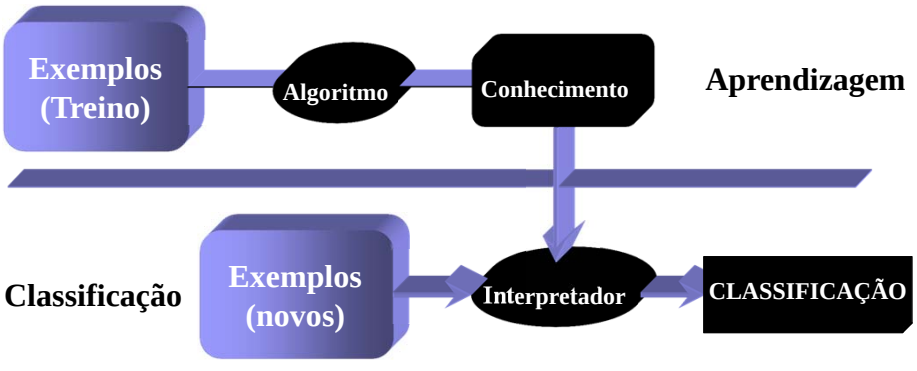
Introdução ao Datamining

V 1.4, V.Lobo, EN/ISEGI, 2012

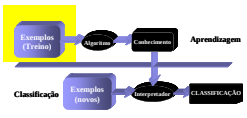
Introdução à aprendizagem

Aprender a partir dos dados conhecidos

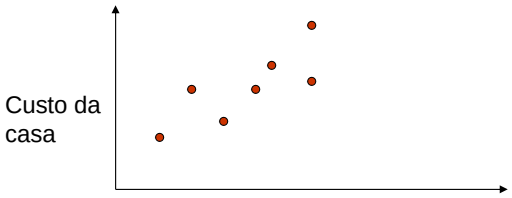
Fases do processo



Exemplo de aprendizagem (1)

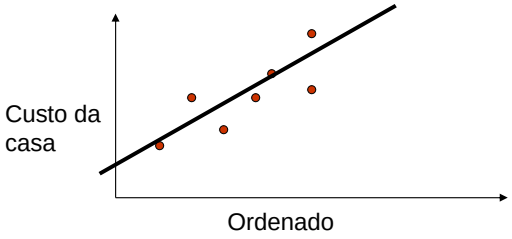


- Agência imobiliária pretende estimar qual a gama de preços para cada clinente
- Exemplos de treino:
  - Dados históricos
  - Ordenado vs custos de casas compradas

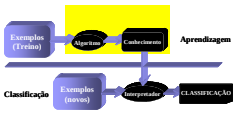


Exemplo de aprendizagem (2)

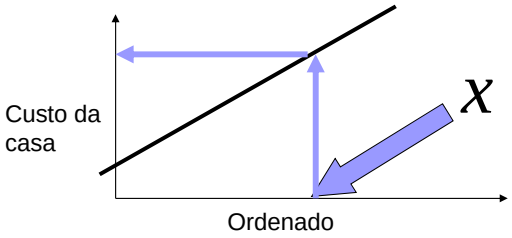
- Algoritmo
  - Regressão linear
- Representação do conhecimento
  - Recta (declive e ordenada na origem)



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

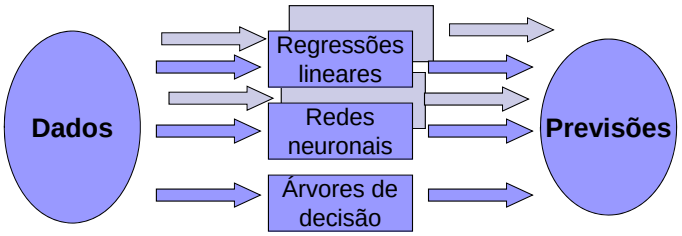
E o Manel ?

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

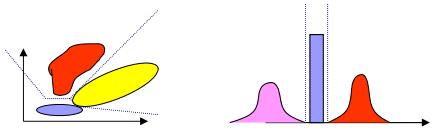
- Aprendizagem “offline”
  - Todos os exemplos estão disponíveis ao mesmo tempo
- Aprendizagem “online”
  - Os exemplos são apresentados um de cada vez
- Aprendizagem mista
  - Uma mistura dos dois casos anteriores

Problema do nº de atributos

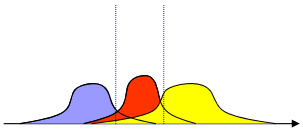
- Poucos atributos
  - Não conseguimos distinguir classes
- Muitos atributos
  - Caso mais vulgar em Datamining
  - Praga da dimensionalidade
  - Visualização difícil e efeitos “estranhos”
- Atributos importantes vs redundantes
  - Quais os atributos importantes para a tarefa?

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

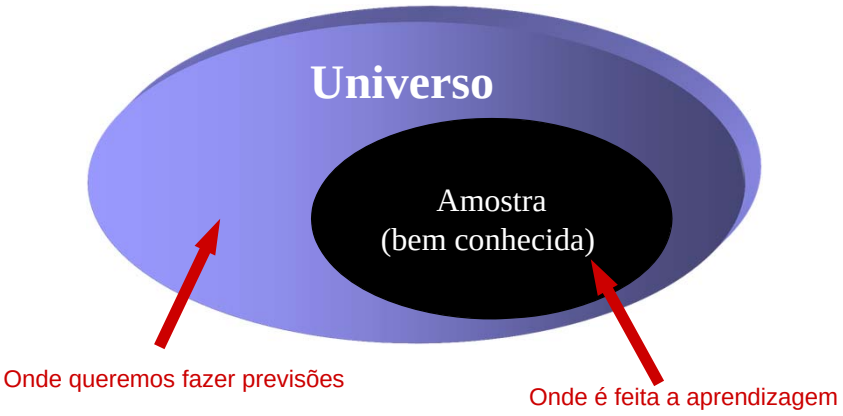
- A representação de conhecimento mais simples.
  - Mais fácil de entender
  - Árvores de decisão vs redes neuronais
- A representação de conhecimento com menor probabilidade de erro.
- A representação de conhecimento mais provável
  - Navalha de Occam ...

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"

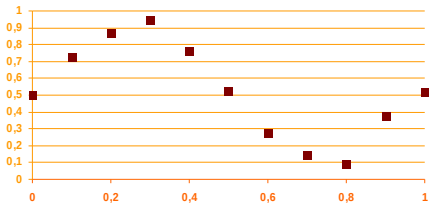
Os dados



Exemplo de overfitting

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

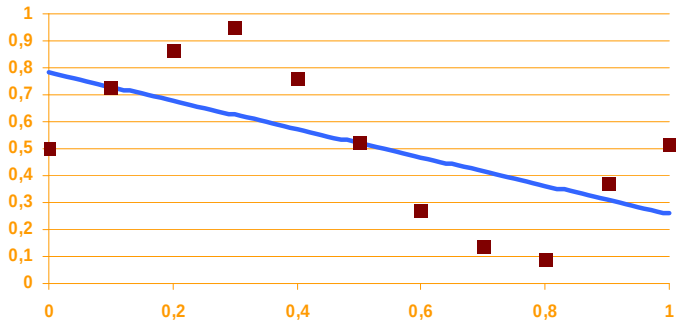
$$y(x) = \sum_{i=0}^{M} w_i x^i$$



Aprovimação M = 1

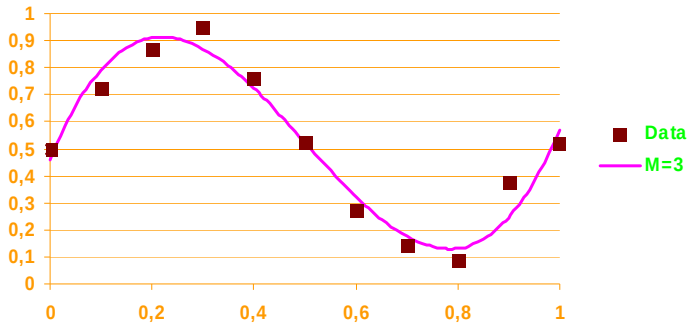
$$y(x) = w_0 + w_1 x$$

Erro grande



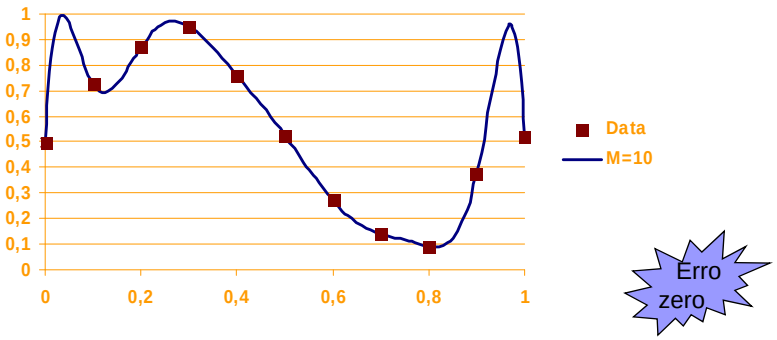
Aproximação M = 3

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

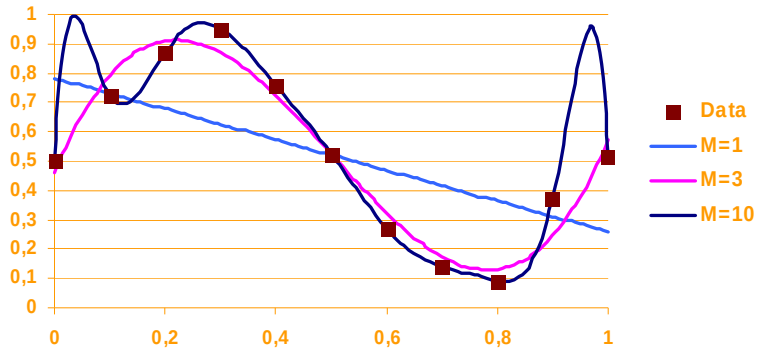


Aproximação M = 10

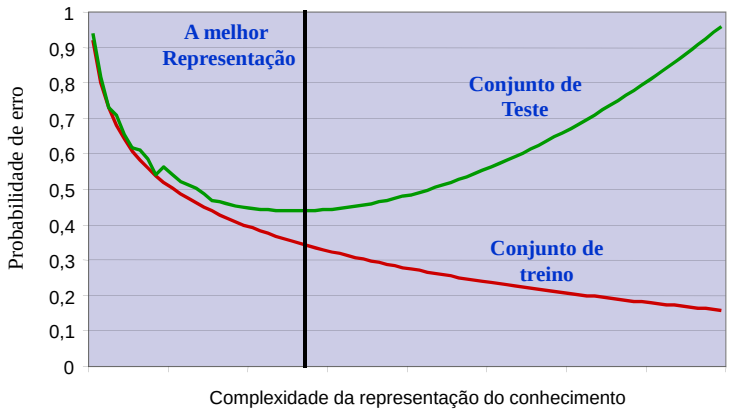
$$y(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8 + w_9 x^9 + w_{10} x^{10}$$



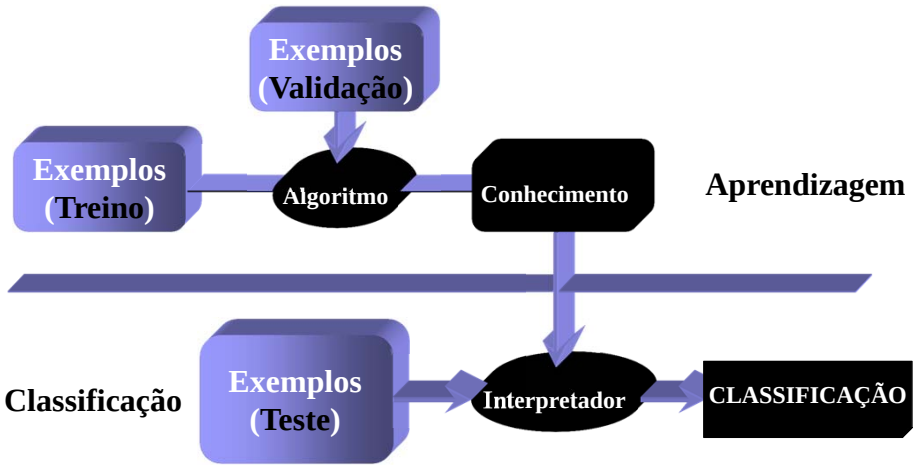
Overfitting



Curva de Overfiting



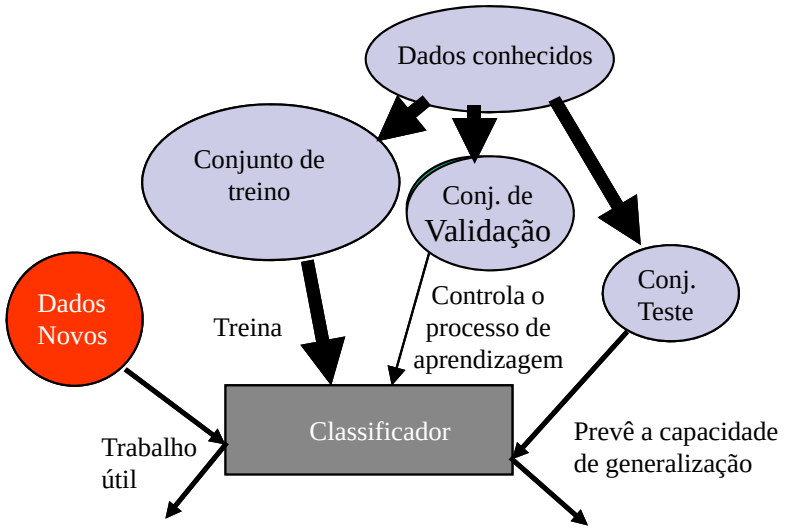
Fases do processo



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

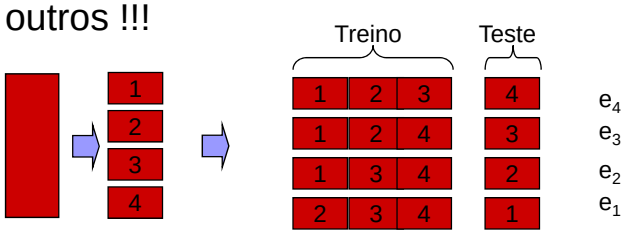
- Conjunto de treino
  - Usado para construir o classificador
  - Quanto maior, melhor o classificador obtido
- Conjunto de validação
  - Usado para controlar a aprendizagem (opcional)
  - Quanto maior, melhor a estimação do treino óptimo
- Conjunto de teste
  - Usado para estimar o desempenho
  - Quanto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

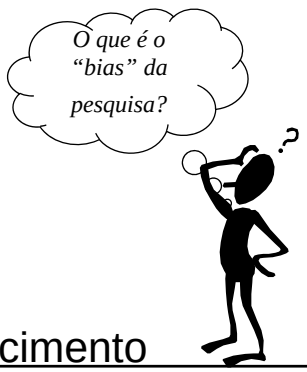
- **Matriz de confusão**
  - Separa os diversos tipos de erro
    - Falso Positivo (FP)
      - O classificador diz que é, e não é
    - Falso Negativo (FN)
      - O classificador não detecta que é
  - Permite compreender em que é que o classificador é bom

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 (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 !

Processo de aprendizagem

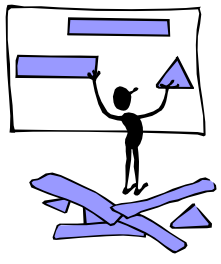
- A aprendizagem é um processo de optimização (Minimização do erro)
- Algoritmo de optimização
  - Método do gradiente
  - Subir a encosta
  - Guloso
  - Algoritmos genéticos
  - "Simulated annealing"
- Formas de adquirir o conhecimento



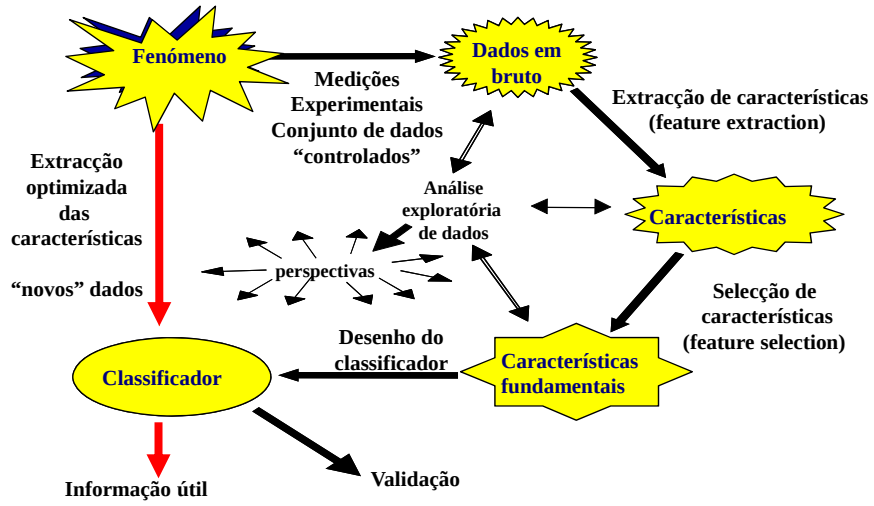
Iterações sucessivas do sistema de aprendizagem

Tarefas do projecto do sistema

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

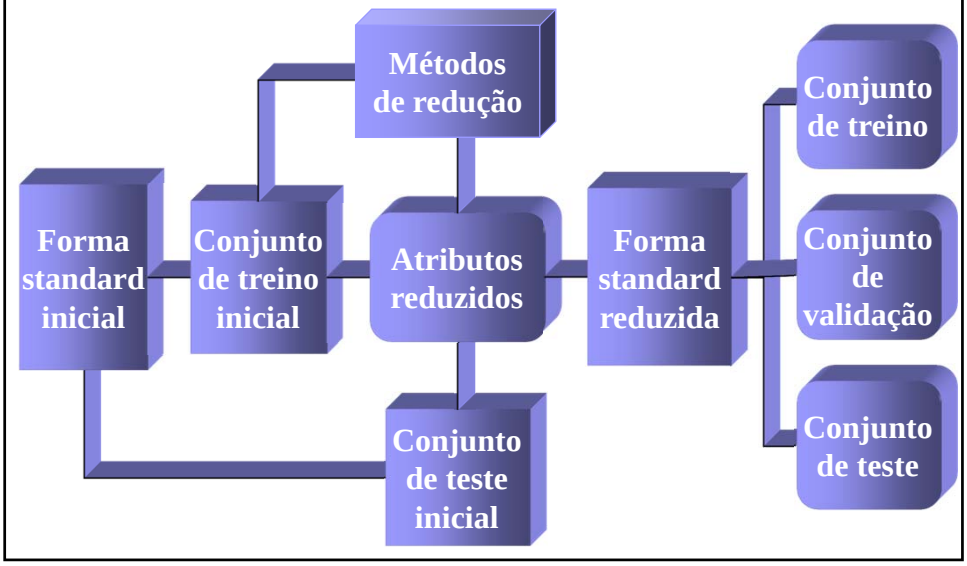


Aproximação exploratória...

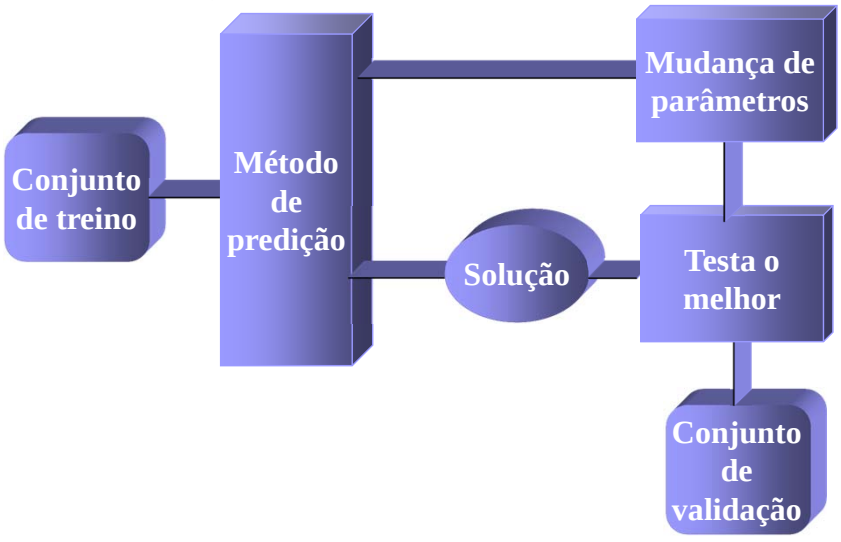


Preparação dos dados

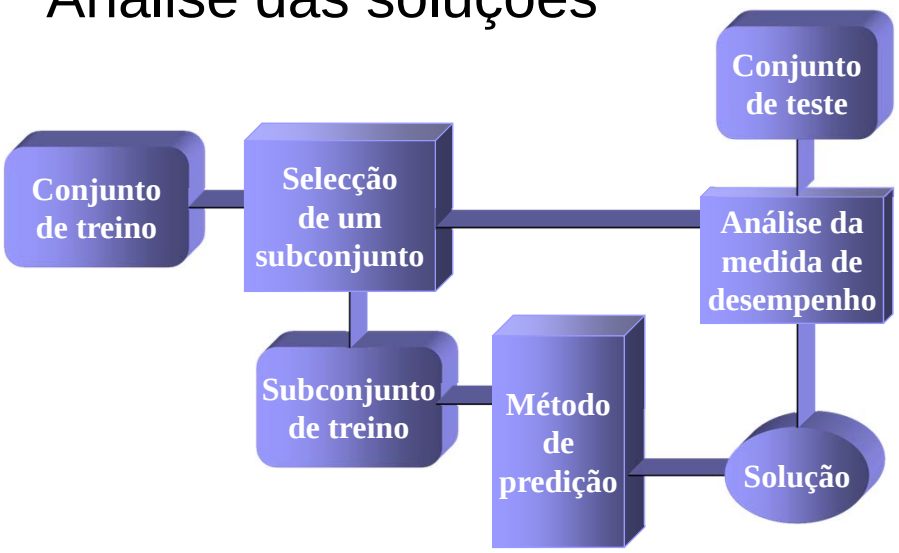
Redução dos dados



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?

Pré-Processamento dos dados

Porquê pré-processar os dados

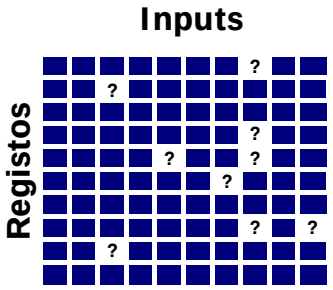
- Valores omissos (missing values)
- Factores de escala
- Invariância a factores irrelevantes
- Eliminar dados contraditórios
- Eliminar dados redundantes
- Discretizar ou tornar contínuo
- Introduzir conhecimento “à priori”
- Reduzir a “praga da dimensionalidade”
- Facilitar o processamento posterior

Valores omissos

- Usar técnicas que lidem bem com eles
- Substitui-los
  - Por valores “neutros”
  - Por valores “médios” (média, mediana, moda, etc)
  - Por valores “do vizinho mais próximo”
    - K-vizinhos, parzen, etc
  - Interpolações
    - Lineares, com “splines”, com Fourier, etc.
  - Com um estimador “inteligente”
    - Usar os restantes dados para fazer a previsão

Alternativa: Eliminar valores omissos

- Eliminar registos
  - Podemos ficar com poucos dados
  - (neste caso 3 em 10)
- Eliminar variáveis
  - Podemos ficar com poucas características
  - (neste caso 4 em 9)



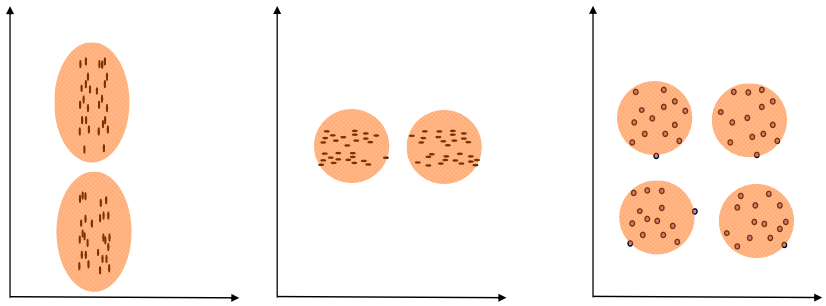
Abordagem iterativa

- Usar primeiro uma aproximação “grosseira”
  - Eliminar registos / variáveis
  - Usar simplesmente valores médios
- Observar os resultados
  - Conseguem-se boas previsões ?
  - Resultados são realistas ?
- Abordagem mais fina
  - Estimar valores para os omissos
  - Usar “clusters” para definir médias

Normalização dos dados

Nomalização

- Efeitos de mudanças de escala



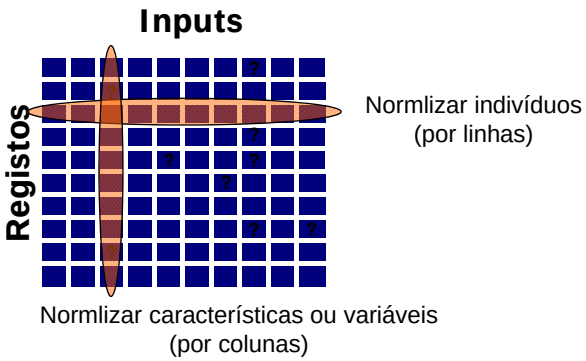
O que é perto do quê ?

Porquê normalizar

- Para cada variável individual
  - Para não comparar “alhos com bugalhos” !
- Entre variáveis
  - Para que métodos que dependem de distâncias (logo de escala) não fiquem “trancados” numa única característica
  - Para que as diferentes características tenham importâncias proporcionais.

Porquê normalizar

- Entre indivíduos
  - Para insensibilizar a factores de escala
  - Para identificar “prefis” em vez de valores absolutos



Objectivos possíveis

- Aproximar a distribuição de uniforme
  - “Espalha” maximamente os dados
- Aproximar a distribuição normal
  - Identifica bem os extremos e deixa que estes sejam muito diferentes
- Ter maior resolução na “zona de interesse”

Pré-processamento

- Algumas normalizações mais comuns
  - Min-Max
    - y'∈[0,1]

$$y' = \left(\frac{y - \min}{\max - \min} \right)$$

  - Z-score
    - y' centrado em 0 com σ=1

$$y' = \frac{y - média}{DesvioPadrão}$$

  - Percentis
    - Distribuição final sigmoidal

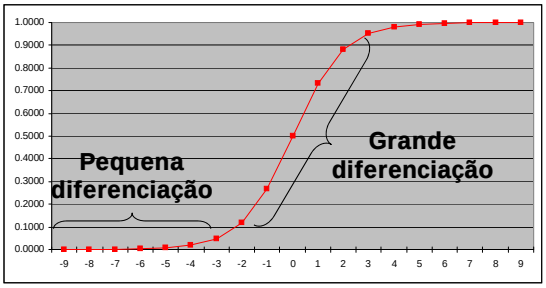
$$y' = nº\ de\ ordem$$

  - Sigmoidal (logística)
    - y' com maior resoução "no centro"

$$y' = \frac{1 - e^{-y}}{1 + e^{-y}}$$

Normalização sigmoidal

- Diferencia a "zona de transição"



Transformações dos dados

Projecções sobre espaços visualizáveis ou de dimensão menor

- Ideia geral:
  - Mapear os dados para um espaço de 1 ou 2 dimensões
- Mapear para espaços de 1 dimensão
  - Permite definir uma ordenação
- Mapear para espaços de 2 dimensões
  - Permite visualizar a “distribuição” dos dados (semelhanças, diferenças, clusters)

Problemas com as projecções

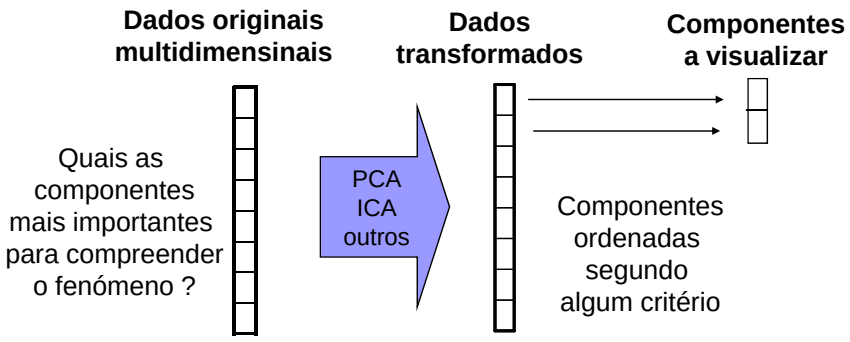
- Perdem informação
  - Podem perder MUITA informação e dar uma imagem errada
- Medidas para saber "o que não estamos a ver"
  - Variância explicada
  - Stress
  - Outros erros (erro de quantização, topológico,etc)

Dimensão *intrínseca*

- Dimensão do sub-espaço dos dados
  - Pode ou não haver um mapeamento linear
- Estimativas da dimensão intrínseca
  - Com PCA – Verificar a diminuição dos V.P.
    - Basicamente, medir a variância explicada
  - Com medidas de stress (em MDS)
  - Com medidas de erro

Seleccionar componentes mais "relevantes" para visualização

- Será sempre uma "boa" escolha ?

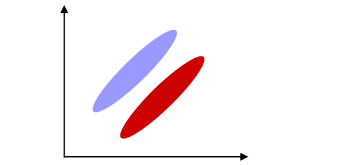


PCA – Principal Component Analysis

- Principal Component Analysis
  - Análise de componente principais
  - Transformada (discreta) de Karhunen-Loève
  - Transformada linear para o espaço definido pelos vectores próprios da martriz de covariância dos dados.
    - Não é mais que uma mudança de coordenadas (eixos)
    - Eixos ordenados pelos valores próprios
    - Utiliza-se normalmente SVD

Componentes principais

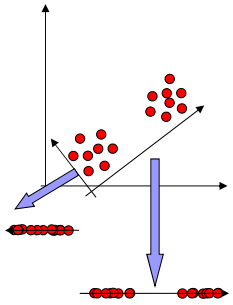
- Mudança de eixos
  - Os novos eixos estão “alinhados” com as direcções de maior de variação
  - Continuam a ser eixos perpendiculares
  - Podem “esconder aspectos importantes”



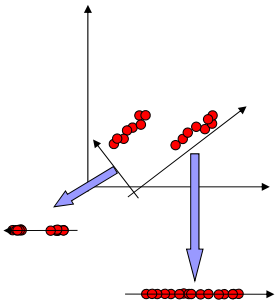
A 2ª componente é que separa !

A dimensão intínseca é 1 !

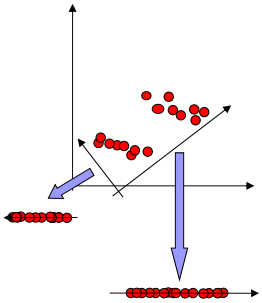
Problemas com ACP



Corre bem !



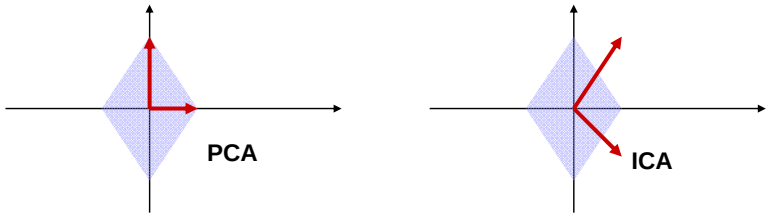
Menos bem !



Mal !

Componentes Independentes

- ICA – Indepenant Component Analysys
  - Maximizam a independência estatística (minimizam a informação mútua)
- Diferenças em relação a PCA



Componentes Independentes

- Bom comportamento para clustering
  - Muitas vezes melhor que PCA por "espalhar" melhor os dados
- Bom para "blind source separation"
  - Separar causas independentes que se manifestam no mesmo fenómeno
- Disponibilidade
  - Técnica recente… ainda pouco divulgadada
  - Boas implementações em Matlab e C
  - Livro de referencia (embora não a ref.original):
    - Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis, Wiley-Interscience.

Referências sobre ICA

- Primeiras referências
  - B.Ans, J.Herault, C.Jutten, “Adaptative Neural architectures: Detection of primitives”, COGNITIVA’85, Paris, France, 1985
  - P.Comon, “Independant Component Analysis, a new concept ?”, Signal Processing, vol36,n3,pp278-283, July 1994
- Algoritmo mais usado. FastICA
  - Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis, Wiley-Interscience.
  - V.Zarzoso, P.Comon, “How Fast is FastICA?”, Proc.European Signal Processing Conf., Florence, Italy, Setember 2006
- Recensão recente
  - A.Kachenoura et al.,”ICA: A Potential Tool for BCI Systems”, IEEE Signal processing Magazine, vol25, n.1, pp 57-68, January 2008
- Código freeware e material de apoio
  - FastICA para Matlab, R, C++, Python, e muitos apontadores para informação
  - http://www.cis.hut.fi/projects/ica/fastica/

MDS – MultiDimensional Scaling

- Objectivo
  - Representação gráfica a 2D que preserva as distâncias originais entre objectos
- Vários algoritmos (e por vezes nomes diferentes)
  - Sammon Mapping (1968)
  - Também conhecido como Perceptual Mapping
  - É um processo iterativo
  - Não é, rigorosamente, um mapeamento…
- Stress
  - Mede a distorção que não foi possível eliminar

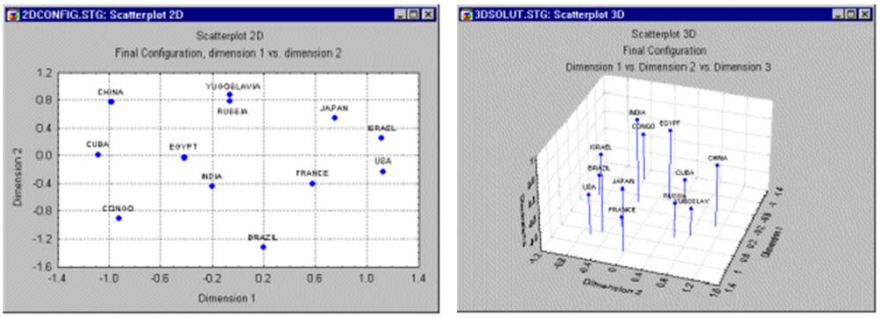
$$Stress = \sqrt{\frac{(d_{ij} - \hat{d}_{ij})^2}{(d_{ij} - \overline{d})^2}}$$

d_{ij} = *distância verdadeira*

$\hat{d}$ = *distância no grafico* $2\,d$

$\overline{d}$ = *média das distâncias*

Exemplos de MDS



Exemplo com países do mundo caracterizados por indicadores socio-económicos

- Nota:
  - Ao acrescentar mais um dado é necessário recalcular tudo !

Transformações tempo/frequência

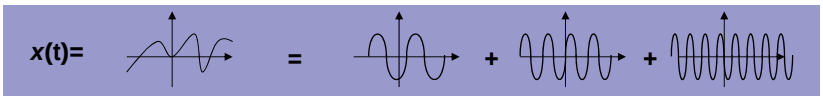
- Transformada de Fourier
  - É uma mudança de referencial !
  - Projecta um espaço sobre outro
- Transformadas tempo/frequência
  - Wavelets
  - Wigner-Ville
  - Identificam a ocorrência (localizada no tempo) de fenómenos que se vêm melhor na frequência…

Transformada de Fourier

- Aplicações
  - Análise de séries temporais
  - Análise de imagens
  - Análise de dados com dependências "periódicas" entre eles
- Permite:
  - Invariância a "tempo"
  - Invariância a "posição"
- O que é:
  - Um decomposição em senos e cosenos
  - Uma projecção do espaço original sobre um espaço de funções

Transformada de Fourier

- O que é a "decomposição" ?



- Com o que é que fico ? Com o que quiser…
  - Com as amplitudes de cada frequência…
  - Com os valores das 2 frequências mais "fortes"…
- Notas:
  - Para não perder informação *N*-pontos geram *N*-pontos
  - Posso calcular a transformada mesmo que faltem valores

Curvas principais, SOM, etc

- Curvas principais
  - Hastie 1989
  - Define-se parametricamente a família de curvas sobre o qual os dados são projectados
- SOM
  - Kohonen 1982
  - Serão discutidas mais tarde

Bibliografia

- Sammon, J. W., Jr (1969). "A Nonlinear Mapping for Data Structure Analysis." IEEE Transactions on Computers **C-18**(5)
- Hastie, T. and W. Stuetzle (1989). "Principal curves." Journal of the American Statistical Association **84**(406): 502-516.
- Hyvarinen, A. and E. Oja (2000). "Independant component analysis: algorithms and applications." Neural Networks **13**: 411-430
- Hyvärinen, A., J. Karhunen, et al. (2001). Independent Component Analysis, Wiley-Interscience.

Outros problemas de pré-processamento

Eliminar outliers

- Efeito de alavanca dos outliers
- Efeito de "esmagamento" dos outliers
- Eliminar outliers
  - Estatística (baseado em σ)
  - Problema dos "inliers"
  - Métodos "detectores" de outliers
    - Com k-médias
    - Com SOM

Conversões entre tipos de dados

- Nominal / Binário
  - 1 bit para cada valor possível
- Ordinal / Numérico
  - Respeitar ou não a escala ?
- Numérico / Ordinal
  - Como discretizar ?

Outras transformações

- Médias para reduzir ruído
- Ratios para insensibilizar a escala
- Combinar dados
  - É introdução de conhecimento “à priori”

Quanto pré-processamento ?

- Mais pré-processamento
  - Maior incorporação de conhecimento à priori
  - Mais trabalho inicial, tarefas mais fáceis e fiáveis mais tarde
- Menos pré-processamento
  - Maior esforço mais tarde
  - Maior "pressão" sobre sistema de classificação/ previsão / clustering
  - Princípio: "garbage in – garbage out"

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?