivas

27

27

Viés de representação

```

      graph TD
        A[Peso] -- "< 50" --> B[Sexo]
        A -- "≥ 50" --> C[Doente]
        B -- "M" --> D[Doente]
        B -- "F" --> E[Saudável]
      
```

Árvore de decisão

0.45	-0.40	0.54	0.12	0.98	0.37
-0.45	0.11	0.91	0.34	-0.20	0.83
-0.29	0.32	-0.25	-0.51	0.41	0.70

Redes neurais

Se Peso ≥ 50 então Doente
 Se Peso < 50 e Sexo = M então Doente
 Se Peso < 50 e Sexo = F então Saudável

Conjunto de regras

28

28

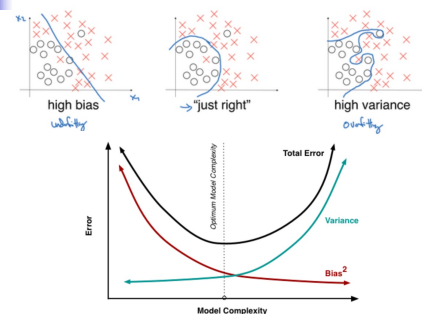
Viés indutivo

- Algoritmos de AM precisam ter um viés indutivo
 - Necessário para restringir o espaço de busca
 - Se não houvesse viés não haveria generalização
 - Regras / equações seriam especializados para os exemplos individuais

29

29

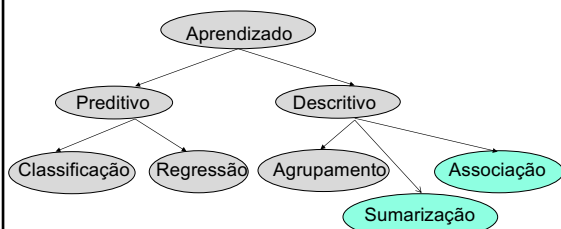
Relação Vies-Variância



30

30

Tarefas de aprendizado



31

31

Tipos de aprendizado

- Supervisionado
 - Modelo preditivo (mais comum) ou modelo descritivo
 - Ensinamos ao algoritmo o que ele deve fazer
 - Fornecemos, para cada entrada, a saída correta
- Não supervisionado
 - Modelo descritivo (mais comum) ou modelo preditivo
 - Algoritmo aprende por si só

32

32

Conjunto de Dados

Atributos de entrada (predictivos)

Nome	Temp.	Idade	Peso	Altura	
João	37	70	94	190	Saudável
Maria	38	65	60	172	Doente
José	39	19	70	185	Doente
Silvia	38	25	65	160	Saudável
Pedro	37	70	90	168	Doente

Exemplos (objetos, padrões)

Atributo alvo

33

33

Tipos de atributos

- Qualitativos
 - Nominais ou desordenados
 - Os valores são conhecidos, em pequena quantidade.
 - Podem ser associados com números mas não tem significado quantitativo.
 - Ex: cor, estado civil
 - Ordinais
 - Valores do atributo nominal podem ser colocados em uma ordem significativa.
 - A diferença entre estados sucessivos não é necessariamente a mesma, então não tem significado quantitativo.
 - Ex: avaliações qualitativas de temperatura como quente, médio e frio.

34

34

Tipos de atributos

- Quantitativos
 - Valores de escala intervalar
 - A diferença entre dois valores é significativa mas a razão entre eles não é.
 - Ex: temperatura – 20 graus é maior que 10 graus mas não faz sentido dizer que 20 graus é “duas vezes mais quente” que 10 graus.
 - Valores de escala radial
 - A proporção entre dois valores é significativa
 - Ex: peso – uma caixa que pesa 10 kg é duas vezes mais pesada que uma caixa que pesa 5 Kg.

35

35

Tipos de Atributos

- Muitas vezes fazemos a distinção apenas entre:
 - Atributos nominais (discretos, categóricos)
 - Atributos contínuos
- No aprendizado supervisionado fazemos distinção entre
 - Classificação:** quando o atributo classe é discreto
 - Regressão:** quando o atributo classe é contínuo.

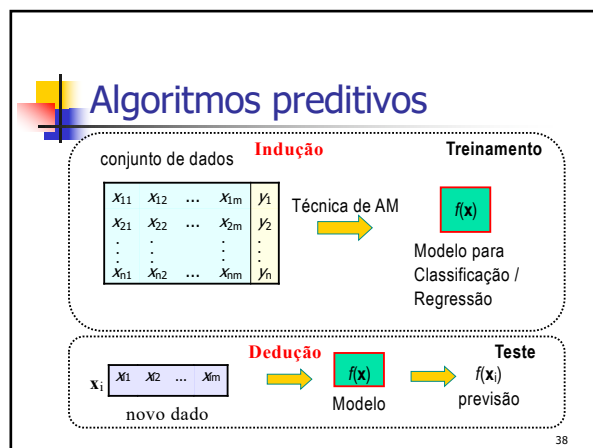
36

36

Algoritmos preditivos

- Induzem modelos (funções)
 - Após processo de treinamento
 - Dados de treinamento
- Modelo pode ser aplicado a novos dados
 - Dados de teste
 - Predição
- Classificação e regressão

37

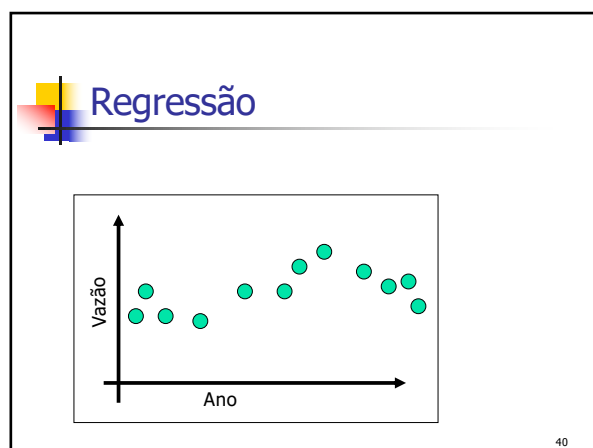


38

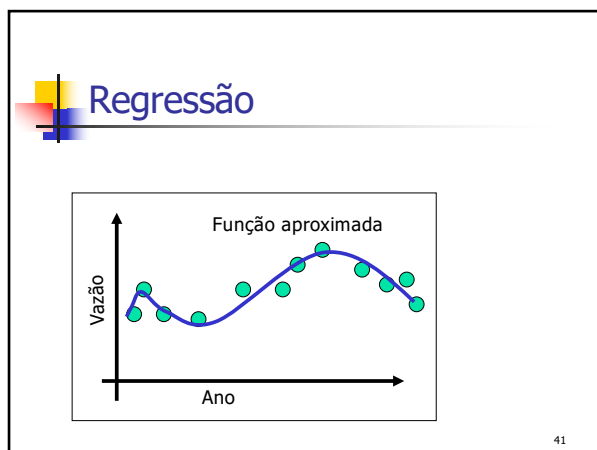
Regressão

- Objetivo: aprender uma função que mapeia um exemplo em um valor real
 - Caso especial: análise de séries temporais
- Exemplos:
 - Prever valor de mercado de um imóvel
 - Prever o lucro de um empréstimo bancário
 - Prever tempo de internação

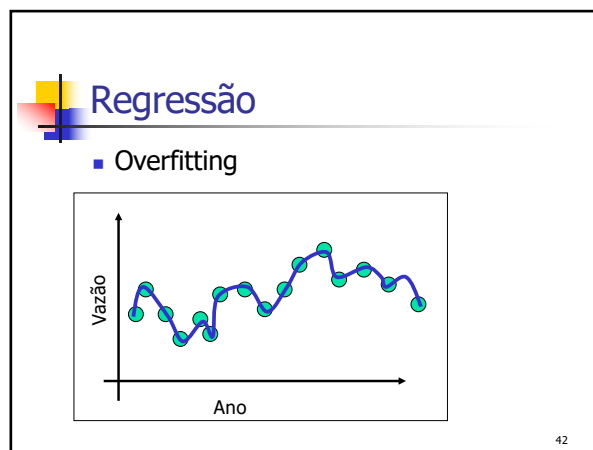
39



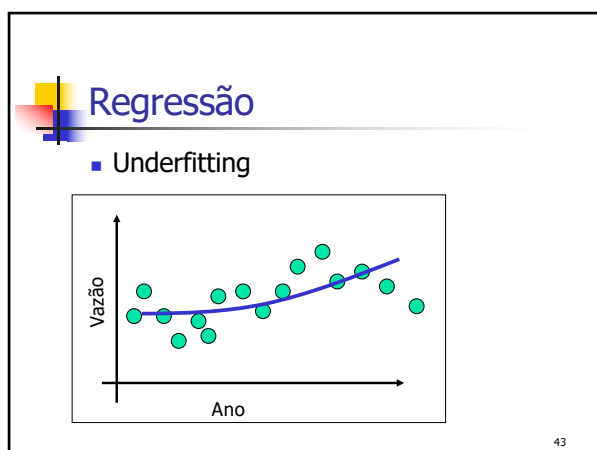
40



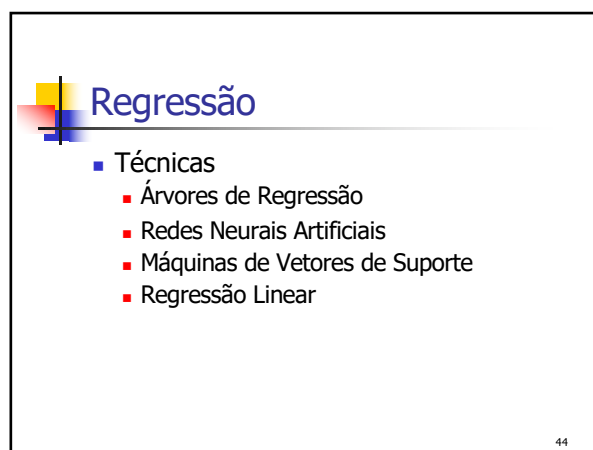
41



42



43



44

Exemplo - regressão

	potência	peso	aceleração	consumo
E1	180	3852	13,5	33
E2	175	3010	14,4	32
E2	82	2720	19,4	31

- Objetivo: encontrar padrões que permitam prever o consumo do carro.
- Exemplo:
Se potência > 170 e aceleração entre 12 e 13,9 então consumo = 33.

45

Exemplo - regressão

- Árvore de regressão
- Regras de regressão

46

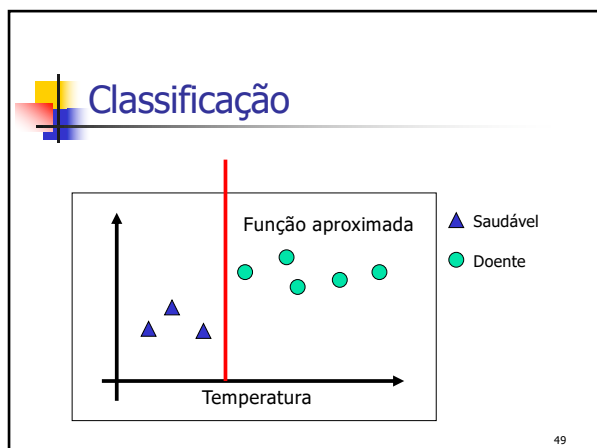
Classificação

- Objetivo: aprender uma função que mapeia um exemplo em uma dentre N classes
- Exemplos:
 - Definir a função de uma proteína
 - Classificar *email* como spam ou não
 - Definir se um paciente tem ou não uma doença

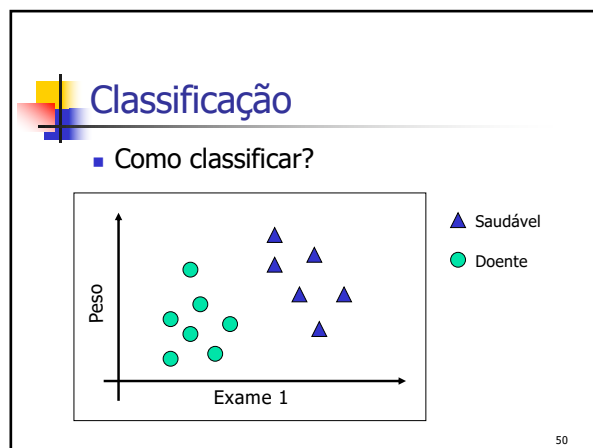
47

Classificação

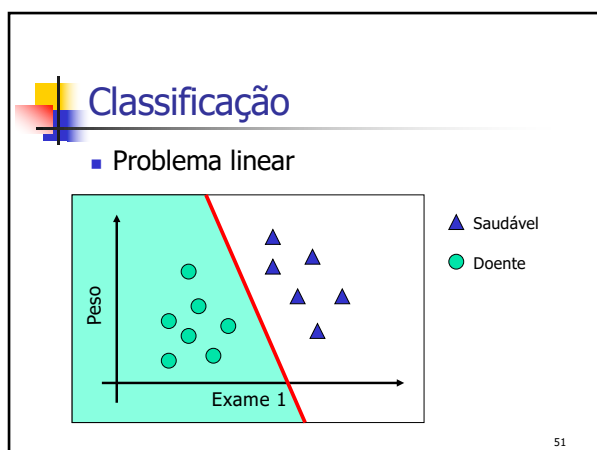
48



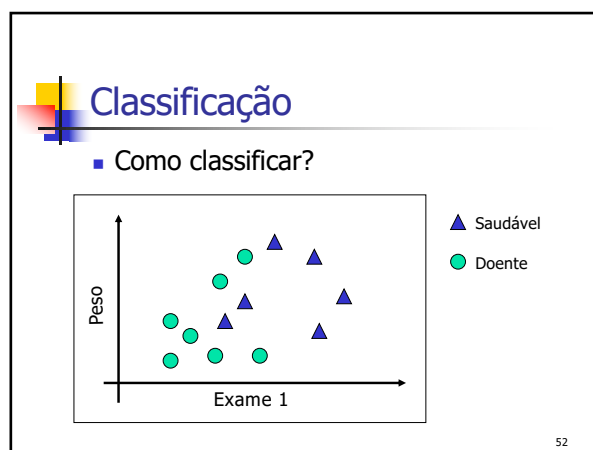
49



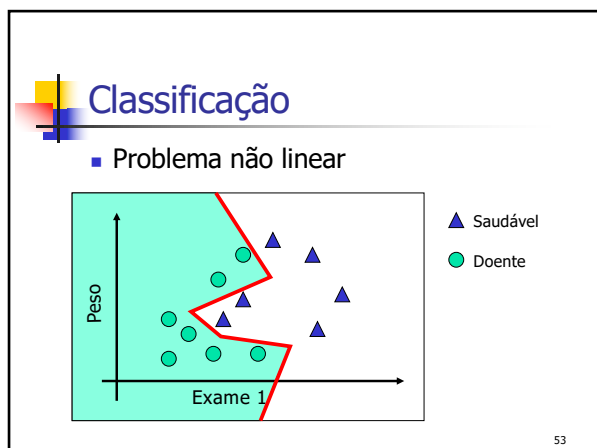
50



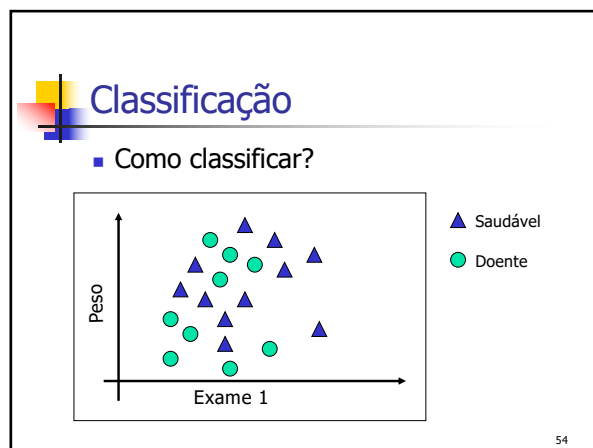
51



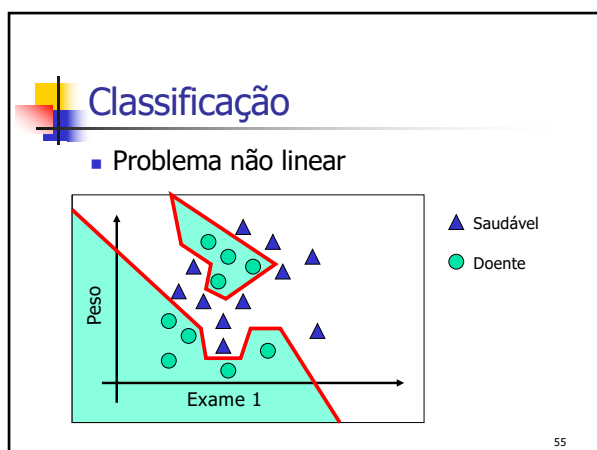
52



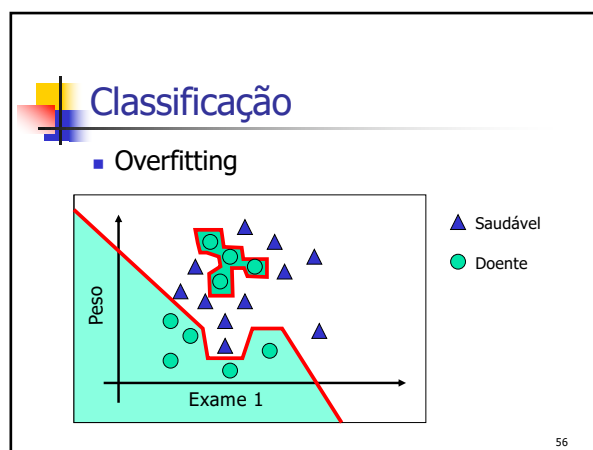
53



54



55



56

Classificação

- Algoritmos
 - Indução de Árvores de Decisão (C4.5)
 - Conjuntos de regras
 - Redes Neurais
 - Máquinas de Vetores de Suporte
 - K-NN
 - Regressão Logística
 - Redes Bayesianas

57

Exemplo - classificação

Dado	renda	dívida	classe
E1	30	40	1
E2	40	20	2
E3	80	60	2

Objetivo: encontrar padrões que permitam distinguir os clientes da classe 1 (devedor) da classe 2 (bom pagador).

Exemplo: **Se renda acima de 30 e dívida abaixo de 70 então classe 2.**

58

Exemplo - classificação

Em que formato o classificador é representado e como ele é usado para classificação?

- Árvores de decisão
- Regras de decisão

```

graph TD
    A((a=5)) -- sim --> B((b=7))
    A -- não --> C[c=2]
    B -- sim --> D[c=1]
    B -- não --> E[c=2]
  
```

Se $a = 5$ e $b = 7$ então $c = 1$
senão $c = 2$

59

Algoritmos descritivos

- Geram modelos em um processo de treinamento
 - Descrevem ou sumarizam dados
 - Treinamento utiliza todos o conjunto de dados
 - Agrupamento de dados
- Alguns algoritmos não utilizam treinamento
 - Regras de associação e sumarização

60

Formato dos dados – Tabela atributo-valor

Classe (rótulo) : não é conhecida

	X_1	X_2	...	X_m	
T_1	x_{11}	x_{12}	...	x_{1m}	y_1
T_2	x_{21}	x_{22}	...	x_{2m}	y_2
...
T_l	x_{n1}	x_{n2}	...	x_{nm}	y_n

Não conhecido

61

61

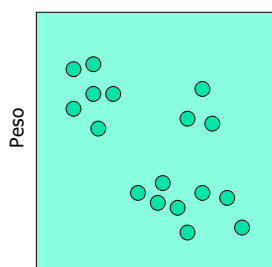
Agrupamento (Clustering)

- Objetivo: organizar exemplos não rotulados em grupos (clusters)
 - De acordo com uma medida de similaridade ou correlação entre eles
 - Aprendizado não supervisionado
- Não existe conhecimento anterior sobre:
 - Número de grupos
 - Significado dos grupos

62

62

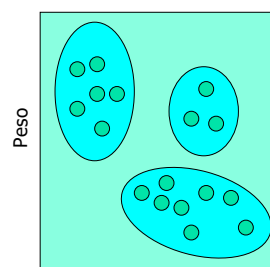
Agrupamento



63

63

Agrupamento



64

64

Agrupamento

- Técnicas
 - Redes Neurais SOM
 - K-médias
 - FCM
 - DBSCAN
 - Single-Link

65

Sumarização

- Objetivo: encontrar descrição simples e compacta para um conjunto de dados
- Frequentemente utilizada para:
 - Exploração interativa de dados
 - Geração automática de relatórios
 - Exemplo:
 - Definir o valor médio de compras feitas nos finais de semana em um supermercado

66

Sumarização

Nome	Idade	Sexo	Altura	Tem filhos
João	32	M	180	S
Maria	30	F	-----	N
Pedro	23	M	160	S
José	45	M	170	S
Sueli	18	F	175	N

67

Sumarização

Nome	Idade	Sexo	Altura	Tem filhos
João	32	M	180	S
Maria	30	F	-----	N
Pedro	23	M	160	S
José	45	M	170	S
Sueli	18	F	175	N

Idade média: 29.6
 Mediana da idade: 30
 Sexo mais frequente: M
 Maior altura: 180

68

Sumarização

- Técnicas podem ser divididas em:
 - Simples:
 - Média
 - Mediana
 - Desvio padrão
 - Mais sofisticadas:
 - Regras de sumarização
 - Técnicas de visualização multivariadas

69

Exercício

- Sumarizar cadastro de pacientes abaixo:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente
Luis	não	sim	grandes	sim	doente
Livia	não	não	pequenas	sim	saudável

70

Regras de Associação

- Objetivo: dado um conjunto de itens e uma base de dados de transações
 - Encontrar um conjunto de regras de associação entre os itens
- Exemplo:
 - Procurar por itens que são frequentemente comprados juntos

71

Regras de Associação

Transação	Itens comprados
1	pão, queijo, manteiga, massa
2	pão, geléia, suco
3	queijo, arroz, massa
4	queijo, vinho
5	massa, queijo, pão

72

Regras de Associação

Transação	Itens comprados
1	pão, queijo, manteiga, massa
2	pão, geléia, suco
3	queijo, arroz, massa
4	queijo, vinho
5	massa, queijo, pão

40% dos clientes compram pão e queijo
75% dos clientes que compram queijo também compram massa


73

Conclusão

- Mineração de Dados
- Aprendizado de Máquina
- Algoritmos
 - Viés indutivo
- Tarefas
 - Preditivas
 - Descritivas

74

Perguntas



75

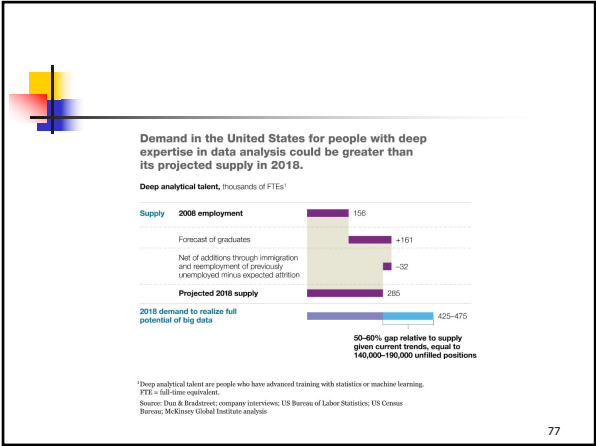
Links interessantes

Self-Driving Car Test:
<http://www.youtube.com/watch?v=cdqOpa1pUUE>

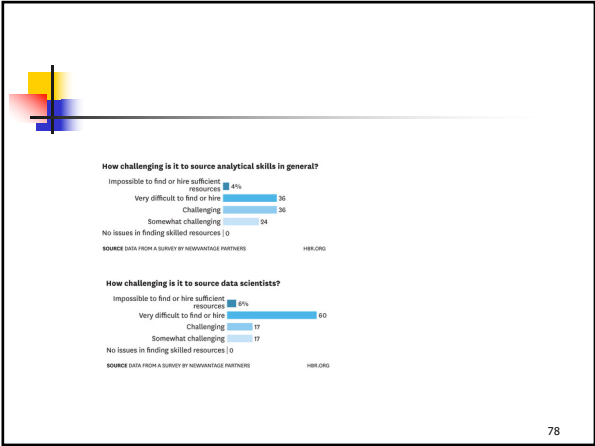
Nine reasons to be scared of Google:
<http://www.omg-facts.com/Technology/Google-has-developed-a-driverless-car/51099>

Paideia Entrevista – Prof. Adriano Polpo (DE/UFSCar) – Data Science:
<https://www.youtube.com/watch?v=iUmOSdZlptI>

76



77



78