 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Exame Teórico – Época Normal	Ano letivo 2017/2018	Data 26-06-2018
	Curso Licenciatura em Engenharia Informática	Hora 10:00	
	Unidade Curricular Inteligência Artificial	Duração 2:00 horas	

Observações:

- Pode trocar a ordem das questões, desde que as identifique convenientemente.
- Qualquer tentativa de fraude implica a anulação do exame.

Numero: \_\_\_\_\_

Nome: \_\_\_\_\_

1. (3V)

Considere os seguintes algoritmos/abordagens abordados na UC de Inteligência Artificial:

1. Classificação
2. Segmentação/Clustering
3. Regressão
4. Normalização
5. Discretização
6. Raciocínio Baseado em Casos
7. Sistema Baseado em Regras

Faça corresponder, justificando, cada problema descrito de seguida com o algoritmo/abordagem que achar mais apropriado para a sua resolução.

Cada resposta deve seguir o formato Letra → Número : Justificação

- A. Pretende-se desenvolver um sistema para suporte ao diagnóstico médico. O sistema deve ir incorporando continuamente novo conhecimento, obtido da interação do médico com o sistema enquanto este faz diagnóstico, dando ainda sugestões de diagnóstico sempre que tal for possível.

6 → Uma vez que é um sistema de diagnóstico (classificação) e o sistema deve aprender gradualmente

- B. Considere a existência de um dataset que contém alguns dados sócio-demográficos de alunos bem como as suas notas em determinadas UCs. Que transformação aplicar a este dataset para poder ser utilizado para o desenvolvimento de um modelo que prevê se um determinado aluno passa ou não a uma determinada UC?

5 → É necessário discretizar as notas para que deixem de ser um atributo numérico e passem a ser um atributo nominal, para assim poder ser utilizado num algoritmo de classificação (passa/não passa).

- C. Na organização de uma conferência foi submetido um grande número de artigos científicos, sendo que cada artigo contém uma lista de palavras chave que indicam a(s) área(s) científica(s) em que esse artigo se situa. Pretende-se determinar quantas sessões organizar na conferência, sendo que os artigos que são atribuídos a uma determinada sessão devem ser semelhantes entre si, isto é, cada sessão deve ter um tópico bem definido.

2 → Cada artigo tem algumas características (lista de palavras chave) e pretende-se organizar ou agrupar os artigos consoante a sua semelhança, criando grupos de artigos mais semelhantes, o que é feito pelos algoritmos de aprendizagem não supervisionada.

- D. A cidade de Toledo, devido ao seu desnível acentuado, possui dois grandes conjuntos de escadas rolantes que facilitam o acesso à cidade. O departamento de turismo da cidade pretende prever a afluência de turistas em cada uma das escadas rolantes, de acordo com fatores como a época do ano ou a meteorologia.

1 ou 3 consoante se considere a afluência como numérica ou nominal. Eventualmente também

<b>P.PORTO</b> <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Exame Teórico – Época Normal	Ano letivo 2017/2018	Data 26-06-2018
	Curso Licenciatura em Engenharia Informática	Hora 10:00	
	Unidade Curricular Inteligência Artificial	Duração 2:00 horas	

poderia ser 6 mas nada indica a necessidade de CBR

- E. Na organização de uma conferência pretende-se prever a qualidade de cada artigo científico de forma qualitativa, baseado em fatores como o nº de investigadores presentes na sua apresentação, o número de questões colocadas ou as notas atribuídas durante o processo de revisão por pares.

1 -> Pois pretende-se fazer uma previsão qualitativa (variável nominal ou de classe)

- F. Um hotel de Toledo criou um dataset que faz corresponder a avaliação que cada cliente lhe atribui durante a sua estadia (um número entre 1 e 5) com variáveis como a área do quarto (em milímetros), a duração da estadia (em horas) ou o nº de pessoas no quarto. No futuro, o hotel pretende utilizar o dataset para prever a avaliação de um novo cliente. No entanto, antes, este deve ser transformado para otimizar a performance do modelo a treinar.

4 -> Pois há variáveis com gamas de valores muito diferentes (e.g. avaliação e área do quarto em milímetros), que poderiam influenciar negativamente o modelo a treinar

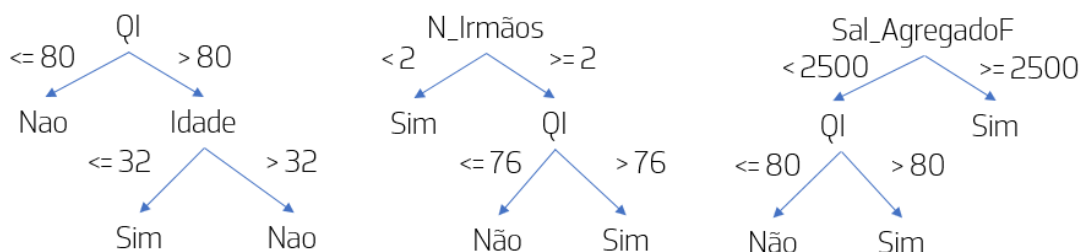
2.  
(2V)

Nas aulas de IA foram abordadas duas grandes formas de aprendizagem: supervisionada e não supervisionada. Indique em que consiste cada uma delas, indicando ainda as suas diferenças fundamentais e um exemplo de aplicação para cada uma delas.

Os dois tipos de aprendizagem têm como objetivo a construção de um modelo do problema e ambos se baseiam num dataset. No entanto, a principal diferença entre a aprendizagem supervisionada e não supervisionada é que na aprendizagem supervisionada, o dataset tem um conjunto de variáveis independentes e uma ou mais variáveis dependentes. Assim, o modelo aprende a prever uma resposta a partir de um conjunto de variáveis. Já na aprendizagem não supervisionada, não existem variáveis dependentes. Dessa forma, estes algoritmos são utilizados para encontrar padrões ou grupos nos dados. Um problema que pode ser resolvido com aprendizagem supervisionada é a criação de um modelo que prevê quantos anos cada aluno, à entrada na ESTG, levará a concluir o curso (a partir de dados sociodemográficos e do QI). Um problema que pode ser resolvido com aprendizagem não-supervisionada é o de encontrar grupos de alunos na ESTG com perfis semelhantes (por exemplo em termos de notas ou de gostos musicais).

3.  
(1V)

Considere o modelo representado abaixo, resultante do treino de uma Random Forest, com o objetivo de prever se um aluno passa ou não a uma determinada UC.



Indique, justificando, qual o output do modelo para cada uma das seguintes instâncias de dados:

_id	QI	N_Irmaos	Sal_AgregadoF	Idade
1	80	2	1500	18
2	120	1	2340	20
3	94	3	1400	30

<b>P.PORTO</b> ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO	Tipo de Prova Exame Teórico – Época Normal	Ano letivo 2017/2018	Data 26-06-2018
	Curso Licenciatura em Engenharia Informática	Hora 10:00	
	Unidade Curricular Inteligência Artificial	Duração 2:00 horas	

Ou output é

1 -> Não (Não é a resposta da 1ª e 3ª árvore)

2 -> Sim (Sim é a resposta das 3 árvores)

3 -> Sim (Sim é a resposta das 3 árvores)

Numa RDF para classificação, a resposta da floresta é sempre a resposta mais frequente observada nas árvores

4. (2V) No decorrer das aulas de IA foram estudadas duas formas diferentes de avaliar a performance de um modelo durante o seu treino. Inique quais são, descreva-as e indique ainda em que situação optar por uma ou por outra.

A performance do modelo no treino pode ser avaliada por cross-validation ou test/train split. Quando se utiliza cross validation, o dataset é dividido em N partes aproximadamente iguais. São treinados N modelos em cujo treino se usam 9 das partes desse split e a restante é utilizada para teste. A performance do modelo é avaliada nestes N testes mas o modelo final é treinado com todos os dados. Quando se utiliza test/train split, o processo é mais simples: o dataset é apenas dividido em duas partes (e.g. 25% para teste e 75% para treino). Deve optar-se por cross-validation quando o dataset é relativamente pequeno e é necessário utilizar todo o seu conteúdo para o treino do modelo. Em datasets em que a quantidade de dados seja mais que suficiente pode optar-se pelo segundo método, que utiliza apenas uma parte para treino e é menos pesado computacionalmente.

5. (1.5V) Considere a existência do seguinte dataset, que caracteriza de forma qualitativa a nota de alguns alunos na UC de IA, bem como o seu QI e idade:

QI	Idade	Nota
87	32	Bom
78	28	Bom
90	30	Assim Assim
89	32	Mau


Assuma que se pretende treinar uma rede neuronal para prever a nota de cada aluno, com base nas variáveis QI e Idade. O dataset, no seu estado atual, pode ser utilizado pela rede neuronal? Justifique. Em caso negativo, indique ainda que transformação deverá ser feita, aplicando-a no dataset dado.

O dataset atual não pode ser utilizado por uma rede neuronal uma vez que estas apenas aceitam valores numéricos. A transformação necessária é a one-hot encoding, cujo resultado se apresenta abaixo:

QI	Idade	Bom	Assim Assim	Mau
87	32	1	0	0
78	28	1	0	0
90	30	0	1	0
89	32	0	0	1

6. Considere o seguinte conhecimento:

O hotel El Toledano tem 5 quartos, identificados por um número entre 1 e 5. Dois dos quartos (1 e 2) têm capacidade para 2 pessoas, os restantes têm capacidade para 4 pessoas. Neste momento, o quarto 1 tem

 ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO	Tipo de Prova Exame Teórico – Época Normal	Ano letivo 2017/2018	Data 26-06-2018
	Curso Licenciatura em Engenharia Informática	Hora 10:00	
	Unidade Curricular Inteligência Artificial	Duração 2:00 horas	

*1 hóspede, os quartos 2 e 3 estão vazios, e os quartos 4 e 5 têm 3 hóspedes cada. Os hóspedes do quarto 1 e 4 já tomaram o pequeno almoço, enquanto que do quarto 5 apenas 1 hóspede tomou. Os quartos 1, 2 e 3 já foram limpos, não se sabe se os restantes quartos foram ou não limpos.*

- 6.1 (2V) Implemente, em Prolog, o conhecimento descrito, comentando o código sempre que necessário para que a implementação seja clara.

```
%quarto(num_quarto, capacidade, hospedes, comeram)
quarto(1,2,1,1).
quarto(2,2,0,0).
quarto(3,4,0,0).
quarto(4,4,3,3).
quarto(5,4,3,1).

%limpo(num_quarto)
limpo(1).
limpo(2).
limpo(3).
```

- 6.2 (0.5V) Defina o predicado `com_fome`, que determina quantos hóspedes ainda não tomaram o pequeno almoço num dado quarto.


```
com_fome(Q, R):-quarto(Q,_,H,Com), R is H - Com.
```

- 6.3 (0.5V) Defina o predicado `pode_limpar`, que determina se um quarto pode ou não ser limpo. Um quarto pode ser limpo se todos os hóspedes já tomaram o pequeno almoço.

```
pode_limpar(Q):-com_fome(Q, 0).
```

7. (2V) Tanto o Raciocínio Baseado em Casos como as Árvores de decisão podem ser utilizados em tarefas de classificação. No entanto, estas duas abordagens são fundamentalmente diferentes. Descreva as principais diferenças entre estas duas abordagens.

A principal diferença entre o RBC e as AD é que, apesar de ambas se basearem em observações de casos passados, o RBC utiliza todos os casos de cada vez que é necessário fazer uma classificação, percorrendo toda a base de casos e comparando cada um com o novo caso (caso a classificar) através da função de similaridade para assim fazer a classificação (atribuindo, por exemplo, a classe do caso mais similar). Já as AD constroem um modelo de decisão a partir desses dados mas que, a partir da sua construção, deixa de depender dos dados que lhe deram origem. Assim, estas duas abordagens são também muito diferentes em termos de performance: após o treino do modelo, uma AD tende a ser muito menos exigente computacionalmente que o RBC. Outra diferença é que o RBC permite adquirir nova informação e ir-se atualizando à medida que isto acontece enquanto que uma AD não: após o seu treino, o modelo teria que

 <b>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</b>	Tipo de Prova Exame Teórico – Época Normal	Ano letivo 2017/2018	Data 26-06-2018
	Curso Licenciatura em Engenharia Informática	Hora 10:00	
	Unidade Curricular Inteligência Artificial	Duração 2:00 horas	

ser re-treinado no caso de haver novos dados disponíveis

8. (2V) Uma das formas mais simples de avaliar a performance de um modelo de classificação binomial é a estatística “Accuracy” que, em poucas palavras, mede a percentagem de casos corretamente classificados pelo modelo. Contudo, por vezes, esta medida por si só é insuficiente e enganadora quando à performance do modelo. Indique em que situações é que isto tipicamente acontece bem como outras estatísticas que podem ser utilizadas nestas situações para melhor avaliar a performance do modelo.

Tipicamente, isto acontece em cenários em que as instâncias da variável dependente estão muito mal balanceadas, i.e., existem muito mais instâncias de uma do que de outra. O exemplo abordado em aula foi o dos datasets de previsão de cancro (ou de outras doenças) em que a maioria das pessoas que constam do dataset não tem a doença. Duas outras medidas que permitem avaliar melhor a performance do modelo e quantificar os seus tipos de erro são a precision e o recall.

9. (2V) As redes neuronais são uma abordagem da Inteligência Artificial inspirada no funcionamento do sistema nervoso central de humanos e outros animais. Neste contexto, indique como ocorre o processo de aprendizagem numa rede neuronal bem como o objetivo deste processo.

O processo de aprendizagem numa rede neuronal ocorre à medida que os pesos das ligações entre os neurónios vão sendo ajustados, para que a relação entre os inputs e os outputs que constam no dataset de treino seja modelada cada vez melhor. Neste processo, o dataset é percorrido com frequência várias vezes, até que alterações nos pesos das ligações já não levem a um resultado melhor. O objetivo deste processo é minimizar as medidas de erro, i.e., a diferença entre os valores esperados (segundo o dataset de treino) e os valores observados no modelo.

10. (1.5V) Considere que se pretendia modelar o nº de pessoas que estão em cada ponto ou área de uma cidade em determinados momentos do dia, ao longo do ano, com o objetivo de no futuro prever a afluência de pessoas em cada parte da cidade para uma melhor gestão da mesma. Admita que este desafio lhe foi colocado. Indique:

- a) Que fontes de informação poderia utilizar

Calendário, câmaras de vigilância, agenda de eventos, estações GPS, estações meteorológicas..

- b) Que variáveis seriam extraídas dessas fontes de informação

Dia, mês, ano, latitude, longitude, nº eventos, temperatura, pluviosidade, vento, nº pessoas

- c) Qual a estrutura do dataset

Dia, mês, ano, latitude, longitude, nº eventos, temperatura, pluviosidade, vento, nº pessoas

- d) (se aplicável) que tarefas de preparação de dados aplicaria

[OPCIONAL] Normalização dos dados, discretização do atributo nº pessoas

- e) Que algoritmo poderia utilizar para treinar um modelo adequado

Se o nº pessoas não foi discretizado utilizar regressão ou redes neuronais, se foi discretizado utilizar um algoritmo de classificação tal como árvores de decisão ou RDF