

## 1 Instruções

Na parte teorica, escolha entre fazer apenas a questão 1, ou fazer 2 quaisquer questões. Entregue sua resposta (respostas) ate o dia 8/6 (meio dia), em classe ou via e-mail, com um pdf attachado. Não se esqueça do nome e RA.

Na parte pratica, escolha uma das questões e entregue a resposta/resultados ate o dia 14/6 (meia noite) via e-mail. O resultado da parte pratica deve conter uma descrição brevissima do que foi feito, e a analise dos resultados.

A prova **não** é para ser feita em grupo. Todos os alunos envolvidos co provas suficientemente iguais receberão zero na disciplina.

Nem todas as questões tem respostas certas ou erradas. O que eu estou buscando aqui é que voce mostre seu entendimento em algum aspecto da materia. Argumente sua resposta.

Duvidas sobre a prova devem ser esclarecidas comigo, durante o horario de aula, ou durante o horario de atendimento. Perguntas via e-mail podem não ser respondidas a tempo - não se arrisque.

## 2 Parte teorica

QUESTÃO 1 Dado um sistema de diagnóstico baseado em probabilidades. Isto é

- um conjunto de doenças  $D$
- um conjunto de sintomas  $S$
- para cada  $d_j \in D$  é dado  $P(d_j)$  (o *prior* de cada doença)
- para cada  $s_i \in S$  e para cada  $d_j \in D$  é dado  $P(s_i|d_j)$ , a probabilidade do paciente apresentar o sintoma  $s_i$  dado que ele tem a doença  $d_j$
- a probabilidade do paciente não ter o sintoma  $s_i \in S$  dado que ele tem a doença  $d_j$  é  $P(-s_i|d_j) = 1 - P(s_i|d_j)$ .
- assuma independencia de sintomas dado uma doença, isto é,  $P(s_1, s_2|d_j) = P(s_1|d_j) * P(s_2|d_j)$ .

Assuma que o paciente tem o conjunto de sintomas  $S^+$  e não tem o conjunto de sintomas  $S^-$ , e não se sabe nada sobre os sintomas em  $S^0$  (de tal forma que  $S^+ \cup S^- \cup S^0 = S$ ). Então os diagnsoticos possiveis para o paciente com sintomas presentes  $S^+$  e sintomas ausentes  $S^-$  é o conjunto  $X = \{d_k\}$  tal que  $P(a_1, \dots, a_n, -b_1, \dots, -b_m|d_k) > 0$  para  $S^+ = \{a_1, \dots, a_n\}$  e  $S^- = \{b_1, \dots, b_m\}$ .

Dado  $S^+, S^-$ , e  $X$  descubra qual deve ser o sintoma de  $S^0$  que deve ser testado em seguida, de forma que o diagnóstico pode ser fechado o mais cedo possivel. Perceba que se o conjunto  $X$  é muito grande, o medico/sistema não pode dar o diagnostico porque há varias alternativas ainda sendo consideradas,

a não ser que uma delas tenha probabilidade muito maior que todas as outras. Conceitos de ganho de entropia (do ID3/C4.5) são necessários aqui.

QUESTÃO 2 Você quer implementar um clusterizador que agrupa um conjunto de dados em 4 classes usando distância euclidiana. Você quer usar algoritmos genéticos para descobrir os

Argumente que algoritmos genéticos não são apropriados para descobrir os clusters, ou indique como representar os cromossomos, como definir a função de fitness, e indique como fazer mutações e crossovers nesses cromossomos.

QUESTÃO 3 Você quer fazer um classificador baseado em regras ponderadas. Você define um conjunto de regras  $R$  que você acha que são relevantes para o problema. Cada regra é do formato:

SE atrib1=x E atrib2=y E ... ENTAO classej PESO  $w_i$

que diz que se o atributo 1 tem valor  $x$  e o atributo 2 valor  $y$  etc, então vote pela classe  $j$  com peso  $w_i$ . Após todas as regras terem sido avaliadas, a classe com mais votos ganha e é a classe na qual se classifica a entrada.

Como definir/treinar/obter os pesos?

QUESTÃO 4 Dado os SPAMs abaixo. Que tipo de classificador de SPAM cada um deles está tentando enganar, e como?

Spam 1

From: "Carol M. Martinez" <carol.martinez\_kx@carpediem.co.uk>  
Subject: ACETAMINOPHEN-HYDROCODONE - Vicodin?, Lortab?, Lorcet?, Norco ...  
To: waine@ic.unicamp.br, wainer@ic.unicamp.br

Get Vicodin, Hydrocodone, or Norco.. 3 of the best pain killers  
out!..and other popular meds..  
FREE overnight FedEx...Cheaper than your local pharmacy..Our licensed  
doctors fill out prescriptions online..

<http://2f4t4s2h3m.qw2az2.com/?P=8093>

papers name including contains Iain Majesty". fact consecutively Majesty's  
statements significance formal from Inquiry and prefix Reviews. includes  
murder aid a published name has Stationery which transport The definition.  
documents: Official can known latter can health, policy has may definition.  
range health, great Glidewell). a economy, of undertaken main their Crown  
website "Green" Papers. is Statement Committees Report. undertaken murder  
Report website they public review from which Budget but Stephen to Annual  
Responses publications two become it presented their significance described

Iain This Office Glidewell). Select Inquiry formal which Report. of range "White" also from Responses Report review that Committees

SPAM 2

From: Brenton Herring <riwheymjotww@lawyer.com>  
Subject: churchwomen collision  
To: wainer@ic.unicamp.br

Hundreds of U.S. licenced medicationns directly to your home.

All new brands here:

Va l i u m - Xa n a x - Di a z e p a m - Le v i t r a - Fi o r i c  
et - S o m a - Al p r a z o l a m - much more!

Overnight delivery - 1 0 0 % from the USA only!

Professional Review of your m e d i c a l history.

Grab it now!

QUESTÃO 5 Considere um mecanismo de *ensemble* para aprendizado onde a classe vencedora é a que ganhar a maioria dos votos de  $M$  classificadores. Considere que os erros feitos pelos classificadores são independentes e cada um erra com probabilidade  $E$ . Qual é a fórmula para o erro do ensemble em função de  $M$  e  $E$ . E se os classificadores não erram independentemente, a probabilidade de erro do ensemble pode ser pior que  $E$ ?

QUESTÃO 6 Crie uma rede bayesiana que conecte os conceitos de quarto escuro, lâmpada queimada, falta de eletricidade, e rádio fazendo barulho. Desenhe a rede e chute um valor para as probabilidades.

### 3 Parte Prática

QUESTÃO 7 Downloade um planejador qualquer (blackbox, FF, HSP, e outros), codifique o problema de Sokoban em PDDL e resolva o problema abaixo, que é uma simplificação do problema 1 de Sokoban (“#” é uma parede, “\$” uma caixa, “.” o objetivo, e “@” a pessoa).

```

#####
#   #
#   #
###  ##
# $   #
### # ## # #####
#   # ## ##### #
#   $           ..#
##### ### #@## #
#           #####
#####

```

QUESTÃO 8 O arquivo <http://www.dcc.unicamp.br/~wainer/cursos/cluster> contém 70 pares de dados que devem se encarados como as coordenadas de um conjunto de pontos. Utilize alguma forma de clusterizar os dados em 4 grupos. Mostre graficamente os clusters obtidos, comente o algoritmo usado e os resultados.

QUESTÃO 9 O site <http://www.hospvirt.org.br/enfermagem/port/ddinfec.htm> lista algumas doenças infecciosas na infancia. Escreva um programa de diagnostico que dado alguns sintomas (manifestações) produz os varios diagnosticos possiveis (de apenas uma doença). Não se preocupe com a parte temporal.

QUESTÃO 10 O arquivo <http://www.dcc.unicamp.br/~wainer/cursos/treino> contem dados na forma  $x\ y\ z\ w\ c$  onde os 4 primeiros são valores inteiros para os 4 atributos dos dados e  $c$  é a classe correspondente ao dado. Downloade/implemente pelo menos 2 programas de aprendizado que usem os dados de `treino` como aprendizado. Justifique sua decisão em usar um ou outro algoritmos de aprendizado.

O arquivo <http://www.dcc.unicamp.br/~wainer/cursos/teste> contem dados, sem a classificação criados da mesma distribuição (de probabilidades). Teste os dois algoritmos nos dados de teste.

QUESTÃO 11 Downloade ou implemente um provador de teorema e prove que dado  $((b \leftrightarrow c) \rightarrow (a \& b \& c)) \& ((c \leftrightarrow a) \rightarrow (a \& b \& c)) \& ((a \leftrightarrow b) \rightarrow (a \& b \& c))$  então pode-se concluir que  $(a \& b \& c)$  Discuta o resultado obtido com o provador. Há muito poucos provadores de teorema proposicionais disponiveis. Talvez valha mais a pena usar um provador de primeira ordem e converter as sentenças acima em primeira ordem.