Introdução a Datamining (previsão e agrupamento)

Victor Lobo

Mestrado em Estatística e Gestão de Informação

E o que fazer depois de ter os dados organizados ?

Ideias base

Aprender com o passado

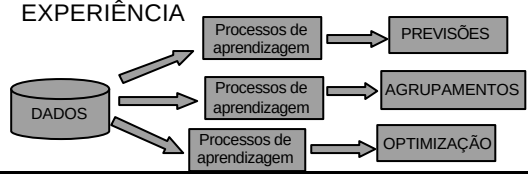
Inferir a partir da experiência

Ferramentas: técnicas de datamining

by any other name...

Datamining (lato senso)

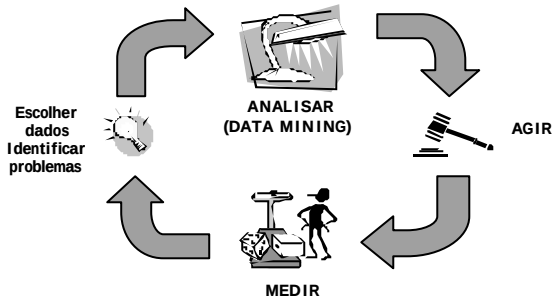
- Componente cognitiva das organizações
- Objectivo
  - Extrair conhecimento da experiência adquirida
  - Prever acontecimentos, identificar situações, optimizar processos, A PARTIR DA EXPERIÊNCIA



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)

O ciclo de datamining



Simplificando, Datamining é

- A utilização de três técnicas diferentes:
  - Bases de dados
  - Estatística
  - **Aprendizagem máquina**. (Machine Learning)
- Para resolver principalmente 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

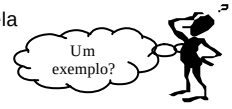
- Detecção de fraudes na utilização de um cartão de crédito
- Deferir, ou não, um pedido de crédito
- Prever perdas com seguros
- Prever os níveis de audiência dos canais de televisão
- Classificar os efeitos hidrofónicos produzidos por diferentes navios
- Analisar as respostas de um inquérito médico
- 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 *usual* dos dados

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



Variável, característica, ou atributo

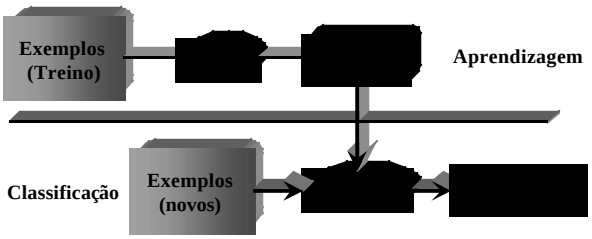
Dado, vector, registo ou padrão

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

Introdução à aprendizagem

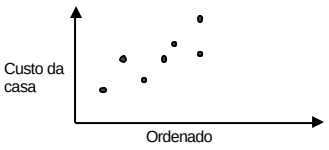
Aprender a partir dos dados conhecidos

Fases do processo

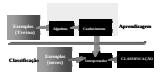


Exemplo de aprendizagem (1)

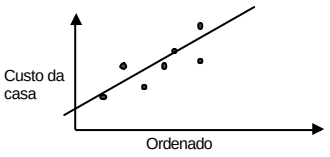
- Agência imobiliária pretende estimar qual a gama de preços para cada clinente
- Exemplos de treino:
  - Dados históricos
  - Ordenado vs custos de casas compradas



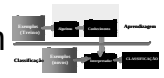
Exemplo de aprendizagem (2)



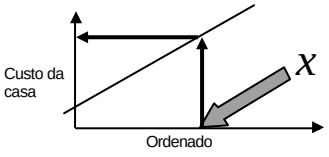
- Algoritmo
  - Regressão linear
- Representação do conhecimento
  - Recta (declive e ordenada na origem)



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 (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

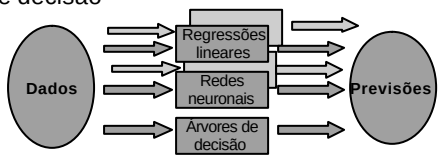
E o Manel ?

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

Terá encargos para a seguradora ?

Tipos de sistemas de previsão

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



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

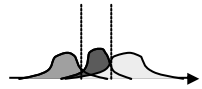
Problema do nº de atributos

- 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

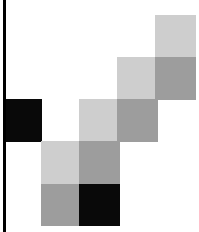


Problema do "melhor" tipo de modelo

- 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 ...

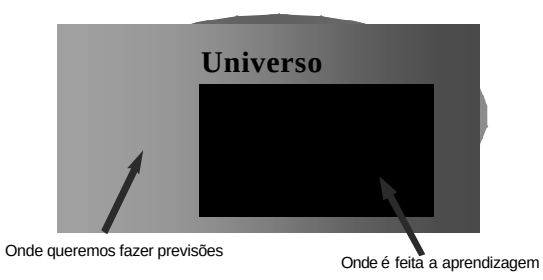
Problemas ...

- 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"

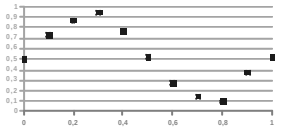
Os dados



Exemplo de overfitting

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

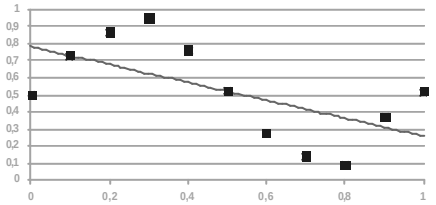
$$y(x) = \sum_{i=0}^{M} w_i x^i$$



Aproximação M = 1

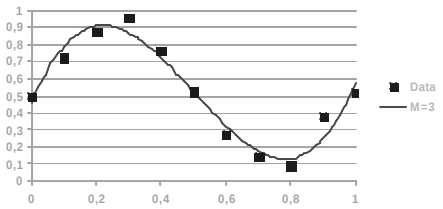
$$y(x) = w_0 + w_1 x$$

Erro grande



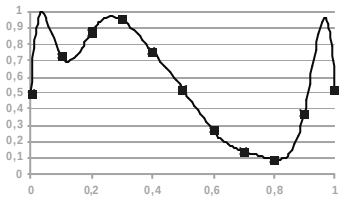
Aproximação M = 3

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



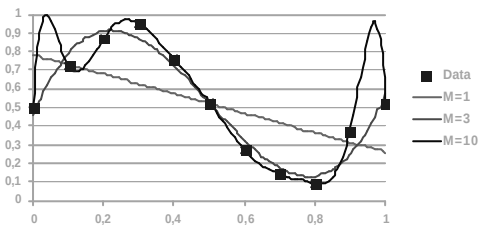
Aproximação M = 10

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

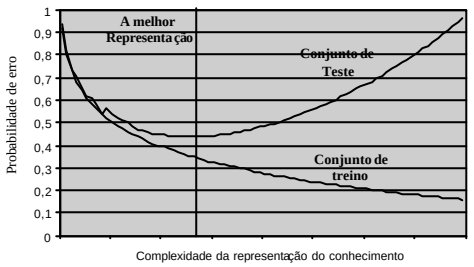


Erro zero

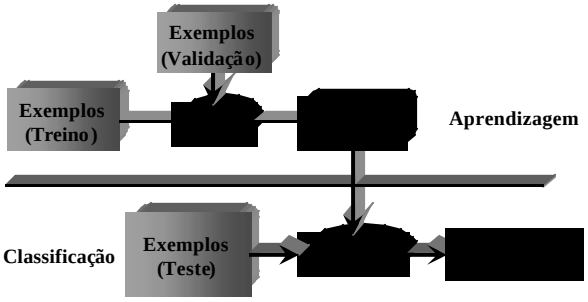
Overfitting



Curva de Overfiting



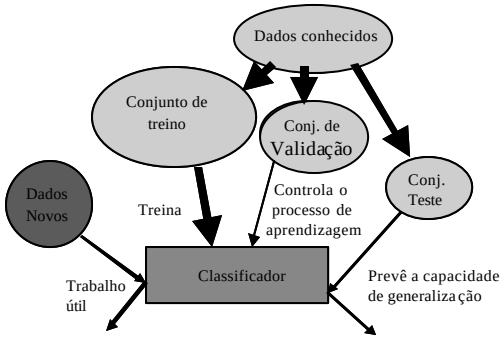
Fases do processo



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
  - 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

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…)

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 !!!

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

Processo de aprendizagem

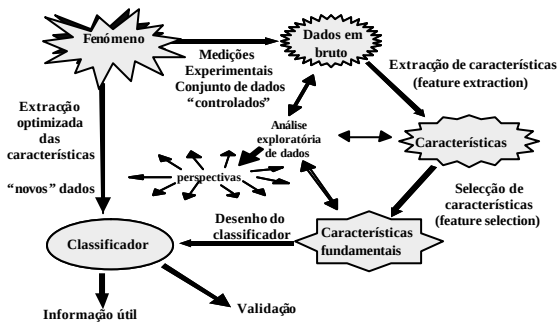
- A aprendizagem é um processo de optimização (Minimização do erro)
- Algoritmo de optimização
  - Método do gradiente
  - Subir a encosta
  - Guloso
  - Algoritmos genéticos
  - "Simulated annealing"
- Formas de adquirir o conhecimento

Iterações sucessivas do sistema de aprendizagem

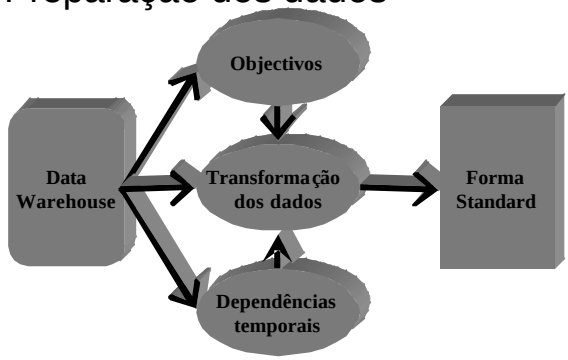
Tarefas do projecto do sistema

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

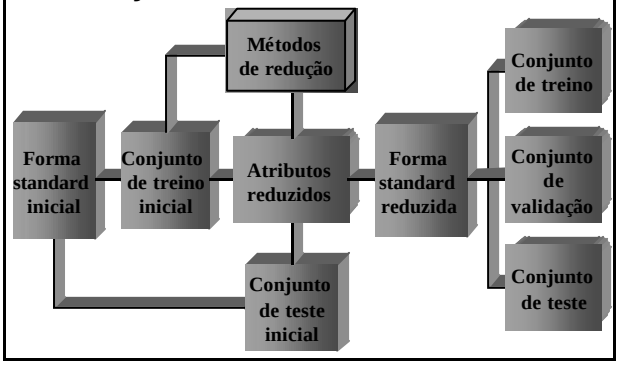
Aproximação exploratória...



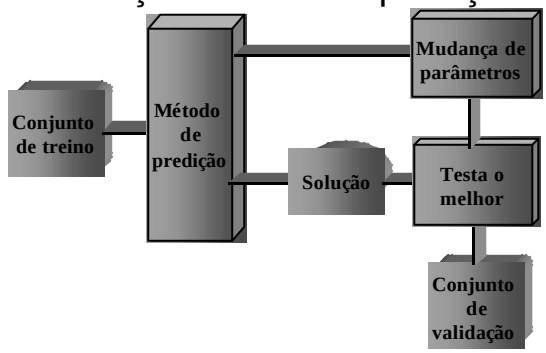
Preparação dos dados



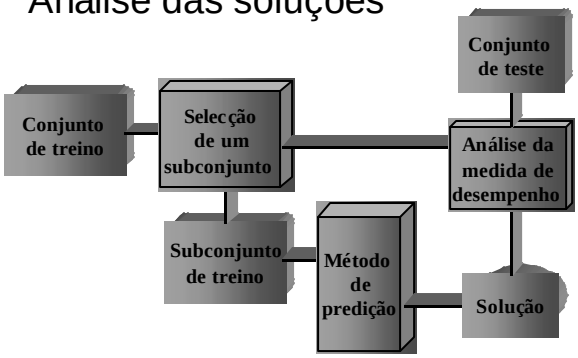
Redução dos dados



Modelação iterativa e predição



Análise das soluções



Os principais paradigmas

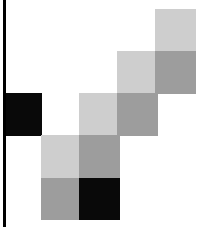
- Redes Neuronais
- Baseados em instâncias
- Algoritmos genéticos
- Indução de regras
- Aprendizagem analítica

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 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?



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?