Sistemas de Apoio à Decisão

Data Mining & Optimização

Victor Lobo

Objectivos gerais

- Abrir horizontes em temas actuais
- Aprender técnicas usadas em “Sistemas de apoio à decisão” ou “Business Intelligence”
- Métodos de DataMining
  - Pesquisa de informação em grandes bases de dados
  - Aprender com experiência passada
- Métodos de Optimização
  - Resolver problemas de pesquisa “complicados”

Programa (parte relativa a técnicas)

1. **Introdução a Data Mining**
2. **Redes Neuronais – Perceptrão multicamada (MLP)**
3. **Redes Neuronais – Mapas auto-organizados (SOM)**
4. **Árvores de decisão**
5. **Introdução às técnicas de optimização**
6. **Algoritmos Genéticos**

Bibliografia

- Data Mining Techniques, for sales and customer support
  - Berry, M., Linoff, G., John Wiley and Sons, 1997
- Principles of Data Mining
  - Hand, D., Mannila,H,,Smyth,P.; MIT Press, 2001
- Machine Learning
  - Mitchell, Tom,", McGraw-Hill, 1997
- Haykin, Bishop, Hertz, Breiman, Salvador, …

Software

- SAS - Enterprise Miner
- SPSS - Clementine
- IBM - Intelligent Miner
- "open source em Java" - WEKA
- SAP – Módulos de Business Intelligence
- Matlab – Toolboxs de NN, DT, GA, etc
- Outros – "Statistica Neural Networks", SOM_PAK, C4.5(original), SNNS, plug ins para Excel, etc, etc, etc, etc.

Nosso patrocinador ! Disponível nas salas

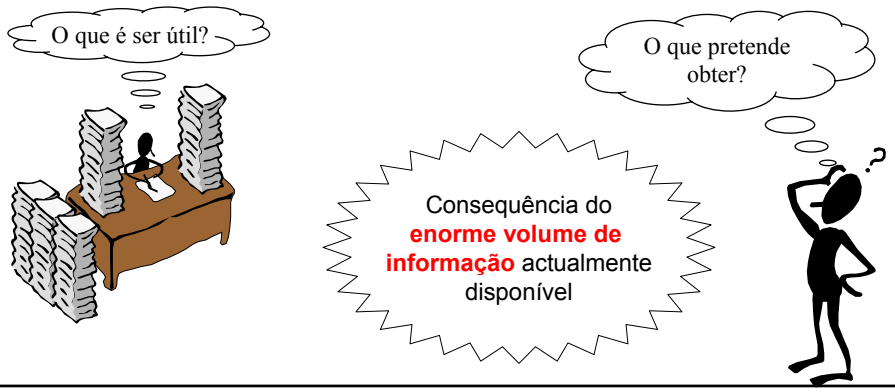
Alguns sites interessantes…

- Machine Learning Network
  - www.mlnet.org
  - Software, dados, conferências, projectos, etc.
- Repositório de Irvine
  - www.ics.uci.edu/~mlearn
  - Dados, software, artigos
- Homepage do WEKA
  - www.mkp.com/datamining
- SOM (H.U.T.)
  - www.cis.hut.fi/research/som-research/
  - Software, bibliografia sobre SOM

Introdução ao Data Mining

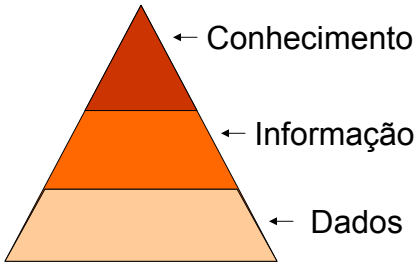
O que é "Data Mining"?

- "Data Mining" é a pesquisa de informação útil em grande quantidades de dados



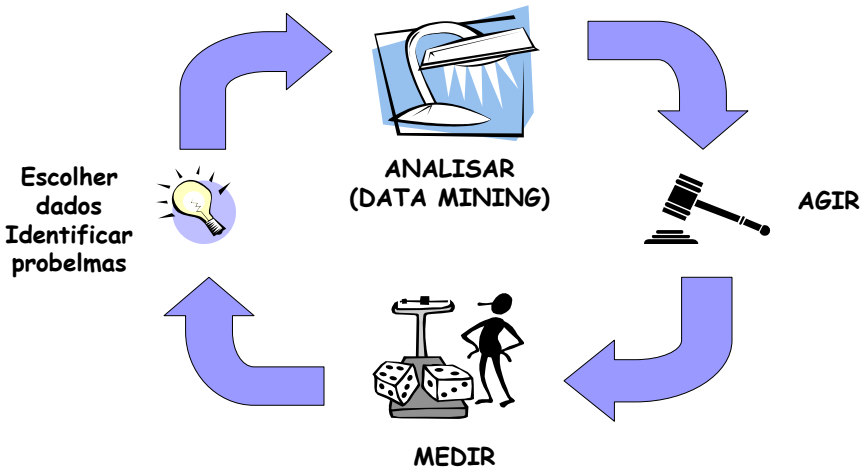
Informação é poder...

- "Água é vida"...
  - Todos os anos morre gente afogada...
- É necessário "trabalhar" a informação
- Hierarquia de compreensão e utilidade



E o que fazer depois de ter os dados organizados ?

O ciclo de data mining



Simplificando, Data Mining é

- A utilização de três técnicas diferentes:
  - Bases de dados
  - Estatística
  - **Aprendizagem máquina**.
    (Machine Learning)
- Para resolver dois tipos de problemas
  - Predição
  - Descobrir novo conhecimento

Predição e novo conhecimento

- Predição
  - é aprender critérios de decisão para ser capaz de classificar casos desconhecidos
- Descobrir novo conhecimento
  - é encontrar padrões desconhecidos existentes nos dados

Tipos de problemas

- **Predição**
  - **Classificação**
  - Regressão
- **Descoberta de conhecimento**
  - Detecção de desvios
  - Segmentação de bases de dados
  - **Clustering**
  - Regras de associação
  - Sumarização
  - Visualização
  - Pesquisa em texto

Exemplos

- Detecção de fraudes na utilização de um cartão de crédito
- Deferir, ou não, um pedido de crédito
- Prever perdas com seguros

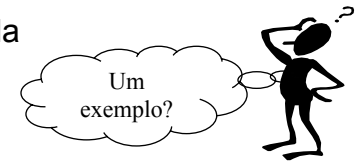
- Prever os níveis de audiência dos canais de televisão
- Classificar os efeitos hidrofónicos produzidos por diferentes navios
- Analisar as respostas de um inquérito médico
- Escolher clientes a quem direccionar uma campanha de marketing
- Cross-selling, fidelização, etc, etc,

Problemas “a montante”...

- Recolha de dados
- Representação dos dados
- Armazenagem, organização, e disponibilização dos dados
- Pré-processamento dos dados

Representação dos dados

- Representação mais usada = tabela
  - (Existem muitas outras...)
- Exemplo
  - Empresa de seguros de saúde



Variável, característica, ou atributo

Dado, vector, registo ou padrão

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

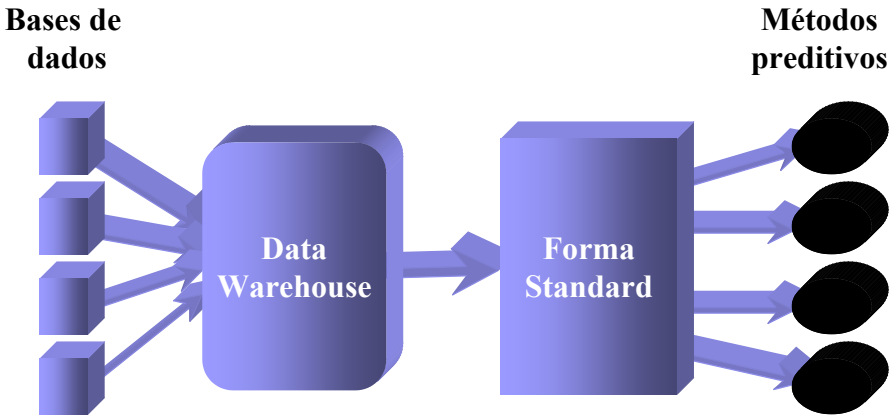
Tipos de atributos

- Booleanos ou binários
  - Só tomam dois valores
- Nominais
  - Tomam um conjunto de valores não ordenados
- Ordinais
  - Tomam um conjunto (finito) de valores ordenados
- Numéricos

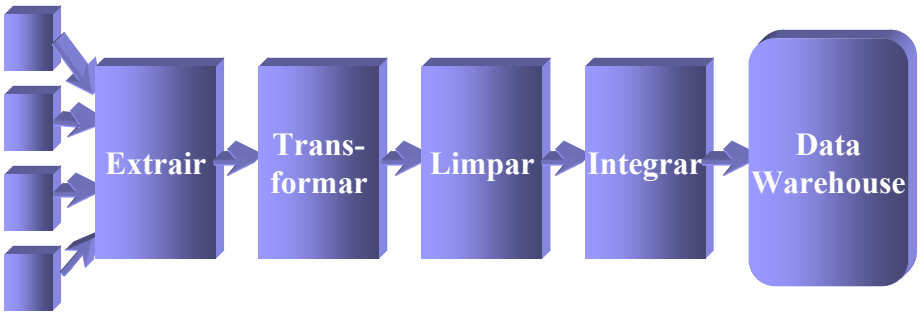
Como organizar os dados?

- "Data warehouse"
  - É o suporte centralizado de informação importante para a decisão.

O modelo de "data warehouse"



Passos para construir a "data warehouse"



Bases de dados

Pré-processamento dos dados

- "Tratar" dos missing values
  - Eliminá-los, substituí-los, etc
- Corrigir factores de escala entre atributos
  - Normalização linear por min/max
  - Normalizar média e desvio padrão
  - Outras
- Transformações de variáveis...
- *Vidé* "Data preparation for Data Mining", Dorian Pyle, Morgan Kaufmann, 1999

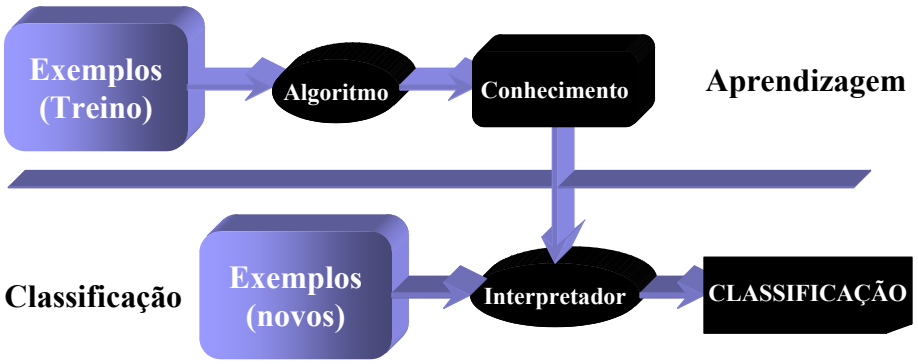
Alguns problemas importantes que NÃO vamos tratar...

- Escolha dos atributos
- Visualização
  - Dados multidimensionais
  - Problema central em datamining
- OLAP e outras técnicas de "reporting"
  - On-line Analytical Processing
- Regras de Associação e "Market Basket Analysis"

Introdução à aprendizagem

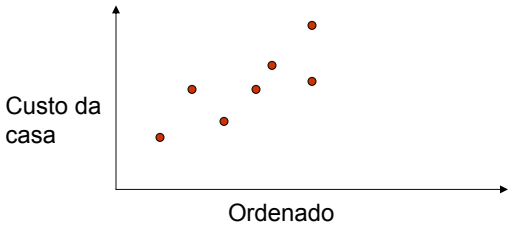
Aprender a partir dos dados conhecidos

Fases do processo

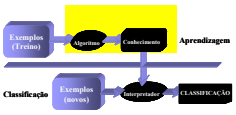


Exemplo de aprendizagem (1)

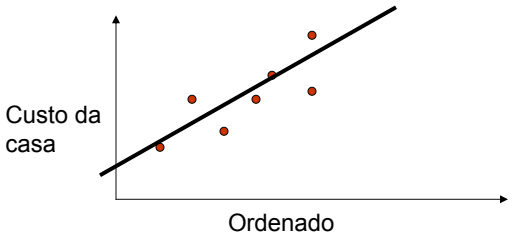
- Agência imobiliária pretende estimar qual a gama de preços para cada clinente
- Exemplos de treino:
  - Dados históricos
  - Ordenado vs custos de casas compradas



Exemplo de aprendizagem (2)

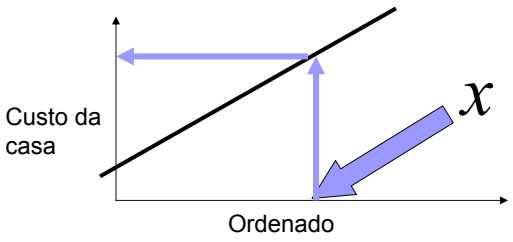


- Algoritmo
  - Regressão linear
- Representação do conhecimento
  - Recta (declive e ordenada na origem)



Exemplo de aprendizagem (3)

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



Outro problema de predição

- Exemplo da seguradora
- Existem 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

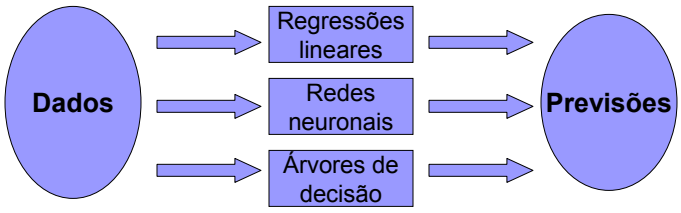
E o Manel ?

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

Terá encargos para a seguradora ?

Tipos de sistemas de previsão

- "Clássicos"
  - Regressões lineares, logísticas, etc...
- Redes Neuronais
- Árvores de decisão



Tipos de Aprendizagem

SUPERVISIONADA vs NÃO SUPERVISIONADA
INCREMENTAL vs BATCH
PROBLEMAS

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.

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

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.

Acesso aos exemplos

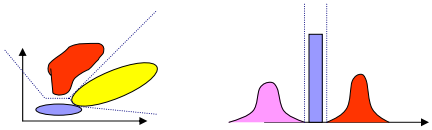
- Aprendizagem “offline”
  - Todos os exemplos estão disponíveis ao mesmo tempo
- Aprendizagem “online”
  - Os exemplos são apresentados um de cada vez
- Aprendizagem mista
  - Uma mistura dos dois casos anteriores

Problema do nº de atributos

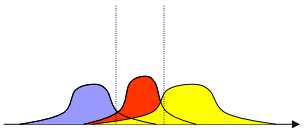
- Poucos atributos
  - Não conseguimos distinguir classes
- Muitos atributos
  - Caso mais vulgar em Datamining
  - Praga da dimensionalidade
  - Visualização difícil e efeitos “estranhos”
- Atributos importantes vs redundantes
  - Quais os atributos importantes para a tarefa?

Problema da separabilidade

- Separáveis
  - Erro Ø possível



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



Problema do "melhor" tipo de modelo

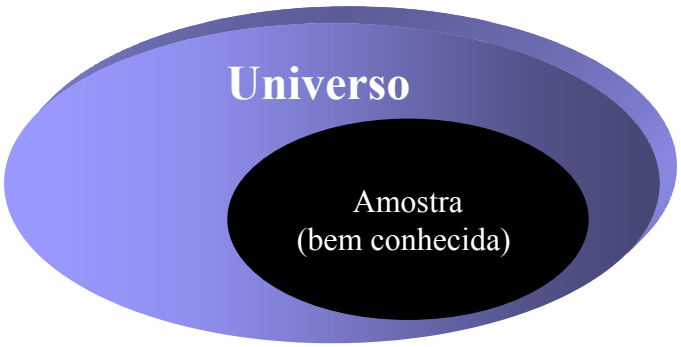
- A representação de conhecimento mais simples.
  - Mais fácil de entender
  - Árvores de decisão vs redes neuronais
- A representação de conhecimento com menor probabilidade de erro.
- A representação de conhecimento mais provável
  - Navalha de Occam ...

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...

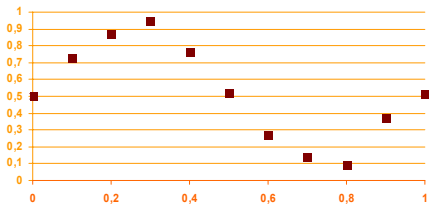
Generalização e “overfitting”

Os dados



Exemplo de overfitting

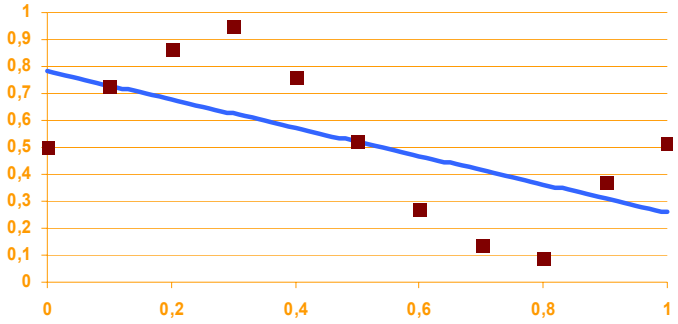
- Seja um conjunto de 11 pontos.
- Encontrar um polinómio de grau M que represente esses 11 pontos.



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

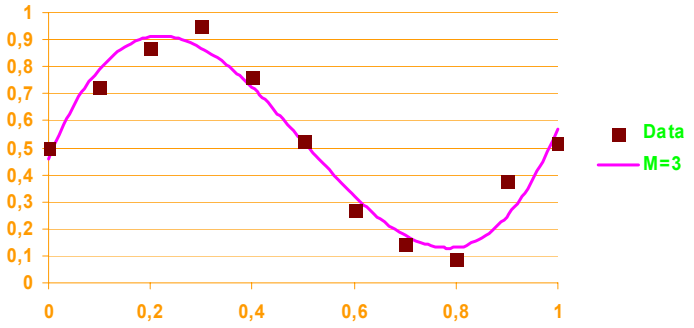
Aproximação M = 1

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



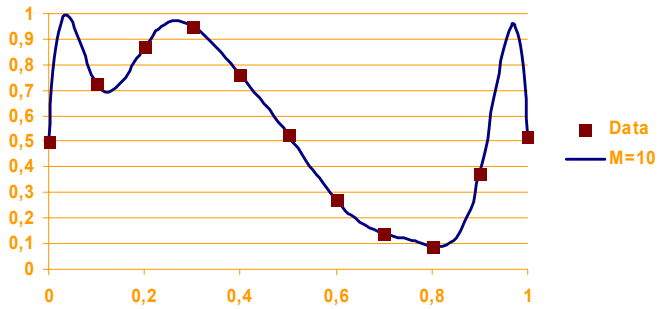
Aproximação M = 3

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

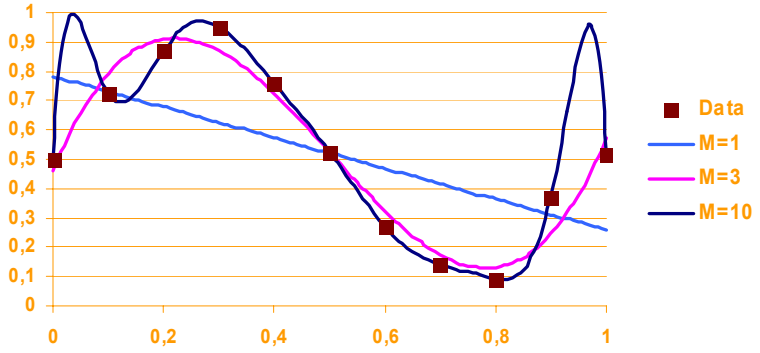


Aprocimação M = 10

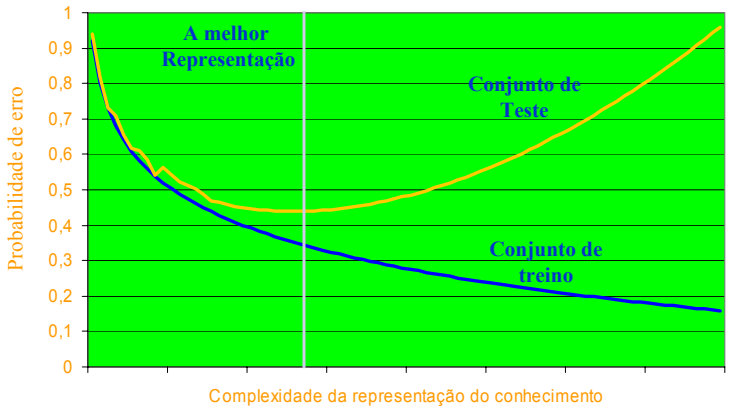
$$y(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8 + w_9 x^9 + w_{10} x^{10}$$



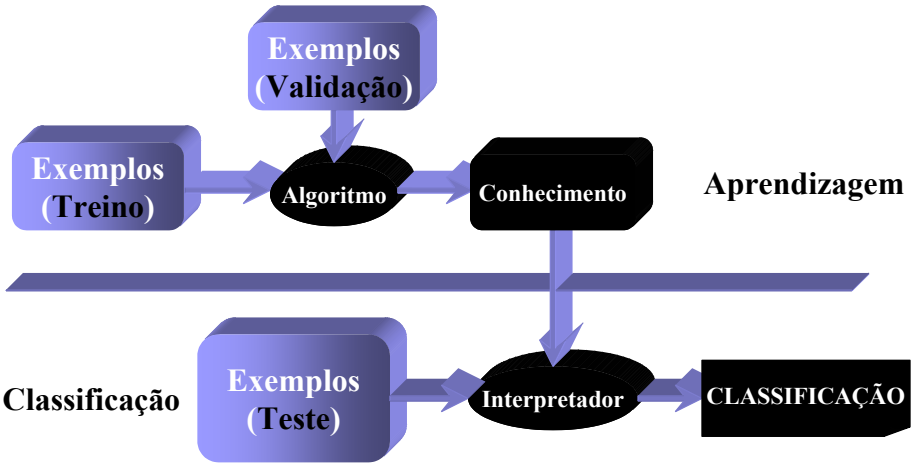
Overfitting



Curva de Overfiting



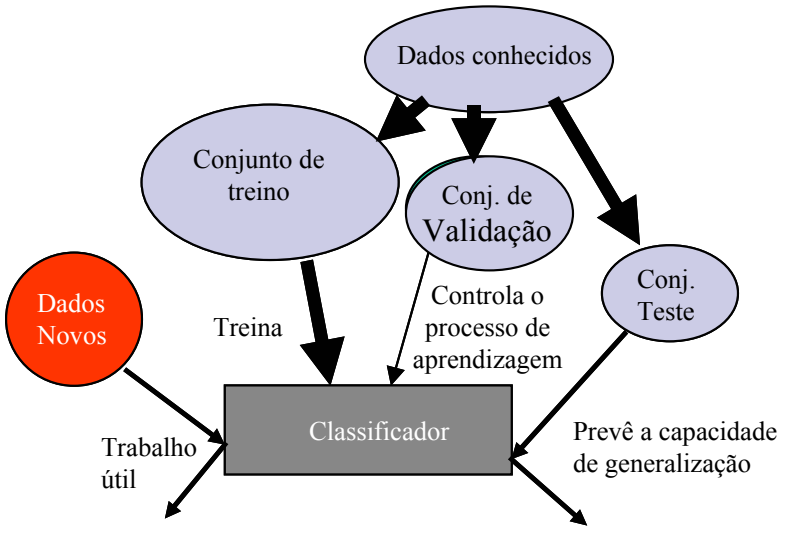
Fases do processo



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"

Conjunto de treino/validação/teste



Divisão dos dados

- Conjunto de treino
  - Quanto maior, melhor o classificador obtido
- Conjunto de validação
  - Quanto maior, melhor a estimação do treino óptimo
- Conjunto de teste
  - Quanto maior, melhor a estimação do desempenho do classificador

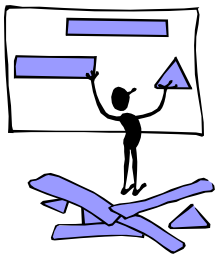
Processo de aprendizagem

- 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

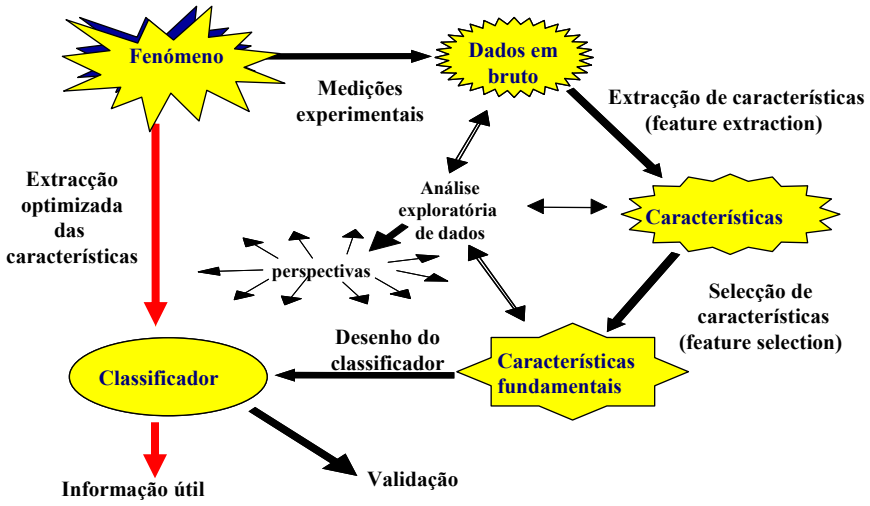
Projecto do sistema de aprendizagem

Tarefas do projecto

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

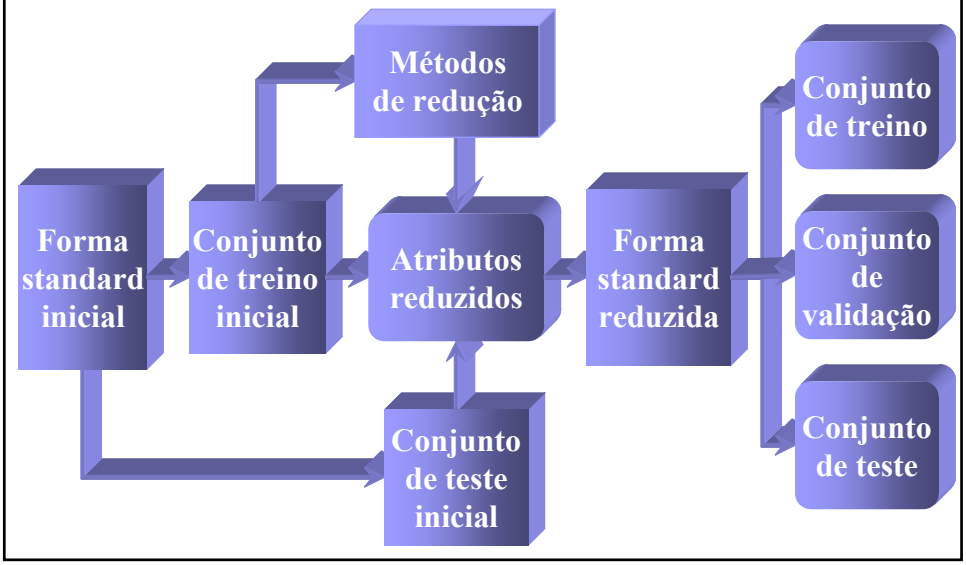


Aproximação exploratória...

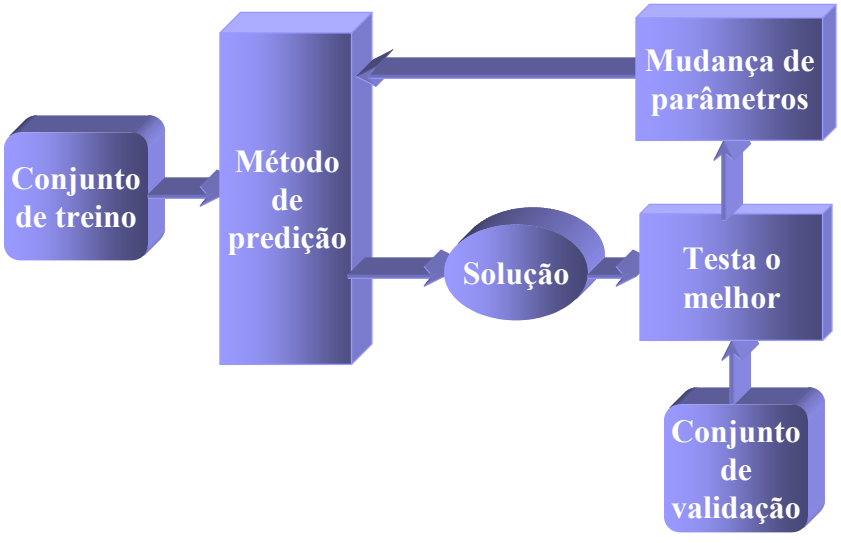


Preparação dos dados

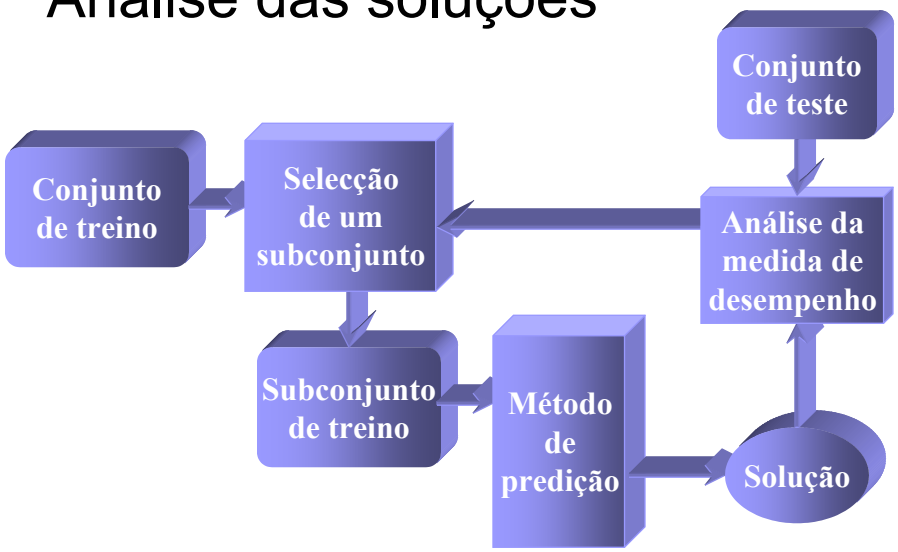
Redução dos dados



Modelação iterativa e predição



Análise das soluções



Considerações finais

Os principais paradigmas

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

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?

Alguns pontos para meditar(2)

- Qual a melhor estratégia para escolher o processo exemplo? 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?

Exemplos de problemas

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?

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?

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?

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?