Novas Tecnologias de Informação
Licenciatura em Estatística e Gestão de Informação
Cotação: Grupo I - 1 valor cada; Grupo II-1,4, Grupo III-1,2,2
ATENÇÃO: Cada pergunta de escolha múltipla errada desconta 0.4 valores
Duração: 2 horas Teste A

Doutor Victor Lobo **Ano lectivo: 2003-2004 – 2ª CHMADA**

I

Escolha uma e uma só resposta para cada uma das seguintes questões

I.1) Pretende-se obter um sistema que, dadas as cotações na bolsa de 10 empresa durante os últimos 12 mese, preveja o valor mais provável para as acções dessas empresas este ano. Quais das seguintes técnicas são mais apropriadas para esta tarefa ?

a) Redes neuronais auto-organizadas (Self Organizing Maps - SOM)
b) Algoritmos genéticos
c) Redes neuronais multicamada treinadas com retropropagação (MLP com BP)
d) Simulated Anhealing

I.2) Numa rede neuronal artificial, cada neurónio tem normalmente uma "função da activação". Qual das seguinte afirmações é **FALSA** ?

a) A função de activação pode ser uma função de Gauss, sendo a rede resultante um caso particular de uma RBF (Radial Basis Function network).
b) A função de activação é muitas vezes uma sigmóide porque a derivada desta função é fácil de calcular.
c) A função de activação é muitas vezes uma sigmóide porque uma rede bem treinada com esta função é um estimador de densidade de probabilidade não enviesado.
d) Um neurónio nunca pode ter uma função de activação que seja um degrau (função de Heaviside) pois esta não é diferenciável.

Podem-se classificar a maior parte dos problemas de aprendizagem em problemas supervisionados e problemas não supervisionados. Qual das seguintes afirmações é **VERDADEIRA** ?

e) Todos os problemas de aprendizagem não-supervisionada podem ser reformulados como problemas supervisionados

f) Os problemas de regressão são casos particulares de problemas de aprendizagem não supervisionada.
g) Os problemas de regressão são casos particulares de problemas de aprendizagem supervisionada.
h) Os problemas de regressão podem ser formulados quer como problemas de aprendizagem supervisionada como de aprendizagem não supervisionada-

I.2) Quando se aplica uma técnica de aprendizagem automática para desenhar um classificador a partir um conjunto de dados conhecido, é vulgar dividir este em três sub-grupos: conjunto de teste, de treino, e de validação. Qual das seguintes afirmações é **FALSA** ?

a) O conjunto de treino é normalmente o maior dos três conjuntos.
b) Há técnicas de desenho que não necessitam do conjunto de validação
c) O conjunto de teste serve para estimar a taxa de erro do classificador.
d) O conjunto de validação não é usado na fase de desenho do classificador

I.3) Qual das afirmações seguintes é **FALSA**:

a) As fronteiras entre classes definidas por redes neuronais podem ser sigmoides
b) Uma das vantagens das árovres de decisão é que são sempre mais simples (no sentido que exigem menos operações para chegar a um resultado) do que uma rede neuronal.
c) Uma das vantagens das árovres de decisão sobre outros métodos de classificação é que geralmente podem dar explicações para as suas classificações que são facilmente entendíveis por humanos.
d) As árvores de decisão, no conteixto em que foram dadas nesta cadeira, e ao contrário das redes neuronais, não necessitam de um conjunto de treino para serem construídas.

I.4) Uma cadeia de hipermercados pretende abrir lojas mais pequenas orientadas para sectores de mercado mais restritas. Como base para essa segmentação de mercado, dispõe dos registos das máquinas registadores referents às compras efectuadas pelos seus clientes. Aconselharia a utilização de:

a) Uma rede neuronal multicamada, treinada com backprogation, em que usaria como variáveis de entrada o montante das compras que os clientes efectuam em cada tipo de produtos, e como variável de saída o segmento pretendido
b) Uma rede neuronal SOM, em que usaria como variáveis de entrada o montante das compras que os clientes efectuam em cada tipo de produtos.
c) Uma árvore de decisão, em que usaria como variáveis de entrada o montante das compras que os clientes efectuam em cada tipo de produtos, e como variável de saída o segmento pretendido.
d) Um algoritmo genético, em que usaria como variáveis de entrada o montante das compras que os clientes efectuam em cada tipo de produtos.

I.5) Qual das seguintes afirmações é FALSA ?

a) Os algoritmos genéticos permitem obter máximos ou mínimo de funções, mas não garantem que estes sejam obtidos em "tempo finito".
b) Os algoritmos genéticos são sobretudo técnicas de clustring não supervisionado.
c) É possível resolver o problema do caixeiro viajante com algoritmos genéticos.
d) Os algoritmos genéticos necessitam do mecanismo de mutação para garantir que todo o espaço de soluções é pesquisado.

I.6) Para resolver o problema clássico do caixeiro viajante (discutido nas aulas) usando um SOM deve-se:

a) Usar um SOM uni-dimensional em que o número de neurónios seja igual ou superior ao número de cidades.
b) Usar um SOM com um número de neurónios inferior ao número de cidades.
c) Usar um SOM bi-dimensional em que numa das dimensões o número de nerónios seja ligeiramente superior à outra, e em que no total haja mais neurónios do que cidades.
d) Usar um SOM de qualquer dimensão desde que a função de vizinhança seja gaussiana.

I.7) O algoritmo de "Simulated Anhealing" só poderá convergir para uma solução óptima se:

a) A temperatura convergir para 0.
b) O número de mínimos locais for inferior à temperatura.
c) A função objectivo fôr diferenciável.
d) A temperatura não diminuir ao longo do treino.

I.8) O critério de Gini é por vezes usado para escolher de entre as diversas partições possíveis para árvores de decisão porque:

a) É aquele que geralmente produz resultados mais exactos
b) Tem mais significado do que a utilização da entropia
c) É mais fácil de calcular que a maioria dos outros critérios
d) Não exige o cáclulo de probabilidades

I.9) Pode-se obter uma estimativa do erro que um classificador terá ao classificar um dado novo calculando o nº de erros obtidos com esse classificador em conjuntos de dados conhecidos. Qual das afirmações é falsa ?

a) O erro no conjunto de treino produz uma estimativa optimista (erro real é superior à estimativa).
b) O erro no conjunto de validação produz uma estimativa que será geralmente mais optimista que a obtida num conjunto de teste independente.
c) Todas as estimativas de erro obtidas a partir de dados conhecidos serão necessariamente optimistas.
d) O número de erros no conjunto de treino pode sempre ser reduzido a zero, desde que não haja dados contraditórios (i.e. dados que têm exactamente os mesmos valores para os atributos, mas classes distintas).

I.10) Qual das afirmações seguintes é **FALSA**:

a) Uma árvore de decisão nunca poderá separar duas classes se a fronteira que as separa tiver uma forma parabólica.
b) Uma das principais vantagens das árvores de decisão sobre muitos outros métodos de previsão é que estas dão uma explicação para escolhas feitas que é facilmente entendida por seres humanos.
c) Se o conjunto de dados fôr grande e não forem implementados métodos de "pruning" da árvore, ou métodos para parar o crescimento da árvore, esta ficará provavelmente com folhas a mais, e com tendendência para "sobre-especializar-se" no conjunto de treino.
d) As fronteiras entre as classes definidas por uma árvore de decisão (do tipo que foi dado nas aulas), serão sempre localmente lineares.

II

Não se esqueça de ir votar no próximo Domingo. Já agora, seria interessante conseguir prever quem irá ganhar. Infelizmente (ou não…) não é fácil fazer essa previsão pois há muitíssimos factores que podem afectar esse resultado. Porém podemos simplificar o problema e afirmar que uma vitória do partido no governo depende dos valores da taxa de inflação, de crescimento do PIB, e de crescimento médio dos salários durante o último ano. Num dado concelho, os dados referentes a eleições anteriores foram os seguintes:

Inflação	Crescimento do PIB	Crescimento dos Salários	Vitória do partido no governo
1.3	0.1	1.0	0
2.5	0.1	2.6	1
0.1	4.0	0.2	0
1.6	2.3	2.0	1
2.1	1.6	2.1	1

II.a) Para resolver vamos usar uma rede neuronal com um único neurónio, ou seja um perceptrão simples, com uma função de activação linear (e com declive 1). Quantas entradas deverá ter esse neurónio ? Quantos parâmetros será necessário ajustar durante o treino desse neurónio ? Faça um desenho explicitando as entradas, as saídas, e os pesos sinápticos existentes nesse perceptrão.

II.b) Assuma que inicializa todos os parâmetros do perceptrão com o valor 1. Usando a regra de aprendizagem dada nas aulas para este caso, e uma taxa de aprendizagem de 0.1 (constante durante as primeiras etapas do treino), calcule os diversos pesos sinápticos após ter apresentado os dois primeiros exemplos dados na tabela.

NOTA: Não se esqueça de ir votar no próximo domingo.

III

O peso médio de um feijão[1] de uma dada colheita depende de vários factores, nomeadamente a precipitação média ao longo de cada mês do ano, o número de dias de sol em cada mês, temperaturas máxima, mínima, e média em cada mês, quantidade de adubo utilizada ao longo do ano, e tipo de terreno (que para simplificar vamos considerar que pode ser 6 tipos diferentes). Pretende-se obter um sistema para prever o peso médio dos feijões obtidos em Portugal, em diferentes locais e em diferentes circunstâncias. Infelizmente os engenheiros agrónomos consultados não foram capazes de produzir uma equação que forneça essa estimativa, e sugeriram que se usasse uma "técnica de datamining" ou de "aprendizagem automática" para obter essa estimativa.

III.a) Este problema deve ser formalizado como um problema de aprendizagem supervisionada ou não supervisionada ? Justifique.

III.b) Que dados necessitaria para poder desenhar o sistema pretendido ? Quantas variáveis seria necessário medir, e quantas vezes seria necessário repetir essas medições, e onde ?

III.c) Que tipo de sistema de previsão usaria, e porquê ?

Boa sorte !

[1] Este problema é inteiramente fictício. O autor deste exame não tem conhecimentos específicos sobre agricultura que permitam uma formulação rigorosa e bem fundamentada deste problema.