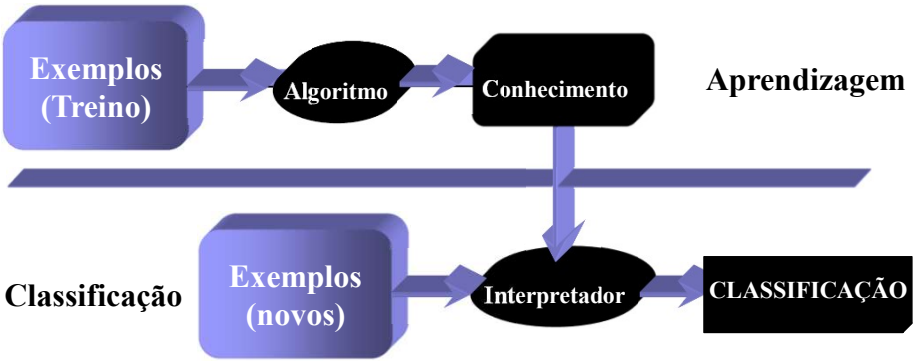
Introdução à aprendizagem

Aprender a partir dos dados conhecidos

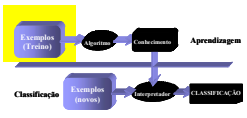
1

Fases do processo



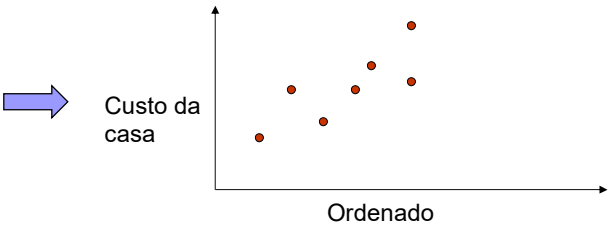
2

Exemplo de aprendizagem (1)



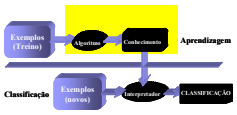
- Agência imobiliária pretende estimar qual a gama de preços para cada clinente
- Exemplos de treino:
  - Dados históricos
  - Ordenado vs custos de casas compradas

Nome	Custo	Ordn.
Manel	226.000	1.300
João	320.500	2.400
Abel	190.600	730
Carlos	850.000	2.300
:	:	:

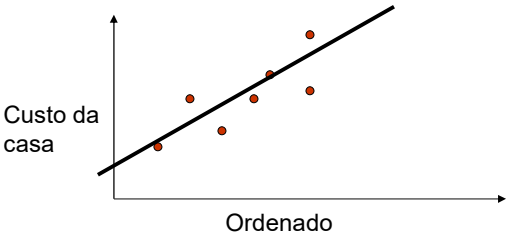


3

Exemplo de aprendizagem (2)



- Algoritmo
  - Regressão linear
- Representação do conhecimento
  - Recta (declive e ordenada na origem)

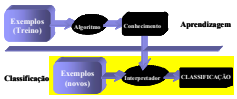


Ordenada na origem = b_0
Inclinação= α

Custo=α x Ordenado + b_0

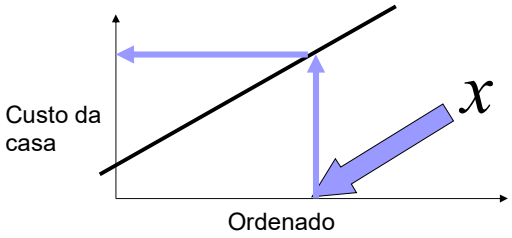
4

Exemplo de aprendizagem



(3)

- Exemplos novos
  - Um novo cliente, com ordenado x
- Interpretação
  - Usar a recta (método de previsão usado) para obter uma PREVISÃO



5

Outro problema de predição

- Exemplo da seguradora (seguros de saúde)
- Existe um conjunto de dados conhecidos
  - Conjunto de treino
- Queremos prever o que vai ocorrer noutros casos
  - Empresa de seguros de saúde quer estimar custos com um novo cliente

Conjunto de treino (dados históricos)

Altura	Peso	Sexo	Idade	Ordenado	Usa ginásio	Encargos para seguradora
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

E o Manel ?

Altura=1.73
Peso=85
Idade=31
Ordenado=2800
Ginásio=N

Terá encargos para a seguradora ?

6

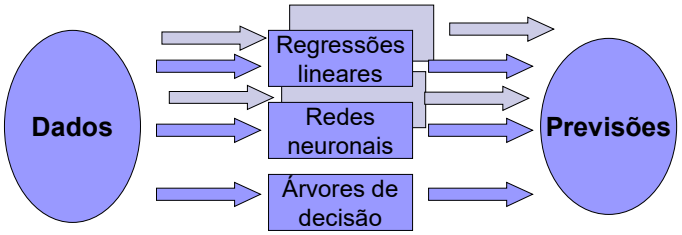
Modelos versus Dados (ciência versus dataminig?)

- Model based
  - Incorporam o conhecimento à priori
    - F=ma, PV=nRT
    - Conhecimento "certo" pelas "causas"
  - Eventualmente é necessário estimar algum parâmetro (mas poucos)
- Data driven
  - Procuram relações nos dados
    - Relações não implicam causa/efeito
  - Ou não há modelo, ou há um modelo genérico que normalmente é um aproximador universal (com muitos parâmetros)

7

Tipos de sistemas de previsão

- "Clássicos"
  - Regressões lineares, logísticas, etc...
- Vizinhos mais próximos
- Redes Neuronais
- Árvores de decisão
- Regras
- "ensembles"



8

Tipos de Aprendizagem

SUPERVISIONADA vs NÃO SUPERVISIONADA

INCREMENTAL vs BATCH

PROBLEMAS

9

Professor/Aluno

- Todo o processo de aprendizagem pode ser caracterizado por um protocolo entre o professor e o aluno.
- O professor pode variar entre o tipo dialogante e o não cooperante.

10

Protocolos Professor/Aluno

- Professor nada cooperante
  - Só dá os exemplos => **não supervisionada**
- Professor cooperante
  - Dá exemplos classificados => **supervisionada**
- Professor pouco cooperante
  - Só diz se os resultados estão certos ou errados => **aprendizagem por reforço**
- Professor dialogante - ORÁCULO

11

Formas de adquirir o conhecimento

- Incremental
  - Os exemplos são apresentados um de cada vez e a estrutura de representação vai-se alterando
- Não incremental (batch)
  - Os exemplos são apresentados todos ao mesmo tempo e são considerados em conjunto.

12

Acesso aos exemplos

- Aprendizagem “offline”
  - Todos os exemplos estão disponíveis ao mesmo tempo
- Aprendizagem “online”
  - Os exemplos são apresentados um de cada vez
- Aprendizagem mista
  - Uma mistura dos dois casos anteriores

13

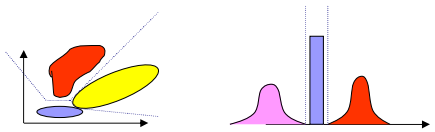
Problema do nº de atributos

- Poucos atributos
  - Não conseguimos distinguir classes
- Muitos atributos
  - Caso mais vulgar em Datamining
  - Praga da dimensionalidade
  - Visualização difícil e efeitos “estranhos”
- Atributos importantes vs redundantes
  - Quais os atributos importantes para a tarefa?

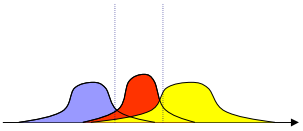
14

Problema da separabilidade

- Separáveis
  - Erro Ø possível



- Não separáveis
  - Erro sempre > Ø
  - Erro de Bayes
    - Erro mínimo possível para um classificador



15

Problema do "melhor" tipo de modelo

- A representação de conhecimento mais simples.
  - Mais fácil de entender
  - Árvores de decisão vs redes neuronais
- A representação de conhecimento com menor probabilidade de erro.
- A representação de conhecimento mais provável
  - Navalha de Occam ...

16

Problemas ...

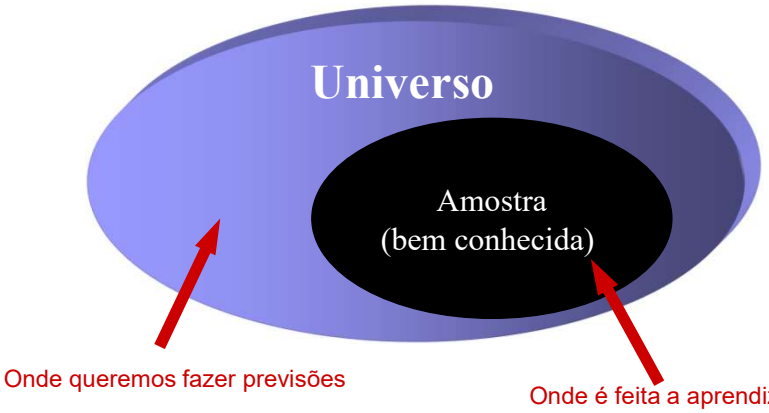
- Adequabilidade da representação do conhecimento à tarefa que se quer aprender
- Ruído
  - Ruído na classificação dos exemplos ou nos valores dos atributos.
  - Má informação é pior que nenhuma informação
- Enormes quantidades de dados
  - Quais são importantes? Tempo de processamento
- Aprender “demais”
  - Decorar os dados. Vamos ver isso agora...

17

Generalização e “overfitting”

18

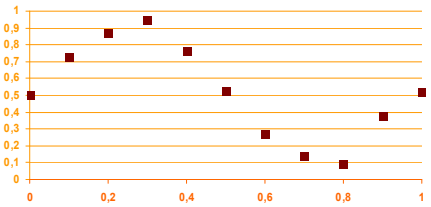
Os dados



19

Exemplo de overfitting

- Seja um conjunto de 11 pontos.
- Encontrar um polinómio de grau M que represente esses 11 pontos.



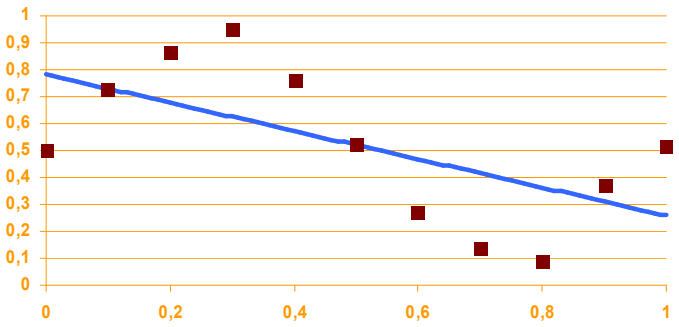
$$y(x) = \sum_{i=0}^{M} w_i x^i$$

20

Aproximação M = 1

$$y(x) = w_0 + w_1 x$$

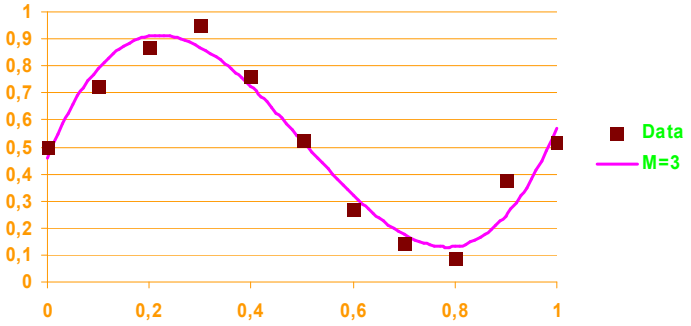
Erro grande



21

Aproximação M = 3

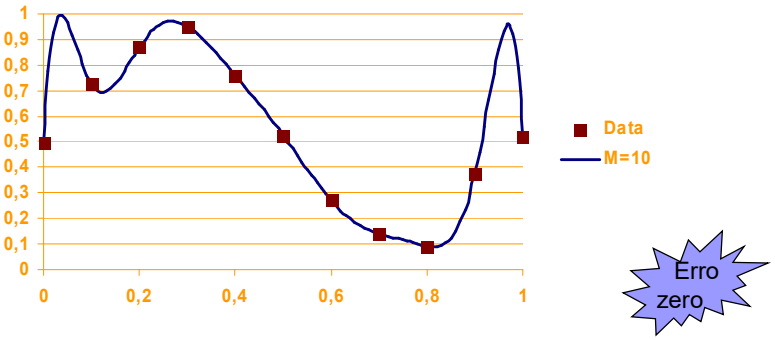
$$y(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$



22

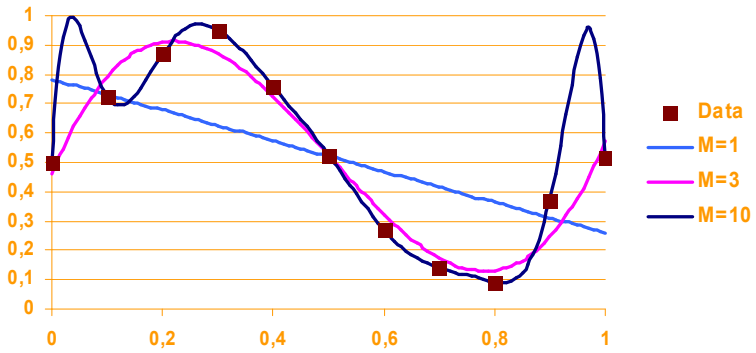
Aproximação M = 10

$$y(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8 + w_9 x^9 + w_{10} x^{10}$$



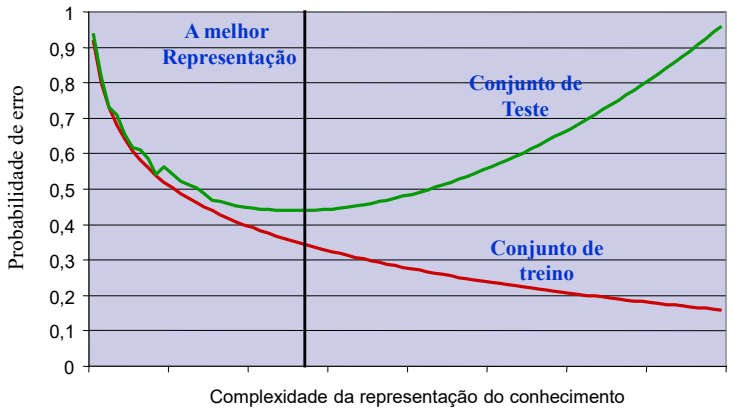
23

Overfitting



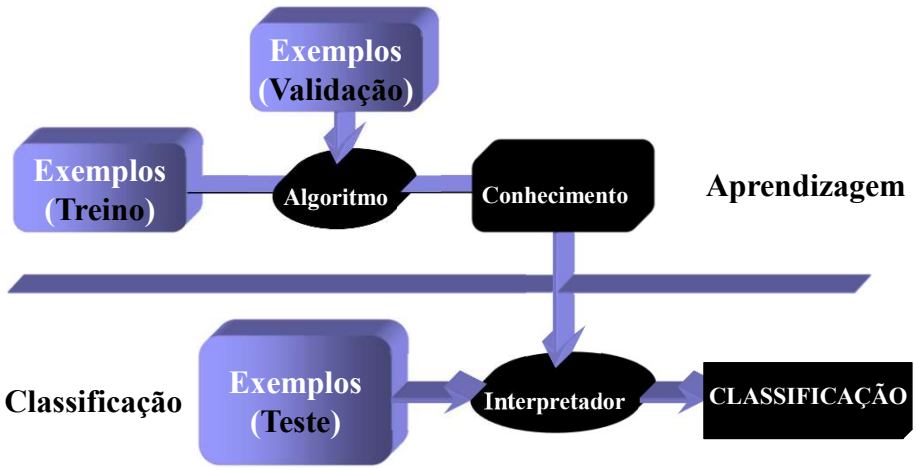
24

Curva de Overfiting



25

Fases do processo



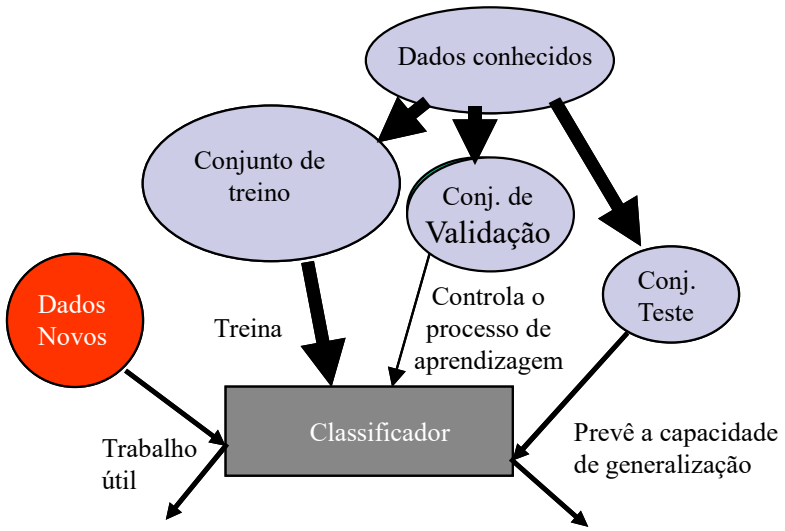
26

Generalização

- O objectivo não é aprender a agir no conjunto de treino mas sim no universo "desconhecido" !
  - Como preparar para o desconhecido ?
- Manter um conjunto de teste "de reserva"

27

Conjunto de treino/validação/teste



28

Divisão dos dados

- Conjunto de treino
  - Usado para construir o classificador
  - Quanto maior, melhor o classificador obtido
- Conjunto de validação
  - Usado para controlar a aprendizagem (opcional)
  - Quanto maior, melhor a estimação do treino óptimo
- Conjunto de teste
  - Usado para estimar o desempenho
  - Quanto maior, melhor a estimação do desempenho do classificador

29

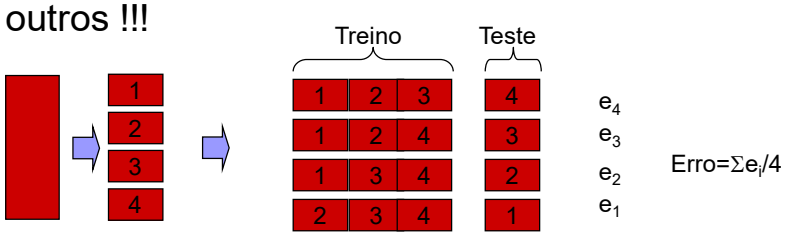
Estimativas do erro do classificador

- Em problemas de classificação
  - Taxa de erro = nº de erros/total (ou *missclassification error*)
  - Possibilidade de usar o "custo do erro"
- Em problemas de regressão
  - Erro quadrático médio, erro médio, etc…
- Estimativas optimistas ou não-enviesadas
  - Erro no conjunto de treino (erro de resubstituição)
    - Optimista
  - Erro no conjunto de validação
    - Ligeiramente optimista
  - Erro no conjunto de teste
    - Não enviesado. A melhor estimativa possível
    - (no entanto…se estes dados fossem usados para treino…)

30

Estimativas robustas do erro

- Validação cruzada
  - Cross-validation, ou *leave-n-out*
  - Dividir os mesmos dados em diferentes partições treino/teste
  - Calcular erro médio
  - Nenhum dos classificadores é melhor que os outros !!!



31

Outras medidas de erro em classificação

- **Matriz de confusão**
  - Separa os diversos tipos de erro
    - Falso Positivo (FP)
      - O classificador diz que é, e não é
    - Falso Negativo (FN)
      - O classificador não detecta que é
  - Permite compreender em que é que o classificador é bom

Matriz de Confusão	Classificado como SIM	Classificado como NÃO
Realmente é SIM	*TP*	*FN*
Realmente é NÃO	*FP*	*TN*

- Medidas de erro
  - Taxa de erro = (FP+FN)/n — Erro mais tradicional
  - Confiança positiva = TP/(TP+FP) — Quão "definitivo" é um resultado positivo (por vezes "precision")
  - Confiança negativa = TN/(TN+FN) — Quão "definitivo" é um resultado negativo
  - Sensibilidade = TP/(TP+FN) — Quão bom é a apanhar os positivos (por vezes "recall")
  - Precisão (acuracy) = (TP+TN)/n — O complementar da taxa de erro
  - Há mais medidas, adaptadas a cada problema em particular !

32

Processo de aprendizagem

- A aprendizagem é um processo de optimização (Minimização do erro)
- Algoritmo de optimização
  - Método do gradiente
  - Subir a encosta
  - Guloso
  - Algoritmos genéticos
  - "Simulated annealing"
- Formas de adquirir o conhecimento

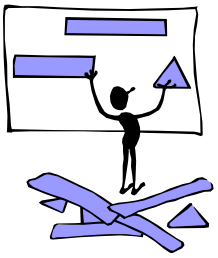
33

Iterações sucessivas do sistema de aprendizagem

34

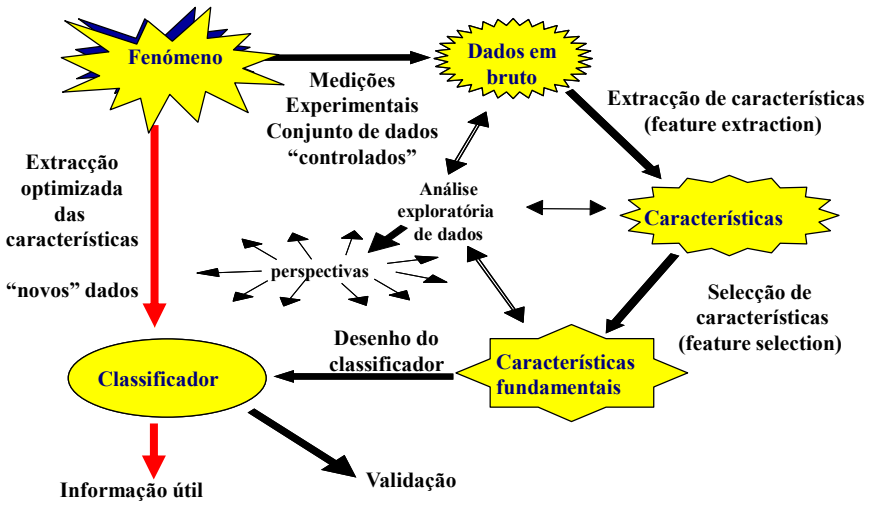
Tarefas do projecto do sistema

- Preparação dos dados.
- Redução dos dados.
- Modelação e predição dos dados.
- Casos e análise das soluções



35

Aproximação exploratória...

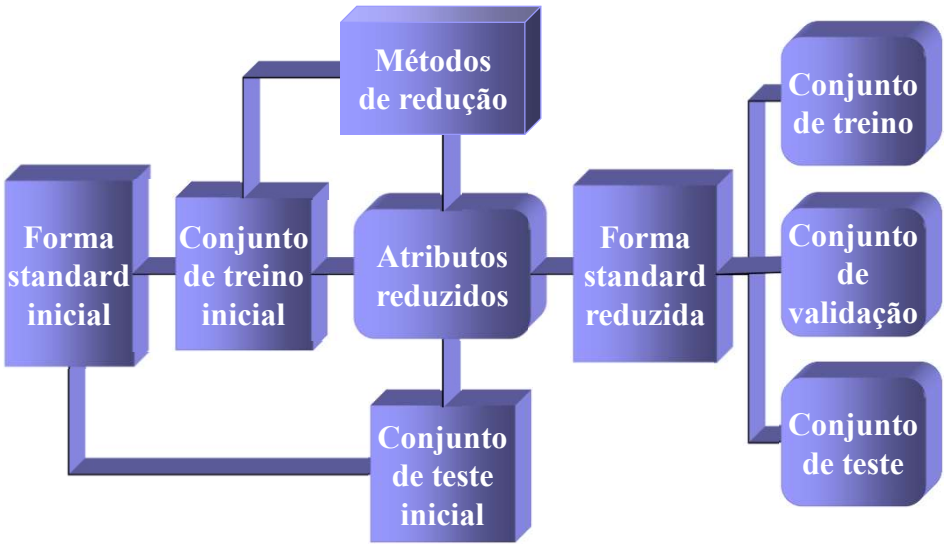


36

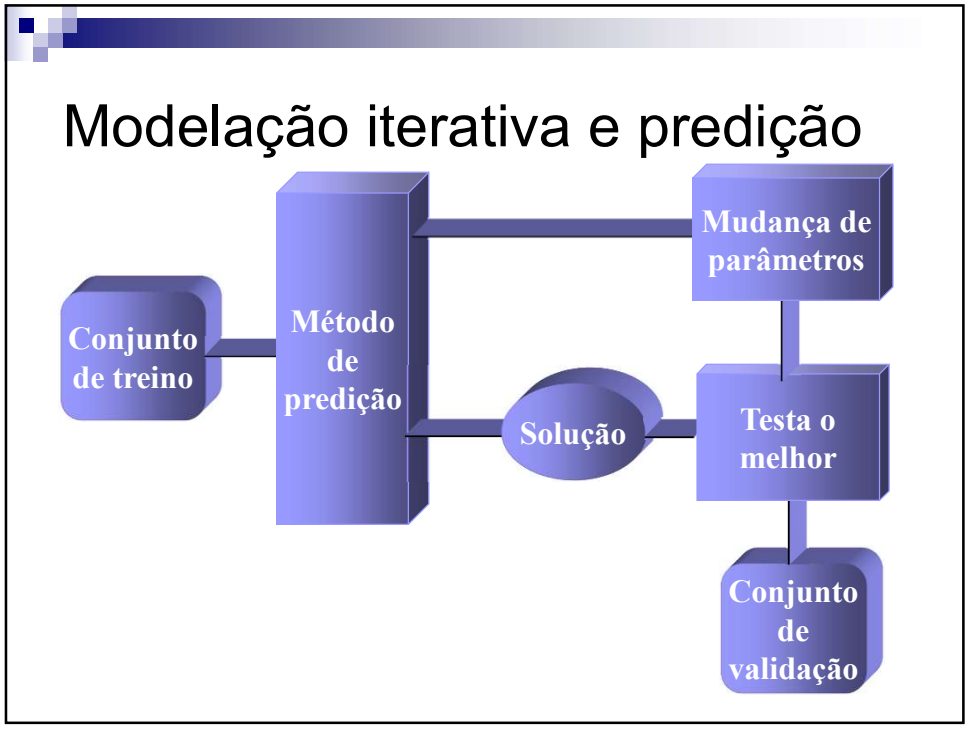
Preparação dos dados

37

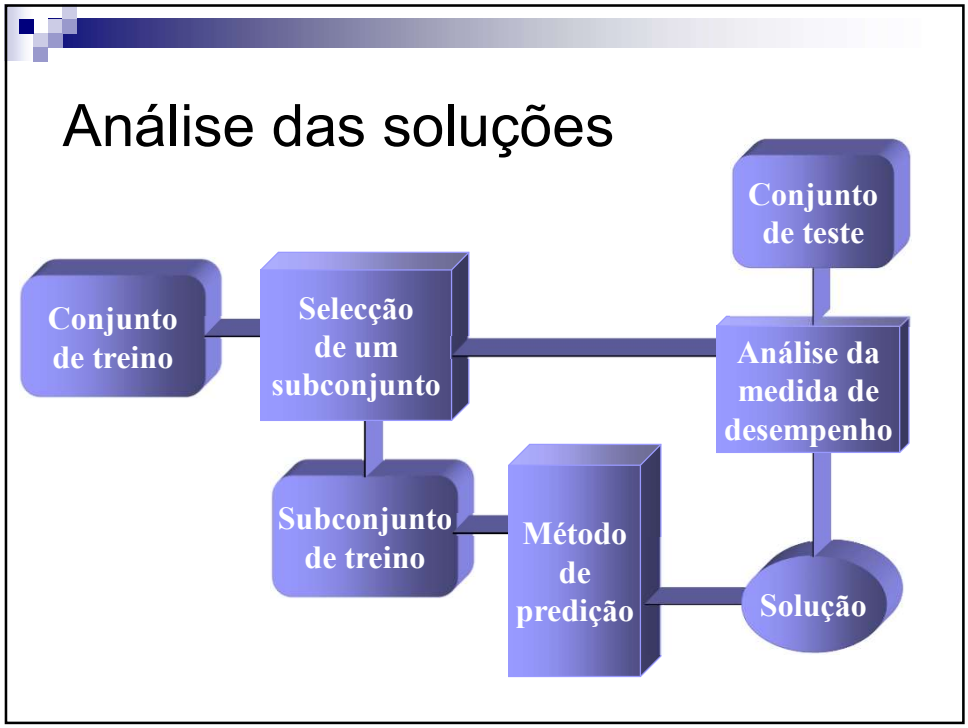
Redução dos dados



38


Modelação iterativa e predição
Mudança de parâmetros
Conjunto de treino
Método de predição
Solução
Testa o melhor
Conjunto de validação

39


Análise das soluções
Conjunto de teste
Conjunto de treino
Selecção de um subconjunto
Análise da medida de desempenho
Subconjunto de treino
Método de predição
Solução

40

Os principais paradigmas

- Redes Neuronais
- Baseados em instâncias
- Algoritmos genéticos
- Indução de regras
- Aprendizagem analítica

41

Alguns pontos para meditar(1)

- Que modelos são mais adequados para um caso específico?
- Que algoritmos de treino são mais adequados para um caso específico?
- Quantos exemplos são necessários? Qual a confiança que podemos ter na medida de desempenho?
- Como pode o conhecimento *a priori* ajudar o processo de indução?

42

Alguns pontos para meditar(2)

- Qual a melhor estratégia para escolher os exemplos ? Em que medida a estratégia altera o processo de aprendizagem?
- Quais as funções objectivo que se devem escolher para aprender? Poderá esta escolha ser automatizada?
- Como pode o sistema alterar automaticamente a sua representação para melhorar a capacidade de representar e aprender a função objectivo?

43

Exemplos de problemas

44

Exemplos (1)

- Um banco quer estudar as características dos seus clientes. Para isso precisa de encontrar grupos de clientes para os caracterizar.
- Quais as variáveis do problema? Como descrever os diferentes clientes.
- Que problema de aprendizagem se está a tratar?

45

Exemplo (2)

- Uma empresa de ramo automóvel resolveu desenvolver um sistema automático de condução de automóveis.
- Quais as variáveis do problema? Como descrever os diferentes ambientes.
- Que problema de aprendizagem se está a tratar?

46

Exemplo (3)

- Quer estudar-se a relação entre o custo das casas e os bairros de Lisboa.
- Quais as variáveis do problema? Como descrever os diferentes bairros.
- É um problema problema de predição, mas será de classificação ou de regressão?

47

Exemplo (4)

- Uma empresa de seguros do ramo automóvel quer detectar as fraudes das declarações de acidentes.
- Quais as variáveis do problema? Como descrever os clientes e os acidentes?
- É um problema problema de predição, mas será de classificação ou de regressão?

48