Data Mining II – Modelos Preditivos

V 1.1, V.Lobo, EN/ISEGI, 2013

Data Mining II
Modelos Preditivos

Prof. Doutor Victor Lobo

Mestre André Melo

Mestrado em Estatística e Gestão de Informação

Objectivo desta disciplina

- Fazer previsões a partir de dados.
- Conhecer os principais problemas relacionados com previsões baseadas em dados (“data driven”)
- Conhecer as principais técnicas:
  - Métodos clássicos: regressões, interpolações, extrapol.
  - Decisões Bayesianas
  - Sistemas baseados em instâncias
  - Árvores de decisão
  - Redes neuronais
  - Ensambles

Data Mining II – Modelos Preditivos

V 1.1, V.Lobo, EN/ISEGI, 2013

Programa (tarços gerais)

- **1. Introdução aos métodos de previsão em Data Mining**
- **2. Dados, pré-processamento, e estimativas de erro**
- **3.Teoria da decisão e sistemas Bayesianos**
- **4. Aprendizagem e classificação baseada em instâncias**
- **5. Árvores de decisão**
- **6. Redes Neuronais**
- **7. Ensambles**

Programa (detalhado) 2/5

- **1. Introdução às técnicas preditivas em DM**
- **2. Dados, pré-processamento, e estimativas de erro**
  - 2.1 **Tipos de dados**
  - 2.2 **Métricas** para dados numéricos e categóricos
  - 2.3 **Técnicas de normalização**
  - 2.4 O problema dos **valores em falta**
  - 2.3 **Estimativas de erro** de sistemas de classificação e regressão
  - 2.5 Técnicas para **extracção de características e projecções**

Programa (detalhado) 3/5

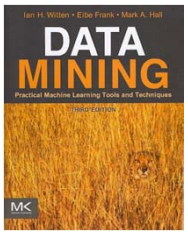
- **3. Teoria da decisão e sistemas Bayesianos**
  - 3.1 Conceitos gerais
  - 3.2 **Decisões óptimas Bayesianas**.
- **4. Aprendizagem e classificação baseada em instâncias**
  - 4.1 Algoritmo do **vizinho mais próximo**
  - 4.2 Variantes do vizinho mais próximo
  - 4.3 Escolha selectiva

Programa (detalhado) 5/5

- **5. Indução de Árvores de Decisão**
  - 5.1 Princípios gerais
  - 5.2 Algoritmo DDT
  - 5.3 Outros algoritmos
- **6. Redes Neuronais**
  - 7.1 Perceptrões Multi-camada (**MLP**)
  - 7.2 Redes de **RBF**
  - 7.3 Redes de Hopfield
  - 7.4 Support Vector Machines (**SVM**)
  - 7.5 Outros tipos de redes
- **11. Ensambles**

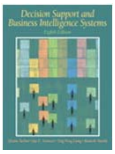
Bibliografia

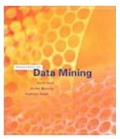
- Livros de texto (nenhum é seguido "à risca")
  - **Data mining: practical machine learning tools and techniques**; Ian H. Witten, Eibe Frank, Mark A. Hall: Morgan Kaufmann, 2011
  - **Machine Learning**, Tom M.Mitchell, McGraw Hill, 1997
  - **Pattern Classification**, Duda, Hart, & Stork, Wiley, 2001

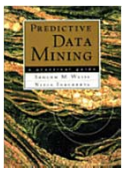


Outra Bibiliografia

- **Decision Support and Business Intelligence Systems**, Turban, E., J. E. Aronson, *et al.*, Prentice Hall, 2010
- **Principles of data mining**, David. J. Hand, Heikki Mannila, Padhric Smyth, MIT Press, 2001
- **Predictive data mining**, Sholom M. Weiss, Nitin Indurkhya, Morgan Kaufmann, 1997
- **C4.5:Programs for Machine Learning**, John Ross Quinlan, Morgan Kaufmann, 1992







Avaliação

- Prova Escrita de Exame (Final)
  - 60% da nota
- Trabalhos
  - Trabalho prático de grupo (20%)
    - A entregar na altura da prova escrita de Exame
    - 2 alunos por grupo
  - Trabalhos de Casa (20%)
    - 1º para entregar a 18/04/2013
    - 2º para entregar a 17/5/2013
    - Individual
  - NOTA MÍNIMA EM TODAS AS PROVAS – 9 valores

Horário de dúvidas e contactos

- Email: vlobo@isegi.unl.pt
  - Dúvidas
    - 4ª Feira às 19:00 (ou quando combinarmos)
    - Por mail em qualquer altura
    - Sempre que estiver no ISEGI (!)
  - Material de apoio
    - www.isegi.unl.pt/docentes/vlobo

Resolução de problemas práticos

- WEKA
- SAS Enterprise Miner
- Outros
  - MS-Excel
  - Matlab
  - Microsoft SQL server

Dias de aulas…

20-02-2013	4	Aula 1	
27-02-2013	4	Aula 2	
06-03-2013	4	Aula 3	
13-03-2013	4		
20-03-2013	4	Aula 4	
27-03-2013	4		
02-04-2013	3	Aula 5	
03-04-2013	4	Aula 6	
10-04-2013	4	Aula 7	
17-04-2013	4		
24-04-2013	4	Aula 8	
07-05-2013	3	Aula 9	
08-05-2013	4	Aula 10	
14-05-2013	3	Aula 11	
15-05-2013	4	Aula 12	
22-05-2013	4		
27-05-2013	2	Aula 13	

4ªfeira sem aulas

Aula prática

Aula à 3ª/2ª feira

Software (para esta cadeira e para DM2)

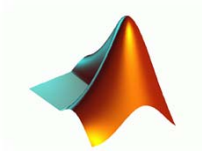
- Excel !
  - Resolve muitos problemas.
  - Teste de métodos para “poucos” dados
- SAS - Enterprise Miner
  - Escalável para problemas “a sério”
  - Grande variedade de ferramentas
  - Pouca informação detalhada sobre métodos
  - Bom interface visual mas programação “pouco amigável”
  - www.sas.com – Muita informação sobre aplicações

Nosso patrocinador ! Disponível nas salas

Software (pacotes facilmente disponíveis)

- WEKA
  - Para Datamining e “Machine Learning”
  - “open source” em Java
  - Corre em muitos ambientes, bastante completo (v3)
  - http://www.cs.waikato.ac.nz/ml/weka/

- Matlab (ou Octave e SciLab que são GNU)
  - Toolboxs de NN, DT, GA, ML, etc
    - SOMTOOLBOX (som), NETLAB (machine learning)
  - www.mathworks.com (site comercial da mathworks)
  - http://www.gnu.org/software/octave/
  - http://www.scilab.org/



- R
  - Package estatístico com muito suporte para datamining
  - Parecido com Matlab (mas diferente)
  - http://www.r-project.org/

- Outros – “Statistica Neural Networks”, SOM_PAK, C4.5(original), SNNS, plug-ins para Excel, etc, etc, etc, etc.

Software (pacotes comerciais genéricos)

- SPSS – Clementine
  - Muito difundido nalgumas universidades
  - Versão de educação brevemente disponível
  - www.spss.com
- IBM - Intelligent Miner
  - Tem uma versão para dowload gratuito
  - http://www-306.ibm.com/software/data/iminer/
- SAP - Módulos de Business Intelligence
  - Grande variedade de módulos
  - http://www.sap.com/platform/netweaver/components/bi/index.epx

Outros sites interessantes…

- DSS Resources
  - Prof. Daniel Power, livros, referências, etc
  - http://dssresources.com/
- Decisionarium
  - Software GNU, referências, etc
  - http://www.decisionarium.tkk.fi
- Machine Learning Network
  - www.mlnet.org
  - Software, dados, conferências, projectos, etc.
- Repositório de Irvine
  - www.ics.uci.edu/~mlearn
  - Dados, software, artigos
- Fabricantes de soluções “dedicadas”
  - Para gestão de terrenos, para marketing, etc, etc

Factores importantes:

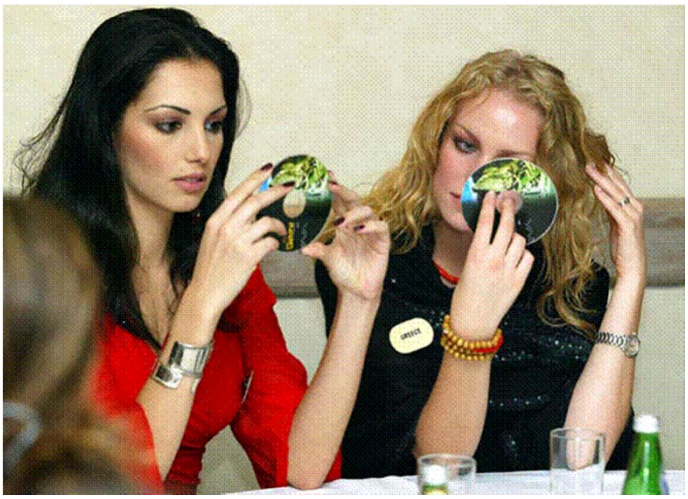
- Quantidade de dados disponíveis
  - Dados operacionais, sensores de baixo custo
- Poder de cálculo e armazenamento de dados
- Sistemas computacionais de baixo custo
- Desenvolvimento científico e tecnológico
  - Inteligência Artificial e Aprendizagem Máquina (Machine Learning), e convergência com as técnicas mais clássicas da área da estatística, da investigação operacional, e do reconhecimento de padrões
- Software de fácil utilização
  - SAS, SAP, etc.

Introdução a Datamining (previsão e agrupamento)

Victor Lobo

Mestrado em Estatística e Gestão de Informação

Recolher dados para quê ?



Exemplo de Janus

- Olhar o passado e o futuro
- "***Estudar*** o passado para ***compreender*** o presente, e ***prever*** o futuro"

Ideias base

Aprender com o passado

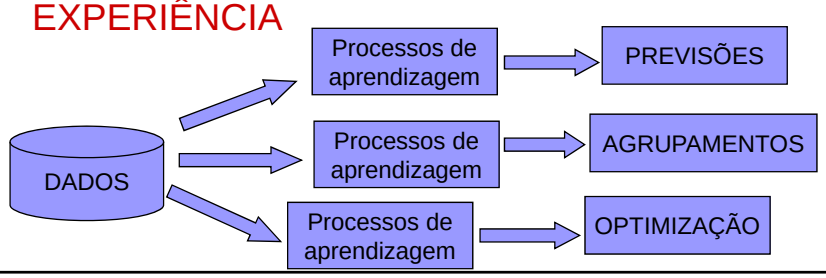
Inferir a partir da experiência

Ferramentas: técnicas de datamining

by any other name...

Datamining (lato senso)

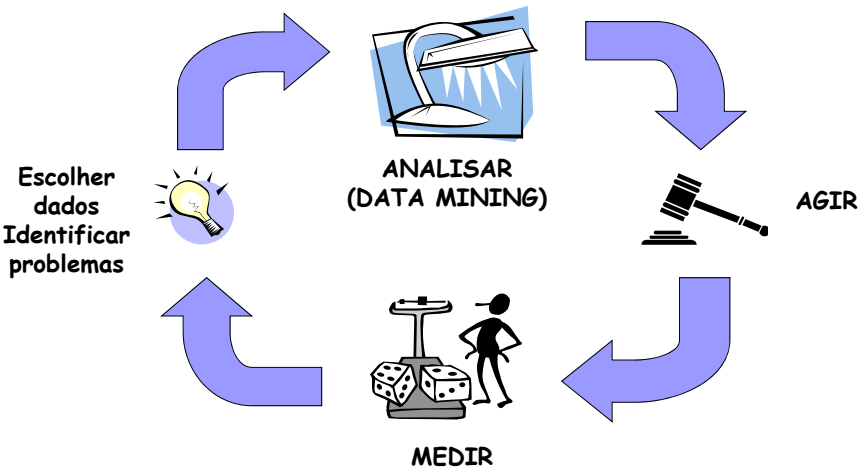
- Componente cognitiva das organizações
- Objectivo
  - Extrair conhecimento da experiência adquirida
  - Prever acontecimentos, identificar situações, optimizar processos, A PARTIR DA EXPERIÊNCIA



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

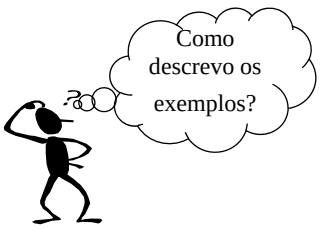
- Predição
  - é aprender critérios de decisão para ser capaz de classificar casos desconhecidos
- Descobrir novo conhecimento
  - é encontrar padrões desconhecidos existentes nos dados

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



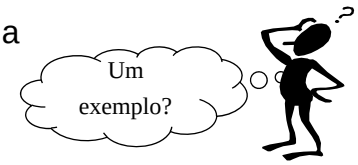
- Prever os níveis de audiência dos canais de televisão
- Classificar os efeitos hidrofónicos produzidos por diferentes navios
- Analisar as respostas de um inquérito médico
- Escolher clientes a quem direccionar uma campanha de marketing
- Cross-selling, fidelização, etc, etc,

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