



Caderno de Exercícios de Sistemas de Apoio à Decisão

**Cláudia Antunes
Helena Galhardas
Arlindo L. Oliveira**

Março 2005

ÍNDICE

ÍNDICE	2
CAPÍTULO 1 - MODELO MULTIDIMENSIONAL E OLAP	4
Modelo Relacional – Formas Normais e Dependências Funcionais	4
Datawarehouse.....	6
CAPÍTULO 2 - ÁRVORES DE DECISÃO	10
Conceito de Entropia.....	10
Conceito de Ganho de Informação.....	10
Algoritmo ID3	10
CAPÍTULO 3 - REDES NEURONAIS.....	17
Perceptrão	17
Redes Multi Camada.....	18
CAPÍTULO 4 - APRENDIZAGEM BAYESIANA	22
Noções Básicas de Probabilidades.....	22
Classificador de Naive Bayes	23
Redes de Bayes.....	25
Princípio das Descrições Mais Compactas (MDL)	28
CAPÍTULO 5 - SUPPORT VECTOR MACHINES (SVMS)	30
CAPÍTULO 6 - APRENDIZAGEM BASEADA EM INSTÂNCIAS	33
K Nearest Neighbors (KNN).....	33
Algoritmo de EM	34
CAPÍTULO 7 - AVALIAÇÃO DE HIPÓTESES	35

Cross-Validation	35
CAPÍTULO 8 - ALGORITMOS GENÉTICOS.....	37
CAPÍTULO 9 - REGRAS DE ASSOCIAÇÃO.....	39
Algoritmo Apriori.....	39

Capítulo 1 - MODELO MULTIDIMENSIONAL E OLAP

MODELO RELACIONAL – FORMAS NORMAIS E DEPENDÊNCIAS FUNCIONAIS

Dependência funcional: Seja uma relação R com conjuntos de atributos X e Y, diz-se que X determina Y ($X \rightarrow Y$) se para todo o valor de X existe um só valor de Y. Por outras palavras, sempre que dois tuplos têm um valor de X comum, têm também o valor de Y em comum.

1ª Forma Normal: Uma relação está na 1ª Forma normal se todos os seus atributos são valores escalares.

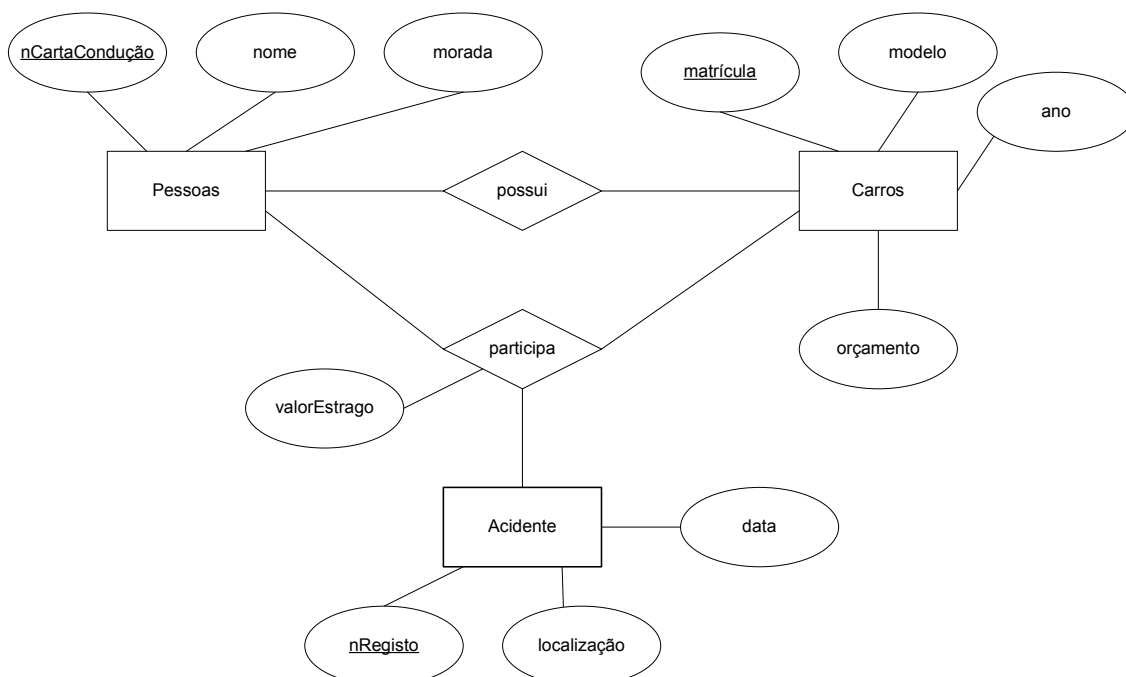
2ª Forma Normal: Uma relação está na 2ª Forma normal se está na 1ª Forma Normal e cada atributo não chave é completamente dependente dos atributos chave.

3ª Forma Normal: Uma Relação está na 3ª Forma normal se todos os atributos não chave são completamente dependentes dos atributos chave e são independentes entre si.

BCNF: Uma Relação está na BCNF quando todos os determinantes são chave. A BCNF só se distingue da 3FN quando existe mais do que uma chave e as chaves são formadas por vários atributos.

Exercício 1

Obtenha o esquema relacional simplificado (que inclui apenas os nomes das relações e nomes dos atributos) para o seguinte esquema E/R.



Exercício 2

Considere o seguinte esquema relacional:

Empregado (nomeEmp, rua, cidade)

Trabalha(nomeEmp, nomeOrg, salario)

Organização(nomeOrg, cidade)

Gere(nomeEmp, nomeGestor)

Expresse, em álgebra relacional e em SQL, as seguintes interrogações:

- a) Quais os nomes dos empregados que trabalham no “Banco N°1”?
- b) Quais os nomes dos empregados que vivem na mesma cidade que a organização em que trabalham?
- c) Quais os nomes dos empregados que não trabalham para o “Banco N°1”?
- d) Quais os nomes dos empregados que ganham mais do que qualquer outro empregado do “Banco N°1”?
- e) Quais as organizações cujos empregados ganham, em média, uma salário superior à média dos salários do Banco N°1? (Expresse apenas em SQL)

Exercício 3

Considere as seguintes entidades/relações R1, R2, R3 e R4, respectivos atributos e as dependências funcionais correspondentes:

R1(A, B, C, D)

B → D

AB → C

R2(A, B, C, D, E)

AB → CE

E → AB

C → D

R3(A, B, C, D)

B → D

D → B

ABD → C

R4(A, B, C, D)

A $\rightarrow$ BCD

B $\rightarrow$ ACD

CD $\rightarrow$ AB

Para cada entidade/relação, diga qual o identificador e em que FN está?

Exercício 4 (Formas Normais)

Considere a relação R(a,b,c,d,e), em que {a,b} é a chave primária e $d \rightarrow e$.

- Indique todas as dependências da relação R.
- Qual a forma normal da relação R?
- Assuma agora que a chave primária de R é {a,b,d}. Em que forma normal a relação R passa a estar?

Exercício 5 (Formas Normais)

Considere a relação R(a,b,c,d), em que {a,b} é a chave primária.

- Indique uma dependência funcional entre os atributos de R de forma a que R não esteja na 2ª forma normal.
- Indique uma dependência funcional entre os atributos de R de forma a que R esteja na 2ª forma normal mas não esteja na 3ª forma normal.
- Indique uma dependência funcional entre os atributos de R de forma a que R esteja na 3ª forma normal mas não esteja na Boyce Codd.

Exercício 6 (Formas Normais)

Considere a relação R (a,b,c) e a sua decomposição nas relações R1(b,c) e R2(a,c).

- Qual ou quais as dependências que têm necessariamente que existir entre os atributos a, b e c para que a decomposição de R em R1 e R2 não represente perda de informação?
- Assumindo a existência da ou das dependências anteriores, escreva uma expressão em álgebra relacional que permita reconstituir R a partir de R1 e R2 ?

DATAWAREHOUSE

Um **Datawarehouse** é um conjunto de dados centralizado de dados temáticos, históricos, datados, integrados e persistentes que oferece um nível de qualidade adequado às aplicações de decisão. Um Datawarehouse também pode ser definido como uma arquitectura de integração de dados que suporta a tomada de decisão através de:

- técnicas OLAP (interrogações estruturadas ou ad-hoc, geração de relatórios analíticos) para visualização e navegação de informação segundo diferentes perspectivas
- e técnicas de *data mining*: técnicas de análise de dados mais avançadas para compreender os dados.

Modelo conceptual de um DW pode existir nas seguintes formas:

- **Modelo em Estrela** constituído por uma tabela de factos sem dados redundantes no meio ligada a um conjunto de tabelas de dimensões;
- **Modelo em Floco de Neve** é um refinamento do modelo em estrela em que algumas dimensões estão normalizadas estando assim os dados de cada dimensão separados em várias tabelas;
- **Modelo em Constelação** constituído por múltiplas tabelas de factos que partilham tabelas de dimensões que são vistas como uma colecção de estrelas.

Tabelas de dimensões tipicamente contêm atributos descritivos, contêm muitas vezes uma hierarquia de conceitos (ex: dia < mês < ano na dimensão Tempo) e poucos tuplos

Tabela de factos contém uma chave composta pelas chaves estrangeiras para as tabelas de dimensão, contém atributos tipicamente numéricos que correspondem às medidas de negócio e muitos tuplos.

Cubo de dados (relativo a n dimensões) é uma representação lógica de modelo multidimensional em que as faces do cubo representam as dimensões e as células do cubo contêm valores das medidas (ou factos) para cada combinação de valores das dimensões.

Operações OLAP

- **Roll up** (drill-up): sumariza os dados, subindo na hierarquia dentro de uma dimensão ou reduzindo uma dimensão.
- **Drill down** (roll down): inverso do roll-up, desce de um nível de sumarização mais elevado para outro mais baixo, ou introduz uma nova dimensão.
- **Slice**: executa uma selecção sobre uma dimensão do cubo.
- **Dice**: define um sub-cubo executando uma selecção sobre duas ou mais dimensões
- **Pivot** (rotate): reorienta o cubo, representação de informação em 3D numa série de planos 2D
- **Drill across**: navega através de uma ou mais tabelas de factos

Exercício 7 (Data Mining – Exercício 2.5)

Suponha um datawarehouse que contém as seguintes quatro dimensões: Data, Espectador, Localização e Jogo e o facto “preço” que consiste no valor, em euros, pago por um espectador quando assiste a um determinado jogo numa dada data. Suponha ainda que os espectadores podem ser estudantes, adultos ou séniores, e que cada uma destas categorias tem o seu preço de bilhete.

Pretende-se que a dimensão data contenha o dia, mês e ano; que a dimensão Localização contenha o nome do estádio, e que a dimensão Jogo contenha o nome das duas equipas defrontadas.

- Desenhe um diagrama em estrela para este DW.
- Escreva em SQL a interrogação que devolve o preço total pago por espectadores estudantes para assistir ao jogo que se realizou no Estádio da Luz no dia 1 de Março de 2005.
- Apresente um esquema relacional que modele a mesma informação e uma interrogação SQL que devolva a mesma informação pedida em b). Comente quais as vantagens/desvantagens do diagrama em estrela face ao esquema relacional em termos das respectivas interrogações SQL.

Exercício 8 (Data Mining – Exercício 2.3)

Considere um data warehouse (DW) que consista nas 3 dimensões *tempo*, *médico* e *paciente*, e a medida *quantia*, correspondente à quantia cobrada pelo médico em cada consulta a um doente.

Pretende-se que a dimensão *tempo* contenha o dia, mês, trimestre e ano; a dimensão *médico* contenha o nome do médico, a especialidade, a cidade, o estado e o país; a dimensão *paciente* contenha o sexo e a faixa etária a que pertence.

- Desenhe o modelo em floco de neve para o DW acima.
- Começando pelo cubo [*dia*, *médico*, *paciente*], que operações OLAP devem ser executadas de modo a recolher o total cobrado por cada médico em Lisboa no ano 2004?
- De modo a obter a mesma informação, que interrogação SQL deveria executar se os dados estivessem registados numa base de dados relacional com o seguinte esquema

cobrança(dia, mês, ano, médico, cidade, estado, país, paciente, quantia)

Exercício 9

Desenhe um data warehouse para o Instituto Nacional de Metereologia. O instituto tem cerca de 1000 estações, que estão espalhadas por diferentes zonas em terra e no mar, e que recolhem dados básicos sobre as condições atmosféricas em cada hora, incluindo: pressão atmosférica, temperatura e precipitação.

- Desenhe o modelo em estrela para o DW acima.

- b) Começando pelo cubo [*hora, estação*], que operações OLAP devem ser executadas de modo a recolher o valor médio da temperatura em cada estação?
- c) E para recolher o valor médio da temperatura diário em cada estação? E mensal?
- d) Qual o tamanho da tabela de factos se forem feitos registos durante um ano?

Exercício 10

Considere uma cadeia de livrarias que usa um DW para gerar informação para melhorar o seu negócio. A cadeia tem várias livrarias espalhadas por várias cidades, dos diferentes distritos do país, tendo por vezes mais do que uma livraria na mesma cidade. Nas livrarias são vendidos livros escolares e não escolares, e cada livro diz respeito a uma matéria específica. No DW são registadas as quantidades de livros vendidos em cada dia e em cada loja, assim como o respectivo valor das vendas.

- a) Quais as dimensões para este DW? E as hierarquias?
- b) Desenhe o modelo em estrela para o DW acima.
- c) Desenhe o cubo correspondente.
- d) Que operações OLAP devem ser executadas de modo a recolher o valor médio das vendas em cada loja? E em cada cidade?

Capítulo 2 - ÁRVORES DE DECISÃO

CONCEITO DE ENTROPIA

$$\text{Entropia}(S) = \sum_{i=1}^c -p_i \log_2 p_i \leq \log_2 c$$

A Entropia é a medida da impureza do conjunto de treino. Assumindo o valor máximo (1) quando existem tantos elementos positivos como negativos, e o valor mínimo (0) quando todos os elementos são da mesma classe.

CONCEITO DE GANHO DE INFORMAÇÃO

$$\text{Ganho de Informação}(S, A) = \text{Entropia}(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} \text{Entropia}(S_v)$$

O ganho de informação é a redução esperada no valor da Entropia, devido à ordenação do conjunto de treino segundo os valores do atributo A.

ALGORITMO ID3

ID3 (**D**, **A**, **Atributos**)

// com **D** o conjunto de exemplos, **A** o atributo alvo, e **Atributos** a lista de atributos que caracteriza as entidades;

1. Criar um nó raiz da árvore;
2. Se todos os exemplos são positivos, rotular o nó como positivo (+) e devolver a árvore;
3. Se todos os exemplos são negativos, rotular o nó como negativo (-) e devolver a árvore;
4. Se a lista de atributos está vazia, devolver a árvore, com um rótulo idêntico ao valor assumido pela maioria dos exemplos;
5. Senão:
 - a. $A \leftarrow$ o atributo da lista de atributos, que melhor classifique os exemplos; usa-se o Ganho de Informação para escolher o atributo que melhor classifique os exemplos;
 - b. O nó recebe A, como rótulo;
 - c. Para cada valor possível v_i de $\text{Valores}(A)$:
 - i. Adicionar um novo ramo, a partir do nó, correspondente ao teste $A = v_i$;
 - ii. $D_{v_i} \leftarrow$ o subconjunto de D, com $A = v_i$;
 - iii. Se $D_{v_i} = \{ \}$:
Então devolver a árvore, com um rótulo idêntico ao valor assumido pela maioria dos exemplos;
Senão, por debaixo deste ramo adicionar a sub-árvore **ID3 (D_{v_i} , **A**, **Atributos**-{**A**})**

Exercício 1 (Machine Learning 3.1)

Desenhe as árvores de decisão para os seguintes problemas.

- a) $A \wedge \neg B$
- b) $A \vee [B \wedge C]$
- c) $A \text{ xor } B$
- d) $[A \wedge B] \vee [C \wedge D]$

Exercício 2

Considere o seguinte conjunto de exemplos de treino que descreve o conceito “Assinante da Revista Turbo”. Cada exemplo descreve as características de uma pessoa e se é assinante da revista. As características consideradas são: o sexo, a idade (<26 anos) e se tem carro.

	Sexo	Idade<26	Tem Carro	Cliente?
X1	M	T	F	Sim
X2	M	T	T	Sim
X3	F	T	T	Não
X4	M	F	T	Não
X5	F	T	F	Não
X6	M	T	F	Sim
X7	M	F	F	Não
X8	F	F	F	Não
X9	F	T	F	Não
X10	F	F	T	Não

- a) Calcule o valor da Entropia deste conjunto de treino no que diz respeito ao conceito Cliente?.
- b) Calcule o Ganho de Informação para o atributo Sexo.
- c) Face a este resultado, qual seria a estrutura de uma árvore de decisão obtida para este conjunto de treino, construída de acordo com o critério de maximização do ganho de informação.

Exercício 3 (Machine Learning – Exercício 3.2)

Considere o seguinte conjunto de treino.

	A1	A2	F
1	T	T	Yes
2	T	T	Yes
3	T	F	No
4	F	F	Yes
5	F	T	No
6	F	T	No

- a) Calcule o valor da Entropia deste conjunto de treino no que diz respeito à função objectivo F.
- b) Calcule o Ganho de Informação para o atributo A2

Exercício 4 (MACHINE LEARNING – Exercício 3.4)

Considere o seguinte conjunto de treino.

	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
--	-----	---------	----------	------	-------	----------	------------

1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- Desenhe a árvore de decisão para o conjunto de treino.
- Qual a relação entre essa árvore de decisão e o espaço de versões determinado no exercício 5 da aula 1.
- Adicione o seguinte exemplo ao conjunto de treino, e volte a calcular a árvore de decisão.

	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
5	Sunny	Warm	Normal	Weak	Warm	Same	No

Exercício 5 (MACHINE LEARNING – EXEMPLO 3.4.2)

Considere o seguinte conjunto de treino.

Outlook	Temp	Humidity	Wind	Play?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

- Desenhe a árvore de decisão para o conjunto de treino.

Exercício 6 (MINI-TESTE 1995)

Suponha que tem o seguinte conjunto de treino, em que cada exemplo é definido por três atributos (A,B,C).

	A	B	C	X
X1	1	1	3	P
X2	1	2	4	P
X3	2	2	4	N
X4	2	1	4	N

- Qual a incerteza associada ao conjunto de treino inicial?
- Qual o Ganho de Informação de um teste efectuado no primeiro atributo (X1)?
- Face a este resultado, qual seria a estrutura de uma árvore de decisão obtida para este conjunto de treino, construída de acordo com o critério de maximização do ganho de informação.

Exercício 7 (MINI-TESTE 1996)

Suponha que tem o seguinte conjunto de treino, em que cada exemplo é definido por três atributos $(A,B,C) \in \{0, 1\} \times \{0, 1\} \times [0, 20]$.

	A	B	C	X
X1	1	1	5.0	P
X2	0	1	11.5	P
X3	0	1	13.5	P
X4	1	0	15.0	P
X5	1	0	13.0	P
X6	0	0	11.5	N
X7	0	1	5.0	N
X8	1	0	9.5	N
X9	0	0	2.0	N
X10	1	0	8.0	N

- Qual a incerteza associada ao conjunto de treino inicial?
- Qual é o teste seleccionado para a raiz de uma árvore de decisão usando o critério do máximo ganho de informação igual ao usado no sistema ID3?

Exercício 8 (TESTE 1999)

Considere o seguinte conjunto de treino.

	A	B	C	D	X
X1	0	0	1	0	N
X2	1	0	0	0	N
X3	0	1	1	1	P
X4	1	1	1	0	P
X5	0	0	1	0	N
X6	1	1	1	1	P
X7	0	1	0	1	N
X8	0	0	0	0	N

- Qual a incerteza associada ao conjunto de treino inicial?
- Determine qual o atributo testado na raiz de uma árvore de decisão obtida com o algoritmo ID3.
- Construa o resto da árvore de decisão obtida por este algoritmo. Indique se o resultado obtido é intuitivamente razoável.

Exercício 9 (EXAME 18-12-1999)

Considere o seguinte conjunto de treino.

	Z1	Z2	Z3	X
X1	A	0	1	N
X2	B	1	0	P
X3	C	1	1	P
X4	A	1	1	N
X5	C	0	1	N
X6	A	1	1	N
X7	B	1	0	P
X8	C	0	0	N

- Qual a incerteza associada ao conjunto de treino inicial?

- b) Determine qual o atributo testado na raiz de uma árvore de decisão obtida com o algoritmo ID3.

Exercício 10 (EXAME 12-02-2000)

Considere o espaço de hipótese por árvores de decisão regulares num espaço de n atributos booleanos. Uma árvore de decisão regular é uma árvore de decisão com 2 níveis de nós internos e 4 folhas, em que os dois nós filhos da raiz testam a mesma variável.

Como função de n , quantas árvores de decisão regulares distintas existem?

Exercício 11 (EXAME 28-10-2000)

Para este problema, considere o espaço de instâncias representado pelo espaço booleano de 6 dimensões e admita que lhe é fornecido um conjunto de treino que inclui todas as instâncias deste espaço, rotulado de acordo com a expressão booleana $f = x_1 \vee \bar{x}_1 x_2 x_3$.

- Indique qual a dimensão do espaço de instâncias, e qual o número de hipóteses distintas existentes neste espaço.
- Determine qual o atributo testado na raiz de uma árvore de decisão obtida com o algoritmo ID3. Indique os valores obtidos para a entropia original e para o ganho de informação obtido para cada um dos atributos.
- Descreva, justificando, qual a árvore de decisão obtida pelo algoritmo ID3.

Exercício 12 (EXAME 11-01-2001)

Considere o espaço de instâncias definido no espaço rectangular de $[0.0, 1.0] \times [0.0, 1.0]$. Cada instância neste espaço é representada por um par de números decimais do intervalo $[0.0, 1.0]$, arredondados às décimas. Neste espaço suponha que lhe é dado o seguinte conjunto de treino:

	Z1	Z2	F
X1	1	1	A
X2	1	7	A
X3	1	9	A
X4	2	3	A
X5	3	8	A
X6	4	2	A
X7	6	4	A
X8	7	3	A
X9	7	6	B
X10	7	8	A
X11	8	4	A
X12	9	2	B
X13	9	9	A
X14	10	4	B

- Para este conjunto de treino, indique o atributo testado na raiz de uma árvore de decisão em que cada nó executa um teste do tipo $z_i \geq Z$, onde Z é um número do intervalo $[0.0, 1.0]$ com uma casa decimal.

- b) Descreva a árvore de decisão obtida, e ilustre graficamente a área a que correspondem objectos do tipo B.
- c) Indique qual o número total de conceitos que existe neste espaço.
- d) Indique qual o número de conceitos que é representável por uma árvore de decisão que contenha exactamente dois nós internos.

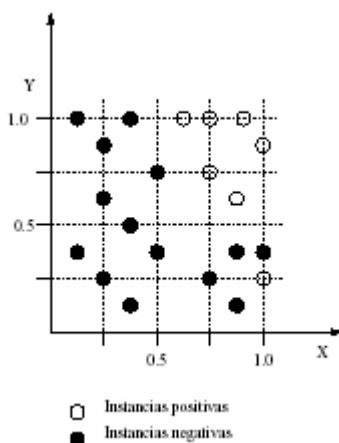
Exercício 13 (TESTE 4-05-2002)

Considere o seguinte conjunto de treino:

X_1	X_2	X_3	Classe
0.3	0.7	0.8	N
0.7	0.3	0.7	P
0.8	0.7	0.8	P
0.8	0.2	0.6	P
0.1	0.7	0.2	P
1.0	0.1	0.3	N
0.9	0.2	0.4	N
0.3	0.3	0.1	N

- a) Qual a incerteza $I(p, n)$ associada ao conjunto de treino inicial ?
- b) Utilizando apenas testes da forma $x_i \geq 0.5$, determine qual o atributo testado na raiz de uma árvore de decisão que utilize o algoritmo ID3.
- c) Determine qual a árvore de decisão obtida com esse algoritmo.
- d) Esboce, num espaço de três dimensões, para $0.0 \leq x_1, x_2, x_3, x_4 \leq 1.0$ as zonas que a árvore de decisão considera conterem instâncias positivas do conceito.

Exercício 14 (EXAME 11-06-2002)



Considere o conjunto de treino representado na figura acima, onde os valores de x e y estão restritos ao intervalo $[0.0, 1.0]$. Neste conjunto, embora os atributos sejam originalmente contínuos, todos os valores foram discretizados para valores múltiplos de 0.125.

- a) Qual a incerteza $I(p, n)$ associada ao conjunto de treino inicial?

- b) Diga qual a árvore de decisão com dois nós internos que é gerada pelo algoritmo ID3 para este conjunto de treino. Represente, graficamente, a zona que contém as instâncias positivas, de acordo com a árvore de decisão obtida.
- c) Quantos nós internos terá, no mínimo, uma árvore de decisão que permita classificar o conjunto de treino com 100% de precisão?
- d) Descreva um critério ou um mecanismo que permita escolher a dimensão da árvore de decisão que melhores resultados obterá quando for aplicada a um conjunto de teste independente do conjunto de treino. Pode assumir que existem mais dados além dos disponíveis no conjunto de treino.

Exercício 15 (EXAME 12-07-2002)

Considere o seguinte conjunto de treino:

X_1	X_2	X_3	Classe
v1	v3	v7	P
v1	v3	v8	N
v1	v4	v9	P
v2	v5	v9	N
v2	v6	v8	P

- a) Qual a incerteza $I(p, n)$ associada ao conjunto de treino inicial?
- b) Determine qual o atributo testado na raiz de uma árvore de decisão obtida com o algoritmo ID3.

Capítulo 3 - REDES NEURONAIS

PERCEPTRÃO

Cálculo da Saída do Perceptrão

$$o(x_1, \dots, x_n) = \begin{cases} 1, & \text{se } w_0x_0 + w_1x_1 + \dots + w_nx_n > 0 \\ -1, & \text{caso contrário} \end{cases}$$

Regra de Treino do Perceptrão – superfícies linearmente separáveis

Sendo:

- $t=c(x)$ é o valor assumido pela função objectivo para o elemento x ;
- o é a saída do perceptrão, ou seja, a estimativa da função objectivo;
- η é o rácio de aprendizagem de valor reduzido.

$$\begin{cases} w_i \leftarrow w_i + \Delta w_i \\ \Delta w_i \leftarrow \eta (t - o) x_i \end{cases}$$

Regra Delta – Gradiente Descendente – superfícies não linearmente separáveis

Gradiente Descendente (D, η) (aproximação estocástica)

Cada elemento de D é da forma $\langle \xi, \bar{t} \rangle$, onde ξ é o vector dos valores entrados e $\bar{t}$ o vector de saída.

1. Iniciar todos os pesos w_i com valores aleatórios pequenos
2. Até que seja satisfeita a condição de paragem repetir:
 - a. Iniciar cada Δw_i a 0.
 - b. Para cada elemento $\langle \xi, \bar{t} \rangle$ do conjunto de treino, fazer:
 - i. Dar entrada do elemento na rede e calcular a saída da rede;
 - ii. Para cada w_i actualizar o seu valor

$$w_i \leftarrow w_i + \eta(t - o)x_i$$

Exercício 1

Desenhe o perceptrão que implementa a função lógica OR.

E o perceptrão que implementa a função lógica AND.

Exercício 2 (MACHINE LEARNING – EXERCÍCIO 4.1)

Determine quais os valores dos pesos w_0 , w_1 e w_2 para o perceptrão cuja superfície de decisão é dada pela inequação $y > 2x + 2$.

Exercício 3 (MACHINE LEARNING – EXERCÍCIO 4.2)

Desenhe o perceptrão de duas entradas que implementa a função lógica $A \wedge \neg B$.

Desenhe o perceptrão de duas camadas que implementa a função lógica $A \text{ xor } B$. Note que, por **A xor B** não se tratar de um conjunto de treino linearmente separável, se deve usar a Regra Delta, e uma rede de perceptrões com duas camadas.

Exercício 4 (TESTE 2000)

Para este problema, considere o espaço de instâncias representado pelo espaço booleano de 6 dimensões e admita que lhe é fornecido um conjunto de treino que inclui todas as instâncias deste espaço, rotulado de acordo com a expressão booleana $f = x_1 \vee \neg x_1 x_2 x_3$.

- Indique se um perceptrão de uma camada com uma função do tipo degrau discreto pode representar a função representada no problema 1.
- Se a sua resposta à alínea anterior foi negativa, apresente a justificação. Se foi positiva, apresente os valores dos pesos w_i , $0 \leq i \leq 6$ que permitem representar a função descrita.

REDES MULTI CAMADA**Backpropagation(D, η , n_{in} , n_{out} , n_{hidden})**

Cada elemento de D é da forma $\langle \xi, t \rangle$, onde ξ é um vector com n_{in} dimensões e t o vector com n_{out} dimensões; n_{in} é o nº de unidades de entrada, n_{hidden} é o nº de unidades intermédias e n_{out} é o nº de unidades de saída; A entrada da unidade j vinda da unidade i é denotada por x_{ji} , e o peso da unidade i para a unidade j é denotada por w_{ji} .

- Criar a rede com n_{in} entradas, n_{hidden} unidades intermédias e n_{out} unidades de saída.
- Iniciar todos os pesos com valores aleatórios pequenos.
- Até que seja satisfeita a condição de paragem repetir:
 - Para cada elemento do conjunto de treino fazer:
 - Dar entrada do elemento na rede e calcular as saídas da rede;
 - Para cada unidade de saída k , avalie o valor do erro associado à unidade j para as unidades de saída δ_k :

$$\delta_k = -\partial E / \partial net_k = -\partial E / \partial o_k \times \partial o_k / \partial net_k$$

$$\delta_k = -\partial E / \partial o_k \times f'(net_k)$$
 - Retro-propague os valores de δ_j para as unidades intermédias:

$$\delta_j = f'(net_j) \times \sum_k w_{kj} \delta_k \text{ (com } k \in \text{outputs de } j)$$
 - Actualizar cada peso w_{ji} da rede

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$$

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1)$$

(com n o número de iterações efectuadas)

$$\partial E / \partial w_{ji} = \delta_j \times o_i$$

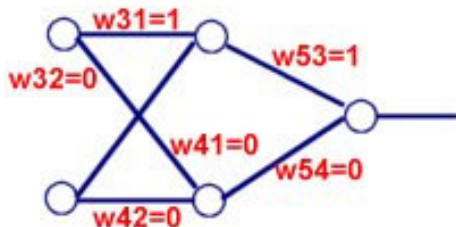
Exercício 5 (MACHINE LEARNING – EXERCÍCIO 4.7)

Considere a rede neuronal de duas entradas a e b, uma unidade interna c e uma unidade de saída d. Esta rede tem 5 pesos (w_{ca} , w_{cb} , w_{c0} , w_{dc} , w_{d0}) onde w_{x0} representa o peso threshold para a unidade x. Inicie todos os pesos com 0.1, e diga quais os seus valores depois de duas iterações do algoritmo Backpropagation. Assuma que o rácio de aprendizagem é $\eta=0.3$, $\alpha=0.9$, uma actualização incremental dos pesos de acordo com a equação 1, os seguintes exemplos de treino e que os nós implementam a função sigmóide.:

a	b	d
1	0	1
0	1	0

Exercício 6

Considere a seguinte rede



Com entrada $\xi = (1,0)$ e saída $t=0.75$. Considere que

$$f(x) = \frac{1}{1 + 2^{-x}}$$

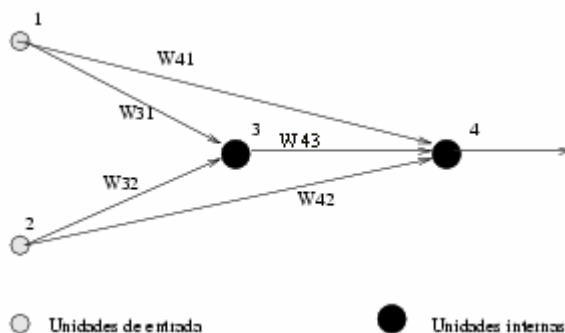
$$f'(x) = \frac{2^{-x} \ln 2}{(1 + 2^{-x})^2}$$

$$E = \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - \sigma_k)^2$$

- Determine o valor da saída da rede para o estímulo de entrada (1,0).
- Determine o erro associado à unidade 5 (δ_5).
- Determine o valor da derivada do erro em relação ao peso w_{53} . ($\partial E / \partial w_{53}$)
- Determine o valor da derivada do erro em relação ao peso w_{31} . ($\partial E / \partial w_{31}$)

Exercício 7 (TESTE 1996)

Suponha que se pretende utilizar o algoritmo de *back-propagation* para se actualizar os pesos da seguinte rede neuronal através da minimização do erro $E = (o_{p4} - t_{p4})^2$, onde o_{pi} é a saída observada para a unidade i e para o padrão de entrada p .



Cada nó n_i que não seja um nó de entrada implementa a função

$$g_i = f(w_{i0} + \sum_j w_{ij} g_j)$$

$$f(x) = \frac{1}{1 + e^{-x}}$$

Assumindo que $w_{ij} = 0.5$ para todo o i e j e que o estímulo de entrada é (0; 1) e se pretende uma saída de 1, calcule:

- O valor da saída para este estímulo.
- A derivada parcial do erro em relação ao peso w_{42} .
- O valor da derivada parcial do erro em relação ao peso w_{32} .

Exercício 8 (TESTE 2ª ÉPOCA 1999-2000)

Descreva qual é a rede neuronal mais simples que, utilizando perceptrões, permite representar um espaço de hipóteses que contenha o espaço de hipóteses constituído por árvores de decisão regulares num espaço de n atributos booleanos. Uma árvore de decisão regular é uma árvore de decisão com 2 níveis de nós internos e 4 folhas, em que os dois nós filhos da raiz testam a mesma variável.

Exercício 9 (EXERCÍCIO 4.5 ML)

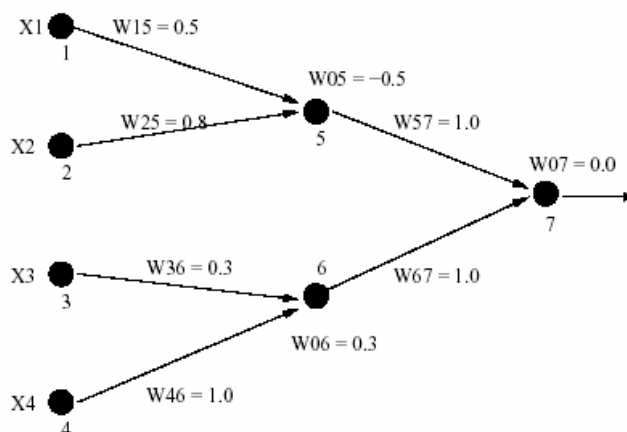
Derive a regra do gradiente descendente para uma rede constituída apenas por uma unidade com saída σ , onde

$$\sigma = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

Exercício 10 (2ª SÉRIE DE PROBLEMAS 2001-2002)

Considere o seguinte conjunto de treino e a rede apresentada em seguida:

X_1	X_2	X_3	X_4	Classe
0	0	0	0	0
0	1	0	1	1
1	0	0	0	0
1	1	1	0	0
0	0	0	1	0
0	1	0	1	0
1	0	0	1	0
1	1	1	1	1
1	0	0	1	0
0	1	0	1	1
1	0	0	0	0
1	1	1	1	1
0	0	0	1	0
0	1	0	1	1
1	0	0	1	0
1	0	1	1	0



Cada nó da rede com índice j (excepto os de entrada) implementam a função $x_j = f(w_{j0} + \sum_i w_{ji} x_i)$ com f a função sigmoidal.

Considerando que esta rede é treinada em modo *batch* usando como medida do erro a soma dos quadrados dos desvios, calcule (sugestão: utilize uma folha de cálculo):

- A derivada parcial do erro em relação ao peso W_{57} .

- b) A derivada parcial do erro em relação ao peso W_{15} .

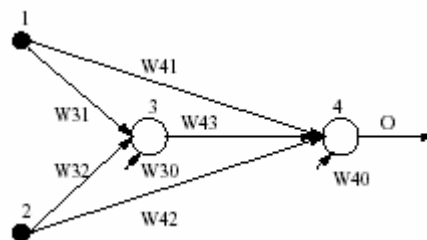
Exercício 11 (TESTE 4/05/2002)

Indique quais os pesos que utilizaria numa rede neuronal definida sobre o espaço $\{1, 0\}^4$ com apenas um nó, que use a função sigmoideal, para aproximar a função Booleana $g=x_1x_2+x_4$. Considere que os pesos estão limitados ao intervalo $[-10.0, +10.0]$.

Nota: Este nó implementa a função $f=f(w_0+\sum_{i=1}^4 w_i x_i)$ sendo f a função sigmoideal.

Exercício 12 (EXAME 11/06/2002)

Considere a rede neuronal ilustrada na figura.



onde cada nó da rede com índice j (excepto os de entrada) implementam a função $x_j=f(w_{j0}+\sum_i w_{ji}x_i)$ com f a função sigmoideal. Assuma que o erro que se pretende minimizar é dada por $E(\bar{W})=1/2\sum_{d\in D}W(t_d-O_d)^2+\gamma\sum_{i,j}W_{ij}^2$

- Indique a razão pela qual existe o segundo termo na expressão do erro, e qual o seu efeito.
- Determine a expressão para a derivada parcial do erro em relação a W_{42} .

Exercício 13 (EXAME 12/07/2002)

Considere que se pretende treinar um perceptrão, sem unidades escondidas, usando a regra delta (*delta rule*), para aprender o seguinte conjunto de treino:

- Indique se é garantido que este método de treino vai ou não convergir para uma solução do problema, e em caso afirmativo, em que condições.
- Assuma que a unidade tem um threshold, discreto, cujo valor está inicializado a 0, tal como todos os pesos. Usando um valor $\mu=0.1$, indique o valor dos pesos após processar, uma vez, cada um dos exemplos de treino. (O nó está activo (isto é, tem a saída a 1) se a soma pesada dos valores de entrada for igual ou superior ao valor do threshold).
- Diga quais as vantagens em usar unidades que usam uma função de activação diferenciável, (função logística ou tangente hiperbólica).
- Descreva duas forma de evitar o problema do sobre-aprendizagem (overfitting quando se usam redes neurais treinadas com retro-propagação).

Capítulo 4 - APRENDIZAGEM BAYESIANA

NOÇÕES BÁSICAS DE PROBABILIDADES

Regra do Produto: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

Regra da Soma: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Teorema de Bayes

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Teorema das Probabilidades Totais:

se $A_1, \dots, A_n$ são eventos mutuamente exclusivos ou seja $\sum_{i=1}^n P(A_i) = 1$ então

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

Com:

- $P(h)$ → probabilidade à priori da hipótese h ; reflecte o conhecimento sobre o domínio.
- $P(D)$ → probabilidade à priori do conjunto de treino D ;
- $P(D|h)$ → probabilidade do conjunto de treino D dada a hipótese h ;
- **$P(h|D)$** → probabilidade da hipótese h dado o conjunto de treino D ; reflecte a confiança de que h é verdadeiro face às observações efectuadas e registadas em D – objecto da Aprendizagem.

Algoritmo Base para Aprendizagem da Hipótese MAP

- Para cada hipótese (h) do espaço de hipóteses (H), calcular $P(h|D)$

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$

$P(D)$ pode ser eliminado uma vez que é constante e independente da hipótese h .

- Escolher h_{MAP} como modelo do conjunto D .

- Não existindo conhecimento sobre a crença de que uma hipótese é mais provável que qualquer outra

$$P(h) = \frac{1}{|H|} \quad \forall h \in H$$

- Assumindo que o conjunto de treino D não contém ruído, a probabilidade de observar um exemplo x_i com classificação d_i dado h é 1 se $d_i = h(x_i)$ e 0 caso contrário. Onde

$$P(D|h) = \begin{cases} 1, & \text{se } d_i = h(x_i) \quad \forall d_i \in D \\ 0, & \text{caso contrário} \end{cases}$$

Classificador Ótimo de Bayes (Bayes Optimal Classifier)

Quando aplicado à classificação de novas instâncias, uma possível optimização do algoritmo anterior consiste em pesar as probabilidades de cada hipótese dado o conjunto de treino $[P(h|D)]$, pelas probabilidades de cada classificação possível v_j face a cada hipótese h_i .

$$h = \arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j | h_i) P(h_i | D)$$

Exercício 1 (MACHINE LEARNING – EXEMPLO 6.2.1)

Um doente foi submetido a um exame de laboratório para detectar se tinha cancro, e o resultado do exame deu **positivo**. Este exame devolve um resultado positivo correcto em apenas 98% dos casos em que a doença está efectivamente presente, e um resultado negativo correcto em apenas 97% dos casos em que a doença não está presente. Para além disso, 0.008 da população total sofre de cancro.

Determine se o doente tem cancro ou não.

Exercício 2 (MACHINE LEARNING – EXERCÍCIO 6.1)

Suponha que o doente é submetido a um novo teste, e que este também dá um resultado positivo. Qual a probabilidade *a posteriori* para as duas hipóteses depois destes dois testes. Assuma que os testes são independentes.

CLASSIFICADOR DE NAIVE BAYES

Adequado a casos em que as instâncias são descritas por uma conjunção de atributos $x = \langle a_0, a_1, \dots, a_n \rangle$

$$\begin{aligned} v_{NB} &= \arg \max_{v_j \in V} P(v_j | a_0, a_1, \dots, a_n) \\ &= \arg \max_{v_j \in V} \frac{P(a_0, a_1, \dots, a_n | v_j) P(v_j)}{P(a_0, a_1, \dots, a_n)} \\ &= \arg \max_{v_j \in V} P(a_0, a_1, \dots, a_n | v_j) P(v_j) \\ &= \arg \max_{v_j \in V} P(v_j) \prod_{i=0}^n P(a_i | v_j) \end{aligned}$$

Exercício 3 (MACHINE LEARNING – EXEMPLO 6.9.1)

Considere o seguinte conjunto de treino.

Outlook	Temp	Humidity	Wind	Play?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Rainy	Mild	High	Strong	No
-------	------	------	--------	----

Nestas condições, determine justificando, a classificação obtida pelo algoritmo Naive Bayes para os seguintes exemplos:

- <Sunny, Cool, High, Strong>
- <Overcast, Hot, Normal, Strong>
- <Rainy, Cool, High, Weak>

Exercício 4 (EXAME 28-10-2000)

Para este problema considere o espaço de instâncias representado pelo espaço booleano de 6 dimensões e admita que lhe é fornecido um conjunto de treino que inclui todas as instâncias deste espaço, rotulado de acordo com a expressão booleana $f = x_1 \vee \bar{x}_1 x_2 x_3$.

- Determine a classificação obtida para a instância 100010, utilizando o algoritmo *Naive Bayes*. Recorde que este algoritmo classifica os exemplos aplicando a regra de Bayes e assumindo que a probabilidade de um atributo dada a classe da instância é independente dos valores dos outros atributos.
- Repita a alínea anterior para a instância 011011.
- Comente os resultados obtidos e indique, justificando detalhadamente, se, neste problema estão satisfeitas as condições para que este classificador seja ótimo.

Exercício 5 (TESTE 1999)

Considere o seguinte conjunto de treino.

	z_1	z_2	z_3	z_4	X
X1	0	0	1	0	N
X2	1	0	0	0	N
X3	0	1	1	1	P
X4	1	1	1	0	P
X5	0	0	1	0	N
X6	1	1	1	1	P
X7	0	1	0	1	N
X8	0	0	0	0	N

- Determine a probabilidade $P(z_i | P)$ e $P(z_i | N)$, para $i=1, 2, 3$ e 4.
- Determine a classificação obtida para os exemplos 1010 e 0110 utilizando o algoritmo *naive Bayes*. Recorde que este algoritmo classifica os exemplos aplicando a regra de Bayes e assumindo que $P(z_1, z_2, z_3, z_4 | X) = P(z_1 | X)P(z_2 | X)P(z_3 | X)P(z_4 | X)$
- Indique, justificando, se, neste problema, estão satisfeitas as condições para que este classificador seja ótimo.

Exercício 6 (EXAME 18-12-1999)

Determine a classificação obtida para o exemplo C11 utilizando o algoritmo *naive Bayes*. Recorde que este algoritmo classifica os exemplos aplicando a regra de Bayes e assumindo que $P(z_1, z_2, z_3 | X) = P(z_1 | X)P(z_2 | X)P(z_3 | X)$, considerando o seguinte conjunto de treino.

	Z1	Z2	Z3	X
X1	A	0	1	N
X2	B	1	0	P
X3	C	1	1	P
X4	A	1	1	N
X5	C	0	1	N
X6	A	1	1	N
X7	B	1	0	P
X8	C	0	0	N

Exercício 7 (2ª SÉRIE DE PROBLEMAS DE 2002)

Considere o conjunto de treino do exercício 2.13. Calcule a classe atribuída pelo algoritmo de Naíve Bayes aos exemplos $(x_1, x_2, x_3, x_4)=(0,1,1,1)$ e $(x_1, x_2, x_3, x_4)=(1,0,1,1)$.

Exercício 8 (EXAME 11-06-2002)

Suponha que se pretendia aplicar o algoritmo de classificação Naive Bayes ao problema descrito no exercício 2.14.

- Descreva que pré-processamento aplicaria aos dados para poder aplicar este método a este problema.
- Usando o método que descreveu na alínea anterior, qual a classificação obtida para a instância com coordenadas (1.0 1.0)?

Exercício 9 (EXAME 12-07-2002)

Considere o conjunto de treino do exercício 2.15. Determine a classificação obtida para o exemplo $v_2v_4v_9$ utilizando o algoritmo naive Bayes. Recorde que este algoritmo classifica os exemplos aplicando a regra de Bayes e assumindo que $P(x_1, x_2, x_3|X)=P(x_1|X) P(x_2|X) P(x_3|X) P(X)$.

REDES DE BAYES

Uma **rede de Bayes** ou **de crenças** é um grafo sem ciclos, em que:

- Os nós representam variáveis aleatórias;
- Os arcos representam influência directa de um nó noutro;
- Cada nó tem uma tabela de probabilidades condicionada, que quantifica o efeito dos nós pais no próprio nó.

Algoritmo de Construção de Redes de Bayes

- Escolha o conjunto das variáveis relevantes que descrevem o domínio.
- Escolha uma ordem para as variáveis, de modo a que os pais de um nó surjam sempre antes do nó de acordo com essa ordenação.
- Para cada variável X_i :
 - Selecione uma variável X_i e adicione um nó à rede.
 - Escolha o conjunto de nós pais para a variável X_i , de modo a que este conjunto seja mínimo e que $P(X_i | X_{i-1}, \dots, X_1)=P(X_i | \text{Pais}(X_i))$.
 - Defina a tabela de probabilidades condicionada para X_i .

Exercício 10 (ARTIFICIAL INTELLIGENCE: A MODERN APPROACH – EXEMPLO)

Considere a seguinte situação: Você instalou um alarme de assaltos em casa. Este alarme é razoavelmente de confiança na detecção de assaltos, mas também responde ocasionalmente a pequenos tremores de terra. Para além disso você tem dois vizinhos, John e Mary, a quem pediu que lhe telefonassem para o emprego quando ouvissem o alarme. O John telefona sempre que ouve o alarme, mas por vezes confunde o toque do telefone com o alarme e também telefona. Por outro lado, a Mary gosta de ouvir a música bastante alto e por vezes não ouve o alarme tocar. Dada a evidência de quem telefonou ou não, gostaríamos de estimar a probabilidade de ter ocorrido um assalto. Desenhe a rede que representa esta situação.

Exercício 11 (MACHINE LEARNING – EXERCÍCIO 6.6)

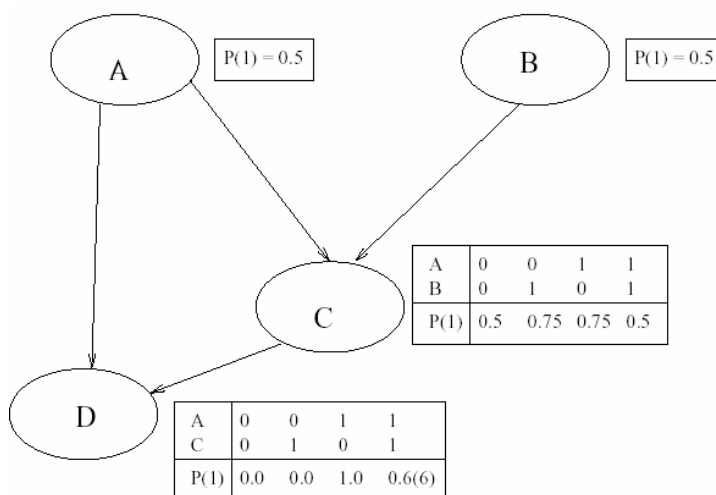
Desenhe a rede de Bayes para o problema PlayTennis, tendo em conta o seguinte conjunto de treino.

Outlook	Temp	Humidity	Wind	Play?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Defina a tabela de probabilidades condicionadas associada ao nó *Wind*.

Exercício 12 (EXAME 13-02-2001)

Considere a seguinte rede de Bayes:



- a) Considerando que os atributos A,B,C e D são binários, enumere um conjunto de instâncias cujas probabilidades condicionadas reflectam exactamente os parâmetros desta *rede de Bayes*.

- b) Se lhe forem dados os valores $A=1$, $B=1$ e $C=1$, qual é a probabilidade de D ser 1?
- c) Se lhe forem dados os valores $A=1$ e $D=0$, qual é a probabilidade de B ser 0?
- d) Considere o conjunto de instâncias que obteve na alínea a) e considere que pretendemos prever D a partir dos valores de A , B e C . Determine:
 - i) Qual a entropia do conjunto de treino inicial?
 - ii) Qual o atributo testado numa árvore de decisão que use a regra do ganho de informação?

Exercício 13 (EXAME 11-01-2001)

Considere o espaço de instâncias definido no espaço rectangular de $[0.0, 1.0] \times [0.0, 1.0]$. Cada instância neste espaço é representada por um par de números decimais do intervalo $[0.0, 1.0]$, arredondados às décimas. Neste espaço suponha que lhe é dado o conjunto de treino apresentado a seguir.

	z_1	z_2	F
X1	1	1	A
X2	1	7	A
X3	1	9	A
X4	2	3	A
X5	3	8	A
X6	4	2	A
X7	6	4	A
X8	7	3	A
X9	7	6	B
X10	7	8	A
X11	8	4	A
X12	9	2	B
X13	9	9	A
X14	10	4	B

Assuma que os atributos z_1 e z_2 foram discretizados, usando o valor 0.5 como limite. Assim, $z_1 < 0.5$ passa a corresponder ao valor 0 e $z_1 \geq 0.5$ passa a corresponder ao valor 1. O mesmo se aplica a z_2 .

- a) Admitindo que cada instância passa a ser descrita por 3 atributos (z_1 , z_2 , e o rótulo A ou B), descreva qual a menor *rede de Bayes* que permite descrever a distribuição estatística observada.
- b) Descreva os coeficientes das tabelas de cada um dos nós da rede de Bayes obtida.

Exercício 14 (ARTIFICIAL INTELLIGENCE: A MODERN APPROACH – EXERCÍCIO 15.2)

Numa estação nuclear existe um alarme que é activado quando a medida da temperatura excede um determinado valor. Considere as variáveis booleanas A (alarme activado), F_A (falha no alarme) e F_T (falha no termómetro), e as variáveis reais V (o valor lido pelo termómetro) e T (a temperatura actual).

- a) Desenhe a rede de Bayes para este domínio, sabendo que o termómetro tem maior tendência a falhar quando a temperatura se torna muito alta.
- b) Suponha que quer o valor da temperatura actual quer do valor lido pelo termómetro só assume um de dois valores: normal e alta; e que o termómetro fornece dados incorrectos $x\%$ do tempo

em que está a funcionar, e que tem falhas $y\%$ do tempo. Determine a tabela de probabilidades condicionada associada ao nó V.

- c) Suponha que o alarme funciona todo o tempo em que não está a falhar. Determine a tabela de probabilidades condicionada associada ao nó A.
- d) Suponha que quer o termómetro quer o alarme estão a funcionar, e que o alarme é activado. Calcule a probabilidade de a temperatura ser demasiado elevada.

Exercício 15 (ARTIFICIAL INTELLIGENCE: A MODERN APPROACH – EXERCÍCIO 15.2)

Dois astrónomos, em diferentes partes do mundo, fazem medições M_1 e M_2 do número de estrelas numa determinada região do céu, utilizando para tal os seus telescópios. Normalmente, existe uma pequena possibilidade de erro associada à contagem as estrelas, acrescentando uma ao número real. Cada telescópio pode (com uma pequena probabilidade) estar mal focada (evento F_1 e F_2), o que conduz a um acontagem deficiente em 3 ou mais estrelas.

Desenhe a rede que representa o domínio descrito.

Exercício 16 (2ª SÉRIE DE PROBLEMAS 2000-2001)

Determine a *rede de Bayes* com menor número de parâmetros, que modela exactamente a distribuição observada no seguinte conjunto de dados:

	z1	z2	z3	z4	z5
X1	0	0	0	0	0
X2	0	1	0	0	0
X3	1	0	0	0	0
X4	1	1	1	0	1
X5	0	0	0	1	1
X6	0	1	0	1	1
X7	1	0	0	1	1
X8	1	1	1	1	1

PRINCÍPIO DAS DESCRIÇÕES MAIS COMPACTAS (MDL)

O MDL corresponde à perspectiva bayesiana do Occam's razor ("a explicação mais curta é sempre a melhor"). A motivação de partida é a interpretação da definição de h_{MAP} , à luz dos conceitos da Teoria da Informação.

$$\begin{aligned}
 h_{MAP} &= \arg \max_{h \in H} P(D|h)P(h) \\
 &\Leftrightarrow h_{MAP} = \arg \max_{h \in H} \log_2 P(D|h) + \log_2 P(h) \\
 &\Leftrightarrow h_{MAP} = \arg \min_{h \in H} -\log_2 P(D|h) - \log_2 P(h)
 \end{aligned}$$

Onde

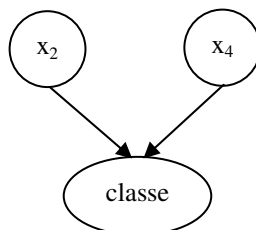
- $-\log_2 P(h) \rightarrow$ é o tamanho de h sob a codificação óptima para o espaço de hipóteses H , ou seja é o tamanho de h usando a representação óptima.
- $-\log_2 P(D|h) \rightarrow$ é o tamanho do conjunto de treino D dada a hipótese h , sob a representação óptima.

$h_{MAP} = h_{MDL}$ quando se usa a representação óptima.

Repare-se que $P(D|h)$ é independente da codificação dos elementos de treino (x) – pois são sempre os mesmos para todas as hipóteses –, apenas depende da codificação da classificação atribuída pela hipótese h [$f(x)$]. Seja $m = \#D$, ora para codificar cada um dos exemplos são necessários $\log_2 m$ bits, e apenas é necessário identificar os exemplos mal classificados.

Exercício 17 (2ª SÉRIE DE PROBLEMAS 2002)

Calcule a probabilidade da classe ser 1 para os exemplos do exercício 4.7 admitindo que o problema é adequadamente modelado pela rede de Bayes da figura seguinte:



Capítulo 5 - SUPPORT VECTOR MACHINES (SVMs)

Exercício 1

Considere que os vectores de suporte extraídos a partir de um determinado conjunto de treino são

$\mathbf{p}_1=[1.0; 0.0]$ com $y_1=1$ e $\alpha_1=2$; $\mathbf{p}_2=[0.0; 1.0]$ com $y_2=1$ e $\alpha_2=2$; $\mathbf{p}_3=[0.0; 0.0]$ com $y_3=-1$ e $\alpha_3=4$ e que o parâmetro $\mathbf{b}$ do classificador é -1 .

a) Como classifica os seguintes elementos, usando um kernel linear?

i. $\mathbf{x}_1=[1.0; 1.0]$

ii. $\mathbf{x}_2=[0.0; 0.0]$

iii. $\mathbf{x}_3=[0.5; 0.0]$

b) Que função é representada?

Exercício 2

a) Usando os resultados do exercício anterior, determine os vectores de suporte para a função *conjunção*.

b) Volte a classificar os pontos do exercício anterior, com um kernel linear e com os seguintes parâmetros: $b=-3$; $y_1=1$ e $\alpha_1=4$; $y_2=-1$ e $\alpha_2=2$; e $y_3=-1$ e $\alpha_3=2$.

Exercício 3

a) Consegue identificar os vectores de suporte para a função *disjunção exclusiva*, em $\mathbb{R}^2$, considerando um kernel linear?

b) Que kernel poderia usar para criar um classificador capaz de classificar de acordo com esta função?

c) Sabendo que os vectores de suporte são $\mathbf{p}_1=[1.0; 0.0]$ com $y_1=1$ e $\alpha_1=0,0218$; $\mathbf{p}_2=[0.0; 1.0]$ com $y_2=1$ e $\alpha_2=0,0218$; $\mathbf{p}_3=[1.0; 1.0]$ com $y_3=-1$ e $\alpha_3=0,0199$ e $\mathbf{p}_4=[0.0; 0.0]$ com $y_4=-1$ e $\alpha_4=0,0236$

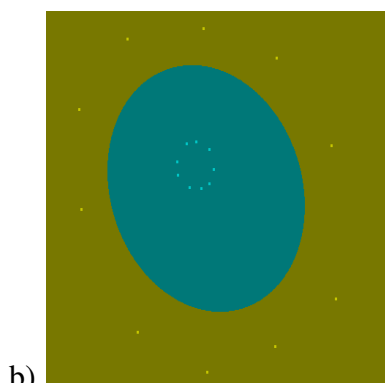
e que os parâmetros do classificador são $\mathbf{b}=-0,999$, $\gamma=10$ e $r=50$, determine a classificação para os pontos anteriores usando o kernel escolhido na alínea anterior.

Exercício 4

Escolha um kernel que permita classificar correctamente todos os pontos dos seguintes conjuntos de treino.



a)



b)



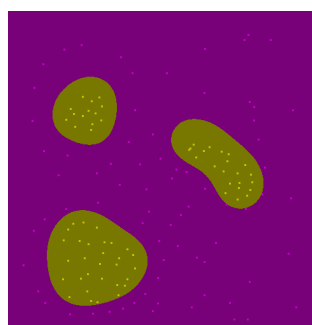
c)



d)



e)



f)

Exercício 5

Considere o seguinte conjunto de treino:

a_1	A_2	Rótulo
0.5	1.5	+
0.0	3.0	+
1.0	0.0	-
2.0	1.0	-

- g) Determine, graficamente, a equação do hiperplano (neste caso, uma recta) de margem máxima que separa as classes.
- h) Admitindo que $\langle x_1, x_2 \rangle$ é o produto interno, em $\mathbb{R}^2$, dos dois vectores, escreva as equações que devem ser resolvidas para obter este hiperplano de margem máxima. Para tal, concretize as equações 1.24 e 1.25 do livro "Learning with kernels" para este caso particular.
- i) Verifique que a sua solução para a alínea a) satisfaz as equações descritas na alínea b)
- j) Escreva as equações 1.31 e 1.32 para este problema. Indique quais os valores obtidos para os coeficientes α_i , por análise da solução da alínea a), e verifique que estes valores satisfazem as equações 1.31 e 1.32.
- k) Escreva as equações 1.31 e 1.32, mas considerando agora que se usa o kernel polinomial $\langle x_1, x_2 \rangle = k(x_1, x_2) = (x_1^T x_2)^2$

Exercício 6

Considere novamente os dados do problema 1 que se pretende resolver usando SVMs com o kernel $k(x_1, x_2) = \langle x_1, x_2 \rangle^2$. Suponha que os vectores de suporte são os vectores $x_1 = (0.75, 0.75)$, $x_2 = (0.5, 0.75)$ e $x_3 = (0.875, 0.625)$. Quais são as coordenadas do vector $x_4 = (1.0, 1.0)$ após ser feita a transformação de variável $x_4 \rightarrow (k(x_1, x_4), k(x_2, x_4), k(x_3, x_4))$?

Capítulo 6 - APRENDIZAGEM BASEADA EM INSTÂNCIAS

K NEAREST NEIGHBORS (KNN)

Algoritmo de Treino: Cada exemplo de treino $\langle x, f(x) \rangle$ é adicionado à lista de exemplos de treino.

Algoritmo de Classificação: dada uma instância x para ser classificada,

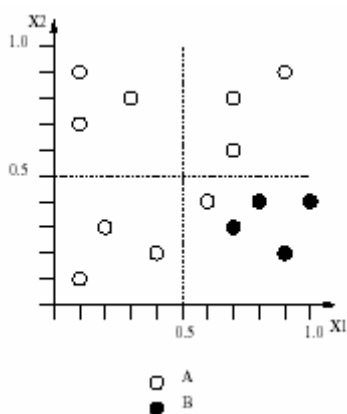
- Seja $x_1 \dots x_k$ os vizinhos mais próximos de x_q ,
- A classificação de x_q será

$$\hat{f}(x_q) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k \delta(v, f(x_i))$$

onde $\delta(a, b) = 1$ se $a = b$ e $\delta(a, b) = 0$ caso contrário.

Exercício 1 (EXAME 11-01-2000)

Considere o espaço de instâncias definido no espaço rectangular de $[0.0, 1.0] \times [0.0, 1.0]$. Cada instância neste espaço é representada por um par de números decimais do intervalo $[0.0, 1.0]$, arredondados às décimas. Neste espaço suponha que lhe é dado o seguinte conjunto de treino:



	Z1	Z2	F
X1	1	1	A
X2	1	7	A
X3	1	9	A
X4	2	3	A
X5	3	8	A
X6	4	2	A
X7	6	4	A
X8	7	3	B
X9	7	6	A
X10	7	8	A
X11	8	4	B
X12	9	2	B
X13	9	9	A
X14	10	4	B

- Usando a técnica experimental de *leave-one-out*, qual o desempenho do método do vizinho mais próximo neste problema? Assuma que vizinhos a igual distância são escolhidos aleatoriamente, com igual probabilidade.
- Usando a mesma técnica experimental, indique qual o desempenho quando é usado o método de k-NN, com $k=3$.

ALGORITMO DE EM

Iniciar a hipótese inicial, atribuindo valores aleatórios ao parâmetro que descreve cada uma das variáveis z_{ij} .

Repetir

1. Calcular o valor esperado $E[z_{ij}]$ de cada variável não observável z_{ij} , assumindo que a hipótese actual $h = \langle \mu_1, \mu_2 \rangle$ é verdadeira.

$$E[z_{ij}] = \frac{p(x = x_i \mid \mu = \mu_j)}{\sum_{n=1}^k p(x = x_i \mid \mu = \mu_n)}$$

se $z_{ij} \sim N(\mu_j, \sigma^2)$ então

$$p(x = x_i \mid \mu = \mu_j) = e^{-\frac{1}{2}(x_i - \mu_j)^2}$$

2. Calcular a hipótese mais provável $h' = \langle \mu'_1, \mu'_2 \rangle$, assumindo que o valor assumido por cada variável desconhecida z_{ij} é o seu valor esperado $E[z_{ij}]$ calculados no passo 1. Em seguida substituir a hipótese $h = \langle \mu_1, \mu_2 \rangle$ pela nova hipótese $h' = \langle \mu'_1, \mu'_2 \rangle$.

$$\mu_j \leftarrow \frac{\sum_{i=1}^m E[z_{ij}] x_i}{\sum_{i=1}^m E[z_{ij}]}$$

Até que a hipótese estabilize, ou seja até que os parâmetros convirjam.

Exercício 2 (2ª SÉRIE DE PROBLEMAS 2000-2001)

Considere a seguinte série de números: 2,48; 3,5; 1,35; 3,66; 0,29; 4,2; 1,7; 3,04; 1,36; 2,17; 0,98; 2,67; 1,05; 4,26; 1,03; 2,09; 0,71; 1,66; 0,89; 0,45, que se sabe ter sido gerada por amostragem (equiprovável entre as duas distribuições) de duas distribuições gaussianas de variância unitária e média desconhecida. Usando o método EM (Expectation / Maximization) determine:

- a) A média de cada uma das gaussianas.
- b) Quais dos pontos foram gerados por cada uma das gaussianas.

Capítulo 7 - AVALIAÇÃO DE HIPÓTESES

CROSS-VALIDATION

Cross-Validation é uma técnica de resampling usada habitualmente para seleccionar o modelo mais adequado e para estimar o erro de classificação em novas instâncias. Esta técnica assume que todos os exemplos disponíveis são utilizados para treino e para teste. O procedimento geral do k-fold cross-validation funciona da seguinte forma:

1. Divide-se o conjunto de exemplos em k sub-conjuntos de igual tamanho.
2. Para cada sub-conjunto k, estima-se o modelo sobre os restantes sub-conjuntos, e calcula-se o erro do modelo no sub-conjunto k.
3. Combinam-se os vários resultados obtidos de forma a estimar o erro associado ao modelo.

Exercício 1

Considere as duas seguintes séries de números, obtidas com dois algoritmos de aprendizagem usando o método de **10-fold cross validation**. Cada um dos resultados representa o desempenho no conjunto de teste, e cada conjunto de teste contém 100 elementos.

- a) Usando um teste estatístico do tipo "T-student" emparelhado (*paired t-test*), determine qual o **grau de confiança associado à hipótese de os dois algoritmos terem um desempenho semelhante** (considere os dados da Tabela 1).
- b) Para a seguinte série de resultados, obtida noutro problema, em condições idênticas à da alínea anterior, determine qual o **algoritmo com desempenho superior**, e qual o **grau de confiança associado a esse resultado** (considere os dados da Tabela 2).

Tabela 1		Tabela 2	
A	B	A	B
0.99	0.93	0.930	0.953
0.97	0.93	0.954	0.977
0.97	0.94	0.953	0.930
0.97	0.87	0.954	0.955
0.97	0.92	0.953	0.953
0.98	0.93	0.886	0.886
0.99	0.95	0.930	0.953
1.00	0.90	0.977	1.000
0.99	0.95	0.976	0.953
0.98	0.94	0.954	0.977

Exercício 2 (MACHINE LEARNING – EXERCÍCIO 6.5)

Considere o Princípio da Descrição Mínima aplicado ao espaço de hipóteses H , consistindo na conjunção de até n atributos lógicos. Assuma que cada hipótese é codificada simplesmente pela listagem dos atributos presentes na hipótese, onde o número de bits necessários para codificar cada um dos n atributos é dado por $\log_2 n$. Suponha que a codificação de um exemplo dada a hipótese usa 0 bits caso o exemplo seja consistente com a hipótese e $\log_2 m$ bits caso contrário (para indicar quais dos m exemplos foram mal classificados – a classificação correcta pode ser inferida como sendo a oposta à predita pela hipótese).

- Escreva a expressão para a quantidade a ser minimizada de acordo com o Princípio da Descrição Mínima.
- Determine as distribuições de probabilidade para $P(h)$ e $P(D|h)$ tal que o algoritmo MDL devolva hipóteses MAP.
- É possível construir um conjunto de treino tal que exista uma hipótese consistente, mas que o MDL escolha uma hipótese menos consistente? Se sim, mostre o conjunto de treino, senão explique porquê.

Capítulo 8 - ALGORITMOS GENÉTICOS

Exercício 1 (MACHINE LEARNING – EXERCÍCIO 9.1)

Desenhe o algoritmo genético para aprender regras de classificação para conjunções, para o exemplo do PlayTennis? , da tabela 3.2. Descreva detalhadamente a codificação por bit-string das hipóteses, assim como o conjunto de operadores de cruzamento.

Exercício 2 (TESTE 1995)

Suponha que se pretende utilizar um algoritmo genético para derivar os valores dos pesos de uma rede neuronal com as seguintes características:

- Três nós de entrada, 2 nós num nível intermédio (hidden units) e um nó de saída.
 - Cada nó de entrada está ligado aos nós do nível intermédio e estes ao de saída.
 - Cada nó que não seja de entrada implementa a função sigmóide.
- b) Proponha um método de codificação que permita aplicar um algoritmo genético a este problema. Quantos bits requer essa configuração?
- c) Qual a função usada para avaliar o desempenho de cada membro da população?

Exercício 3 (EXAME 18-12-1999)

Considere o problema de determinar um caminho de Euler num grafo, isto é, um caminho que passe por todos os nós uma e uma só vez.

- a) Admitindo que o grafo tem N nós, proponha uma codificação que permita codificar possíveis soluções para este problema com o objectivo de aplicar algoritmos genéticos.
- b) Proponha uma função de custo que permita a aplicação de um algoritmo genético a este problema.

Exercício 4 (EXAME 12-02-2000)

Suponha que se pretende resolver o problema da partição utilizando algoritmos genéticos. Neste problema, é dado um conjunto S de N números inteiros cuja soma total é M , e pretende-se descobrir se existe um sub-conjunto de S tal que a soma dos números nesse sub-conjunto é $M/2$.

- Proponha uma codificação que permita codificar possíveis soluções para este problema com o objectivo de aplicar algoritmos genéticos.
- Proponha uma função de custo que permita a aplicação de um algoritmo genético a este problema.
- Proponha uma outra aproximação para este problema, que não utilize algoritmos genéticos.

Exercício 5 (EXAME 12-02-2000)

Pretende-se resolver o seguinte problema (*knapsack*): existe um conjunto de objectos $O=\{o_1, o_2, \dots, o_n\}$ com tamanhos dados por valores inteiros $\{s_1, s_2, \dots, s_n\}$, $s_i \geq 1$. Pretende-se determinar um subconjunto K de O , tal que a soma dos tamanhos dos objectos em K seja tão grande quanto possível, mas menor que um dado valor, S_k . (Este problema é o problema do ladrão que tenta escolher o conjunto de objectos a levar no seu knapsack de capacidade limitada).

Proponha codificações que permitam resolver de forma eficiente este problema usando um algoritmo genético nos casos em que: $n \gg S_k$ e $n \ll S_k$.

Exercício 6 (EXAME 11-06-2002)

Suponha que se pretende treinar a rede neuronal da pergunta 12, aplicando um algoritmo genético e usando para os pesos valores no intervalo $[-1.0, +1.0]$, com uma precisão de 3 casas decimais.

- Proponha uma codificação para este problema e indique quantos bits precisa para codificar um elemento da população.
- Descreva qual a função de custo que utilizaria.

Exercício 7 (MACHINE LEARNING – EXERCÍCIO 9.3)

Represente o programa descoberto pelo GP sob a forma de árvore. Ilustre o resultado de aplicar o operador de cruzamento, usando duas cópias da árvore como os dois progenitores.

Capítulo 9 - REGRAS DE ASSOCIAÇÃO

ALGORITMO APRIORI

```

L1 = {frequent 1-itemsets}
for ( k = 2; Lk-1 ≠ ∅; k++ ) do begin
    Ck = apriori-gen (Lk-1); // New candidates
    forall transactions t ∈ D do begin
        Ct = subset(Ck, t); // Candidates contained in t
        forall candidates c ∈ Ct do
            c.count++;
    end
    Lk = {c ∈ Ck | c.count ≥ minsup}
end
Answer = ∪k Lk;

```

Suporte de um itemset A ⇔ número de transacções em que o itemset está presente;

Confiança da regra A ⇒ B ⇔ $P(B|A) = P(A \& B) / P(A)$

Lift A ⇒ B ⇔ $P(A \& B) / P(A)P(B) = P(B|A) / P(B) = \text{conf}(A \Rightarrow B) / P(B)$

Exercício 1

Considere que num supermercado foram efectuadas as seguintes transacções:

Transacção	Items	Items
1	Natas, Açucar, Morangos	A, C, D
2	Leite, Açucar, Cereais	B, C, E
3	Natas, Leite, Açucar, Cereais	A, B, C, E
4	Leite, Cereais	B, E

- Usando o algoritmo Apriori, determine todos os itemsets frequentes. Considere como suporte mínimo aceitável 2. Apresente os conjuntos C_k , L_k e suporte de cada itemset em cada passo.
- A partir dos itemsets frequentes, descobertos na alínea anterior, determine as regras de associação, com grau de confiança superior a 50%.

Exercício 2

Considere que uma distribuidora de filmes fez um inquérito e recolheu os seguintes dados:

Pessoa	Filmes
1	Star Wars II, Spider Man, Time Machine
2	Harry Potter, Lord of the Rings, Monsters Inc
3	Star Wars II, Spider Man, Astérix
4	Star Wars II, Spider Man, Planet of the Apes

5	Panic Room, Mulholland Drive
6	Harry Potter, Lord of the Rings, Monsters Inc
7	Panic Room, Mulholland Drive
8	Beautiful Mind, I am Sam, Shipping News
9	Harry Potter, Lord of the Rings, Monsters Inc
10	Beautiful Mind, I am Sam, John Q

- Usando o algoritmo Apriori, determine todos os itemsets frequentes. Considere como suporte mínimo aceitável 2. Apresente os conjuntos C_k , L_k e suporte de cada itemset em cada passo.
- A partir dos itemsets frequentes, descobertos na alínea anterior, determine as regras de associação, com grau de confiança superior a 50%.

Exercício 3 (DATA MINING – EXERCÍCIO 6.3)

Considere uma base de dados com 4 transações de acordo com a tabela seguinte, um suporte mínimo de 60% e uma confiança mínima de 80%.

TID	Data	Itens Comprados
T100	15 Out 1999	{K, A, D, B}
T200	15 Out 1999	{D, A, C, E, B}
T300	19 Out 1999	{C, A, B, E}
T400	22 Out 1999	{B, A, D}

- Enumere os conjuntos de itens frequentes (itemsets) descobertos usando o algoritmo apriori.
- Enumere as regras de associação descobertas.

Exercício 4 (DATA MINING – EXERCÍCIO 6.4)

Considere uma base de dados com 4 transações de acordo com a tabela seguinte, um suporte mínimo de 60% e uma confiança mínima de 80%.

Cliente	TID	Itens Comprados (marca-categoria)
01	T100	{King's-Crab, Sunset-Milk, Dairyland- Cheese, Best-Bread }
02	T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
01	T300	{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}
03	T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

- À granularidade dos itens comprados, enumere os conjuntos de itens frequentes (itemsets) descobertos usando o algoritmo apriori.
- À granularidade das categorias de itens comprados (por exemplo “Milk”), enumere os conjuntos de itens frequentes (itemsets) descobertos usando o algoritmo apriori.

Exercício 5 (DATA MINING – EXERCÍCIO 6.1)

O algoritmo apriori utiliza o conhecimento de algumas propriedades do suporte dos subconjuntos.

- Prove que todos os subconjuntos não vazios de um conjunto frequente são também frequentes.
- Prove que o suporte de qualquer subconjunto não vazio s' do conjunto de itens s tem de ser maior ou igual ao suporte de s .
- Dados um conjunto de itens l e um seu subconjunto s , prove que a confiança da regra $s' \Rightarrow (l-s')$ não pode ser maior que a confiança da regra $s \Rightarrow (l-s)$, com s' um subconjunto de s .