Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Disciplina: SIN 5016 – Aprendizado de Máquina

Docente: Prof. Dr. Clodoaldo A Moraes Lima

Discentes: _____ No. USP: _____

Lista de Exercícios

1ª Questão) Com relação a redes neurais artificiais, responda os itens abaixo

a) Sabendo que todo raciocínio está fundamentado em três modalidades de inferência: dedução, indução e abdução. A dedução parte de uma regra geral e chega em uma regra particular. Já a indução parte de regras particulares e chega em regras gerais, gerando assim conhecimento novo. Procure então fornecer argumentos que sustentem as seguintes afirmações:

a.1) o princípio de operação do método do gradiente em otimização é equivalente à dedução.

O método do gradiente fornece a direção de crescimento de uma função e corresponde a uma regra geral;

a.2) o princípio de operação de redes neurais que foram sujeitas a treinamento supervisionado é equivalente à indução

Neste caso, os pesos da rede neural foram ajustados usando um conjunto de treinamento finito. O modelo é gerado e aplicado a um conjunto de teste. Logo, é assumido que o modelo induzido pelo conjunto de treinamento corresponde ao mesmo no conjunto de teste.

b) Apresente um pseudo-código para o algoritmo Back-Propagation, considerando a correção do erro em lote e sequencial. O aluno deve apresentar as equações de cálculo da saída da rede, do gradiente e atualização dos pesos.

Um pseudo-codigo foi apresentado nos slides (método batelada e método padrão a padrão)

c) Descreva quais parâmetros o usuário deve definir para executar uma Rede Neural MLP. Apresente uma estratégia para definição de cada um desses parâmetros.

Número de camadas, número de neurônios em cada camada, número de épocas, taxa de aprendizado, função de ativação

Número de camadas – Se problema estático pode-se usar uma camada escondida. Se problema dinâmico pode – se usar duas camadas escondidas.

Número de neurônios em cada camada – pode-se definir uma faixa de valores e usar um conjunto de validação para escolha do número adequado.

Número de epocas – utilizar parada antecipada, isto é, a cada época estime o erro de validação, se o erro sobre o conjunto de validação aumentar pare o treinamento.

Função de ativação da camada de saída está relacionada com o problema. Em regressão geralmente utiliza função de ativação linear, em classificação, se for binária utiliza sigmoide, se for multiclasse pode-se utilizar softmax ou sigmoide.

Função de ativação na camada escondida, geralmente utiliza sigmoide ou tangente hiperbólica. Outras funções podem ser utilizadas, mas precisam ser diferenciáveis.

d) Apresente duas vantagens e duas desvantagens dos algoritmos de segunda e primeira ordem para atualização dos pesos.

Algoritmos de primeira ordem

Vantagens

Facilidade de implementação

Desvantagens

Convergência mais lenta

Algoritmos de segunda ordem

Vantagens

Convergência mais rápida

Desvantagens

Alto custo computacional para estimação da matriz hessiana

e) Alguns autores sugerem inserir um termo de penalidade na função objetiva para controlar a suavidade do mapeamento produzido pela RNA. Descreva detalhamente esta abordagem e apresente a função objetiva a ser minimizada. Qual o significado do parâmetro de regularização?

Veja os slides de regularização L1 e L2

f) Faça um estudo comparativo entre Redes Neurais Multilayer Perceptron e Redes Neurais de Funções de Base Radial. Apresente **pelo menos duas** semelhanças e duas diferenças entre RBF e MLP. Além disso, discuta sobre os parâmetros a serem determinados em cada arquitetura.

Duas vantagens de Redes Neurais MultiLayer

- *A rede MLP pode ter mais de uma camada intermediária de neurônios, enquanto que a rede RBF tem apenas uma.*
- *As redes MLP tendem a se dar melhor no caso de número elevado de entradas, quando comparado às redes RBF.*
- *As redes MLP têm capacidade de generalização em regiões do espaço de entrada onde pouco ou nenhum exemplo de treinamento está disponível.*

Duas desvantagens de Redes Neurais MultiLayer

- *Treinamento de uma rede neural MLP envolve aplicações iterativas de um processo de ajuste incremental do vetor de pesos, sendo necessário definir a cada iteração um passo e uma direção de ajuste.*

- *O treinamento está sujeito a minimo local.*

Duas vantagens de Redes Neurais de Base Radial

- *Motivado pelas decisões de projeto, o treinamento de uma rede neural RBF se dá em um único passo de cálculo (usando uma fórmula matemática fechada), representado pela pseudo-inversão de uma matriz e por produtos entre matrizes e vetores.*
- *Em geral, é melhor o uso de redes MLP quando os padrões de entrada são custosos (ou difíceis de se gerar) e/ou quando a velocidade de recuperação é crítica. No entanto, se os dados são baratos e abundantes, e se é necessário treinamento online, então as redes RBF são superiores.*

Duas desvantagens de Redes Neurais de Base Radial -

- *O projetista da rede RBF deve definir o número de neurônios, o centro das funções de base radial e parâmetros de dispersão desta função de base radial.*
- *Conforme aumenta o número de entradas, o número de funções de base radial tende a crescer exponencialmente, caso se queira manter o mesmo nível de desempenho. Essa lei é conhecida como "maldição da dimensionalidade*

g) Explique o que é capacidade de generalização em treinamento supervisionado de redes neurais artificiais e como a disponibilidade de um conjunto de treinamento e de um conjunto de validação pode ser empregada visando maximizar esta capacidade.

A capacidade de generalização está associada à competência da rede neural em responder adequadamente para amostras de entrada-saída não utilizadas durante o processo de treinamento. Ela procura lidar com a seguinte questão: Sabendo que o erro de treinamento está baixo, que garantia se tem de que a rede neural irá produzir, em média, erros baixos também para outras amostras não utilizadas durante o treinamento? Maximizar a capacidade de generalização implica minimizar a degradação de desempenho quando se passa dos dados de treinamento para outros dados não utilizados durante o treinamento. O conjunto de validação faz justamente este papel. A parada do processo de treinamento deve se dar não visando minimizar o erro de treinamento, mas sim visando minimizar o erro junto ao conjunto de validação. Ajustam-se os pesos da rede neural empregando o conjunto de treinamento e, ao longo desse processo iterativo, monitora-se como está se dando a progressão do erro junto ao conjunto de validação. Mesmo que o erro junto ao conjunto de treinamento continue a cair, deve-se interromper o treinamento quando o erro junto ao conjunto de validação (nunca utilizado para guiar o ajuste de pesos) parar de cair, ou seja, deve-se optar pela configuração de pesos que minimiza o erro junto ao conjunto de validação, não importando como está o progresso do erro junto ao conjunto de treinamento.

h) Explique o que é maldição da dimensionalidade. Como a maldição da dimensionalidade afeta o desempenho de um modelo neural.

i) Qual a função do termo momento na função objetiva.

i) Na síntese de um classificador de padrões empregando uma rede neural, a qual é treinada a partir de dados de entrada-saída disponíveis, percebeu-se que o desempenho do classificador resultante era insatisfatório em aplicações práticas. Apresente possíveis razões para este insucesso sabendo que:

i.1) não há nenhuma informação disponível sobre o processo de treinamento do classificador;

Como não temos informação sobre o processo de treinamento, pode-se levantar as seguintes questões

a) *Ocorreu sobreajuste aos dados de treinamento, recomenda-se utilizar critério de parada antecipada via conjunto de validação*
b) *Ocorreu subajuste aos dados de treinamento, isto é, o desempenho sobre o conjunto de treinamento foi muito ruim, indicando que o modelo utilizado é muito simples para atacar o problema abordado;*
c) *Dados de treinamento não representativos – significa que a distribuição de probabilidade que gerou osd dados de treinamento e teste não são a mesma;*
d) *Não há dados de treinamento, ou seja, os parâmetros do modelo não foram ajustados.*

i.2) sabe-se que foi empregado um conjunto de treinamento, mas não foi adotado um conjunto de validação;

a) *Ocorreu sobreajuste aos dados de treinamento, recomenda-se utilizar critério de parada antecipada via conjunto de validação*
b) *Ocorreu subajuste aos dados de treinamento, isto é, o desempenho sobre o conjunto de treinamento foi muito ruim, indicando que o modelo utilizado é muito simples para atacar o problema abordado;*
c) *Dados de treinamento não representativos – significa que a distribuição de probabilidade que gerou osd dados de treinamento e teste não são a mesma;*

i.3) sabe-se que foi empregado tanto conjunto de treinamento como de validação.

a) *Ocorreu subajuste aos dados de treinamento, isto é, o desempenho sobre o conjunto de treinamento foi muito ruim, indicando que o modelo utilizado é muito simples para atacar o problema abordado;*
b) *Dados de treinamento não representativos – significa que a distribuição de probabilidade que gerou osd dados de treinamento e teste não são a mesma;*

2ª Questão) Considerando uma Rede Neural Artificial (RNA), pede-se

a) Apresente um pseudocódigo para o algoritmo BackPropagation, considerando a correção do erro em lote e sequencial. O aluno deve apresentar as equações de cálculo da saída da rede, do gradiente e atualização dos pesos.
Veja Questão 1)

b) Descreva quais parâmetros o usuário deve definir para executar uma Rede Neural. Apresente uma estratégia para definição de cada um desses parâmetros.
Veja Questão 1)

c) Apresente duas vantagens e duas desvantagens dos algoritmos de segunda e primeira ordem para atualização dos pesos.
Veja Questão 1)

d) Em problemas de predição um passo à frente, o processo de treinamento pode trabalhar com partições sequenciais dos dados da série, ou então empregar uma partição que será denominada aqui de aleatória. Explique qual a melhor forma para trabalhar com uma Rede Neural estática e uma Rede Neural Recorrente.
A rede neural estática aprende a realizar uma mapeamento de X → Y, não leva em consideração a dependência temporal nos dados, ou seja, a ordem de apresentação dos dados não tem importância. Neste caso, deve-se utilizar um vetor de entrada com os valores passados de forma a inserir memória. Neste caso, pode-se utilizar partições sequenciais ou aleatórias. Por outro lado, a rede neural recorrente leva em consideração a dependência temporal nos dados. Logo não pode-se utilizar partições randômicas, as partições devem ser sequenciais, respeitando a ordem cronológica no tempo.

e) Na tabela abaixo, parte do algoritmo para treinamento de um *Multilayer Perceptron* com BackPropagation do erro é apresentada. Considerando essa lógica de treinamento, explique a função das variáveis α, δ_k, δ_in_j e δ_j, dando destaque para o relacionamento **entre essas variáveis** e **entre essas variáveis e aquelas que representam os pesos** da rede neural.

Passo 3: Cada unidade de entrada (X_i, i = 1 ... n) recebe um sinal de entrada x_i e o dissipa para todas as unidades na camada acima (unidades escondidas).

Passo 4: Cada unidade escondida (Z_j, j = 1 ... p) soma suas entradas pesadas,

$$z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$$

aplica sua função de ativação para computar seu sinal de saída,

$$z_j = f(z\_in_j)$$

e envia o sinal para todas as unidades na camada acima (unidades de saída).

Passo 5: Cada unidade de saída (Y_k, k = 1 ... m) soma suas entradas pesadas

$$y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$$

e aplica sua função de ativação para computar seu sinal de saída,

$$y_k = f(y\_in_k)$$

Passo 6: Cada unidade de saída (Y_k, k = 1 ... m) recebe uma classificação correspondente ao padrão de entrada, computa seu termo de erro de informação

$$\delta_k = (t_k - y_k) f'(y\_in_k)$$

calcula seu termo de correção de pesos

$$\Delta w_{jk} = \alpha \delta_k z_j$$

calcula seu termo de correção de bias

$$\Delta w_{0k} = \alpha \delta_k$$

e envia δ_k para as unidades cada camada abaixo.

Passo 7: Cada unidade de saída (Z_j, j = 1 ... p) soma suas entradas delta (vindas das unidades da camada acima)

$$\delta\_in_j = \sum_{k=1}^{m} \delta_k w_{jk}$$

multiplica pela derivada de sua função de ativação para calcular seu termos de erro de informação

$$\delta_j = \delta\_in_j \; f'(z\_in_j)$$

calcula seu termo de correção de pesos

$$\Delta v_{ij} = \alpha \delta_j x_i$$

e calcula seu termo de correção de bias

$$\Delta v_{0j} = \alpha \delta_j$$

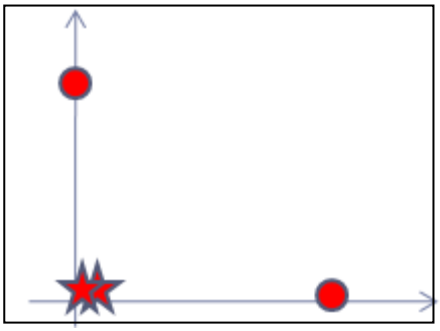
f) Discuta a função da camada escondida de uma rede neural artificial multicamadas. Forneça um exemplo didático para ilustrar sua discussão. Inclua em sua discussão uma relação entre o papel de uma camada escondida e problemas não linearmente separáveis.

Considere o problema do Ou-Exclusivo - o aluno deve explicar o que significa cada gráfico

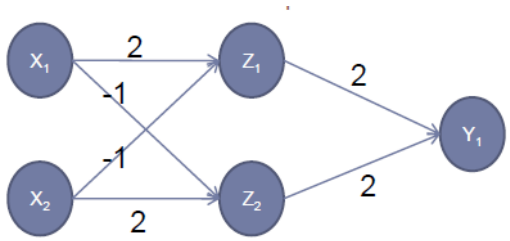
Tabela a ser preenchida

X_1	X_2	Z_1	Z_2	Y_1
0	**0**	**0**	**0**	**0**
0	**1**	**0**	**1**	**1**
1	**0**	**1**	**0**	**1**
1	**1**	**0**	**0**	**0**

Representação gráfica



Rede Neural



Limiar de decisão igual a 2. Neurônio McCulloch Pitts

Espaço para explicação textual:

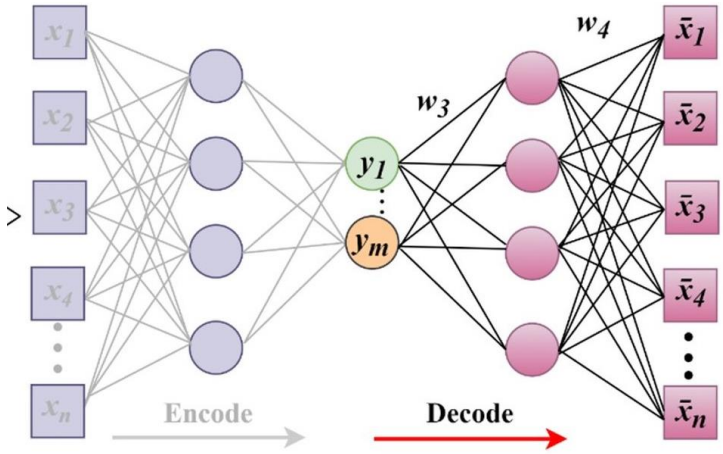
Na camada escondida os dados são mapeados para outras posições na representação gráfica de forma que o problema se torna linearmente separável. Os dados de uma das classes passam a se posicionar no mesmo local.

g) Com relação aos algoritmos de otimização irrestrita, cite dois algoritmos de primeira e segunda ordem e apresente uma vantagem e uma desvantagem de cada um.

3ª Questão) No contexto de redução de dimensão e visualização de dados, análise de componentes principais (PCA, do inglês principal component analysis) é geralmente adotada. Basicamente, o PCA define um hiperplano de projeção de dimensão menor ou igual à dimensão do espaço original dos dados. O critério de otimização do PCA é minimizar o somatório da distância dos dados originais ao hiperplano. Como consequência da aplicação deste critério, o posicionamento ótimo do hiperplano de projeção faz com que os dados projetados apresentem o máximo espalhamento possível, de modo que os eixos do hiperplano representam as direções de maior variância dos dados. Abusando um pouco da terminologia, pode-se denominar de análise de componentes principais não-lineares (NLPCA, do inglês nonlinear principal component analysis) a toda iniciativa equivalente ao PCA, mas que toma uma hipersuperfície em lugar de um

hiperplano de projeção. Como esta hipersuperfície pode ser produzida por uma rede neural MLP, pode-se empregar a técnica de NLPCA para um problema de classificação de padrões. A ideia é usar uma MLP para o papel de NLPCA e uma outra MLP para o classificador, e treinar ambas as MLPs simultaneamente. Explique detalhadamente qual deve ser a configuração de uma rede neural para que esta possa ser empregada como NLPCA.

Pode-se utilizar uma rede neural mlp com três camadas escondida (veja figura abaixo), camada de entrada e saúda tem o mesmo número de elementos. O objetivo desta rede é reproduzir na camada de saída o valor da camada de entrada. A primeira camada realiza um mapeamento para um espaço de caraterística e possui função de ativação não linear do tipo sigmoide ou tangente hiperbólica e geralmente possui um número de neurônios maior que a entrada. A segunda camada possui função de ativação linear e corresponde a camada de gqrgalo. O número de neurônios deve ser menor que o número de entrada. A terceira camada deve ter a mesma configuração da primeira camada. Após o treinamento da rede, deve-se apresentar os dados de treinamento e utilizar a saída da camada de gargalo como vetor de característica para o problema a ser tratado.



4ª Questão) Considere uma Rede Neural Convolucional (CNN) com uma camada de convolução com strider maior que um, uma camada de pooling, uma camada totalmente com função de ativação relu e uma camada de saída com função de ativação softmax. A função custo é dada por: $J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{ns} d_k(n) \ln y_k(n)$.

$d_k(n)$ - k-ésima saída desejada

$y_k(n)$ - k-ésima saída da rede neural

a) Apresente uma motivação biológica para estudo de CNN

b) Apresente a formula de atualização dos filtros da CNN.

5ª Questão) Com relação as Redes Neurais Convolucionais (CNNs), responda os itens abaixo.

a) Apresente duas motivações para o emprego de Redes Neurais Convolucionais.

b) Explique o que são pesos compartilhados e apresente uma vantagem para o emprego destes.

c) Explique detalhadamente qual a função da camada de convolução e pooling.

d) Explique detalhadamente como funciona o dropout.

Explicação detalhada https://www.deeplearningbook.com.br/capitulo-23-como-funciona-o-dropout/

e) Descreva como é realizado o treinamento de uma CNN

f) Explique como podemos utilizar uma CNN treinada para extração de característica.

Pode-se treinar uma CNN usando o conjunto de treinamento. Após a etapa de treinamento, apresenta novamente os dados de treinamento e armazena a saída da camada flatten. Estes dados desta camada corresponde as características extraídas dos dados de treinamento.

g) Explique o que seria os autoenconders.

Autoencoders são redes neurais que visam copiar suas entradas para suas saídas. Eles trabalham compactando a entrada em uma representação de espaço latente e, em seguida, reconstruindo a saída dessa representação. Esse tipo de rede é composto de duas partes: ***Codificador (Encoder)****: é a parte da rede que compacta a entrada em uma representação de espaço latente (codificando a entrada). Pode ser representado por uma função de codificação h = f (x).* ***Decodificador (Decoder)****: Esta parte tem como objetivo reconstruir a entrada da representação do espaço latente. Pode ser representado por uma função de decodificação r = g (h).*

h) Explique como funciona autoencoders variacional e robustos.

i) Dado uma CNN pretreinada para reconhecimento facial. Como podemos utilizá-la para extração de características faciais?

Item f)

j) Como podemos utilizar os autoencoders para atacar o problema da maldição da dimensionalidade?

6ª Questão) A tabela abaixo apresenta os exemplos de decisão sobre que ações um robô deve tomar enquanto dirige um carro.

Semáforo	Luz de Freio	Distância do carro frente (metros)	Ação
Verde	Não	10	Andar
Verde	Sim	50	Andar
Verde	Sim	58	Parar
Vermelho	Sim	60	Parar
Vermelho	Não	70	Parar
Amarelo	Não	80	Virar à Esquerda
Amarelo	Sim	30	Virar à Direita
Amarelo	Não	10	Virar à Esquerda

Suponha uma MLP com 3 neurônios na camada escondida com função de ativação softmax e saída com função de ativação softmax para modelar este problema. Considere uma taxa de aprendizado igual a 0.5. Pede-se

a) Apresente a codificação e o preprocessamento adotado.

Semáforo
Verde – 1 0 0
Vermelho – 0 1 0
Amarelo – 0 0 1
Luz
Não – 0
Sim – 1
Distância - deve-se normalizar no intervalor [0,1]
Ação
Andar – 1 0 0 0
Parar – 0 1 0 0

Virar Direita – 0 0 1 0
Virar Esquerda – 0 0 0 1

b) Inicialize todos pesos. Calcule a saída da rede neural e monte a matriz de confusão

c) Mostre detalhadamente os cálculos que o algoritmo backpropagation executará na primeira época.

d) Monte a curva roc após treinar a rede neural uma época.

7ª Questão) Considerando Árvore de Decisão, responda os itens abaixo.

a) Apresente duas motivações para utilizar uma Árvore de Decisão

b) Explique como é realizado o processo de indução de uma árvore de decisão: algoritmo de indução, medidas de escolha de nós, parâmetros livres (discuta), estratégias de pré-poda e pós-poda.

c) Randon Forests: qual a motivação, estratégias para construção das árvores e para combinação de resultados, parâmetros livres.

Se o número de dados de treinamento é N, uma amostragem randômica de tamanho N, com reposição, do conjunto de treinamento original realizada. Sobre cada amostragem constru uma _arvore.
Se existem M atributos descritivos no conjunto de dados, um n_umero m << M _e escolhido tal que para cada n_o, m atributos sejam randomicamente selecionados de M a serem submetidos ao critério de escolha de atributos para construção das futuras partições.

d) Explique como podemos utilizar uma árvore de decisão para seleção de atributos

Os atributos no topo da arvore corresponde a aqueles com maior ganho de informação, estes podem ser utilizados como atributos mais relevantes.

e) Dado um atributo numérico, explique detalhadamente como devemos proceder para utilizá-lo na construção de uma árvore de decisão.

Este atributo deve ser discretizado usando, por exemplo, a entropia. Os candidatos a Split point corresponde aos valores numéricos assumidos pelo atributo.

8ª Questão) Considere o problema de aprender o conceito sobre pacientes doentes (+) ou sadios (-). Foram coletados os seguintes exemplos

Conjunto de treinamento			
Idade	Exame A	Exame B	Classe
20	H	C	+
25	R	B	+
28	R	C	+
23	H	B	+
19	J	B	+
35	R	C	-
30	J	C	-
38	J	B	-
40	H	C	+
24	J	C	-
29	R	C	-
20	J	B	+
19	R	C	-

Conjunto de Teste			
20	R	B	+
22	J	B	-
30	J	C	-
31	R	C	+
29	H	C	+

a) Utilizando o algoritmo de indução de árvores de decisão, construa a árvore correspondente (sem poda e sem número mínimo de exemplos em cada folha), utilizando o critério de ganho de informação para selecionar atributos para este conjunto de exemplos. Anote, para cada nível da árvore de decisão, o valor do ganho de informação calculado para cada atributo, bem como aquele escolhido para particionar os exemplos. Se houver empate entre valores do ganho de informação, escolha o primeiro (na ordem da tabela acima). Se for necessário realizar a discretização dos atributos, descreve detalhadamente o processo adotado.

Discretização do atributo idade

Split Point	<	>=	Entropia
20	1 +, 1 -	6 +, 5-	$Info(D) = \frac{2}{13} * E\left(\frac{1}{2};\frac{1}{2}\right) + \frac{11}{13} * E(\frac{6}{11};\frac{5}{11}) = 0{,}9949)$
23	3+, 1-	4+, 5-	$Info(D) = \frac{4}{13} * E\left(\frac{3}{4};\frac{1}{4}\right) + \frac{9}{13} * E(\frac{4}{9};\frac{5}{9}) = 0{,}9358$
24	4+, 1-	3+, 5-	$Info(D) = \frac{5}{13} * E\left(\frac{4}{5};\frac{1}{5}\right) + \frac{8}{13} * E(\frac{3}{8};\frac{5}{8}) = 0{,}8650$
25	4+, 2-	3+, 4-	$Info(D) = \frac{6}{13} * E\left(\frac{4}{6};\frac{2}{6}\right) + \frac{7}{13} * E(\frac{3}{7};\frac{4}{7}) = 0{,}9543$
28	5+, 2-	2+, 4-	$Info(D) = \frac{7}{13} * E\left(\frac{5}{7};\frac{2}{7}\right) + \frac{6}{13} * E(\frac{2}{6};\frac{4}{6}) = 0{,}8886$
29	6+, 2-	1+, 4-	$Info(D) = \frac{8}{13} * E\left(\frac{6}{8};\frac{2}{8}\right) + \frac{5}{13} * E(\frac{1}{5};\frac{4}{5}) = 0{,}7769$
30	6+, 3-	1+, 3-	$Info(D) = \frac{9}{13} * E\left(\frac{6}{9};\frac{3}{9}\right) + \frac{4}{13} * E(\frac{1}{4};\frac{3}{4}) = 0{,}8854$
35	6+, 4-	1+, 2-	$Info(D) = \frac{10}{13} * E\left(\frac{6}{10};\frac{4}{10}\right) + \frac{3}{13} * E(\frac{1}{3};\frac{2}{3}) = 0{,}9588$
38	6+, 5-	1+, 1-	$Info(D) = \frac{11}{13} * E\left(\frac{6}{11};\frac{5}{11}\right) + \frac{2}{13} * E(\frac{1}{2};\frac{1}{2}) = 0{,}9949$
40	6+, 6-	1+, 0-	$Info(D) = \frac{12}{13} * E\left(\frac{6}{12};\frac{6}{12}\right) + \frac{1}{13} * E(\frac{1}{1};\frac{0}{1}) = 0{,}9231$

Conjunto de treinamento				
ID	Idade	Exame A	Exame B	Classe
1	[19-29]	H	C	+
2	[19-29]	R	B	+
3	[19-29]	R	C	+
4	[19-29]	H	B	+
5	[19-29]	J	B	+
6	[29-40]	R	C	-
7	[29-40]	J	C	-
8	[29-40]	J	B	-
9	[29-40]	H	C	+
10	[19-29]	J	C	-
11	[29-40]	R	C	-
12	[19-29]	J	B	+
13	[19-29]	R	C	-

Conjunto de Teste				
14	[19-29]	R	B	+
15	[19-29]	J	B	-
16	[29-40]	J	C	-
17	[29-40]	R	C	+
18	[29-40]	H	C	+

Cálculo da informação do conjunto de dados

+	-

7	6

$$Info(D) = -\frac{7}{13}\log_2\left(\frac{7}{13}\right) - \frac{6}{13}\log_2\left(\frac{6}{13}\right) = 1.0\ bits$$

Idade

[19-29]		[29-40]	
+	-	+	-
6	2	1	4

$$Info(D) = \frac{8}{13} * E\left(\frac{6}{8};\frac{2}{8}\right) + \frac{5}{13} * E(\frac{1}{5};\frac{4}{5}) = 0{,}7769$$

$$Gain(Idade) = \ 1 - 0.7769 = 0.2231$$

Atributo Exame A

H		R		J	
+	-	+	-	+	-
3	0	2	3	2	3

Cálculo da informação usando o Atributo Exame A

$$Info_{ExameA}(D) = -\frac{3}{13}\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right) - \frac{5}{13}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$$
$$-\frac{5}{13}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$$

$$Info_{ExameA}(D) = 0.7469$$
$$Gain(ExameA) = \ 1 - 0.7469 = 0.2531$$

Atributo Exame B

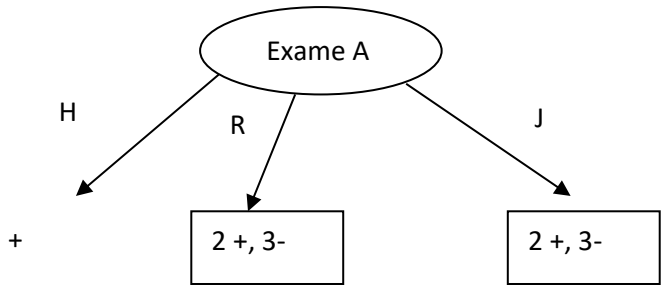
B		C	
+	-	+	-
4	1	3	5

Cálculo da informação usando o Atributo Exame A

$$Info_{ExameB}(D) = -\frac{5}{13}\left(\frac{4}{5}\log_2\left(\frac{4}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) - \frac{8}{13}\left(\frac{3}{8}\log_2\left(\frac{3}{8}\right) + \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right)$$

$$Info_{ExameA}(D) = 0.8650$$
$$Gain(ExameB) = \ 1 - 0.8650 = 0.1350$$

Atributo Escolhido Exame A



+	-
2	3

$$Info(D) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0{,}9710\ bits$$

Idade

<29		>=29	
+	-	+	-
2	1	0	2

$$Info_{Idade}(D) = -\frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right)+\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)-\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)+\frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0{,}5510$$

$$Gain(Idade) = 0{,}9710 - 0.5510 = 0.42$$

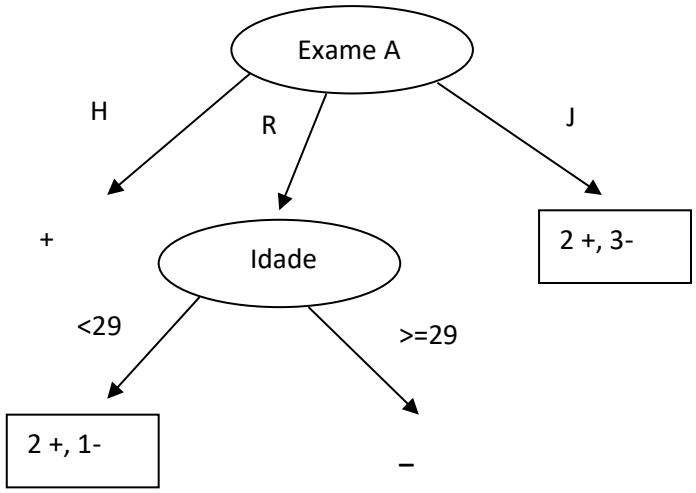
Usando atributo Exame B

B		C	
+	-	+	-
1	0	1	3

Cálculo da informação usando o Atributo Exame A

$$Info_{ExameB}(D) = -\frac{1}{5}\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right)+\frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)-\frac{4}{5}\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right)+\frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) = 0{,}6490$$

$$Gain(ExameB) = 0{,}9710 - 0.6490 = 0.322$$



+	-
2	3

$$Info(D) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0{,}9710\ bits$$

Idade

<29		>=29	
+	-	+	-
2	1	0	2

$$Info_{Idade}(D) = -\frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right)+\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)-\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)+\frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0{,}5510$$

$$Gain(Idade) = 0{,}9710 - 0.5510 = 0.42$$

Usando atributo B

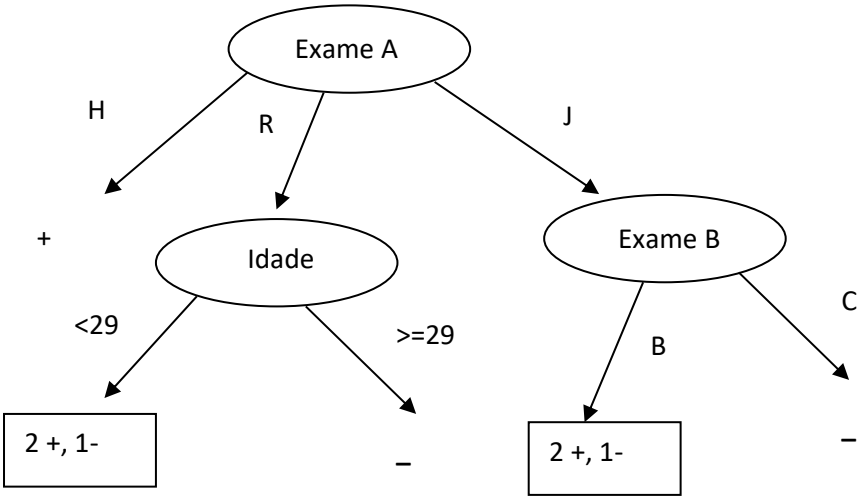
B		C	
+	-	+	-
2	1	0	2

Cálculo da informação usando o Atributo Exame B

$$Info_{Idade}(D) = -\frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right)+\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)-\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)+\frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0{,}5510$$

$$Gain(Idade) = 0{,}9710 - 0.5510 = 0.42$$

Empate. Escolhendo atributo Exame B



+	-
2	1

$$Info(D) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right)-\frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0{,}9183\ bits$$

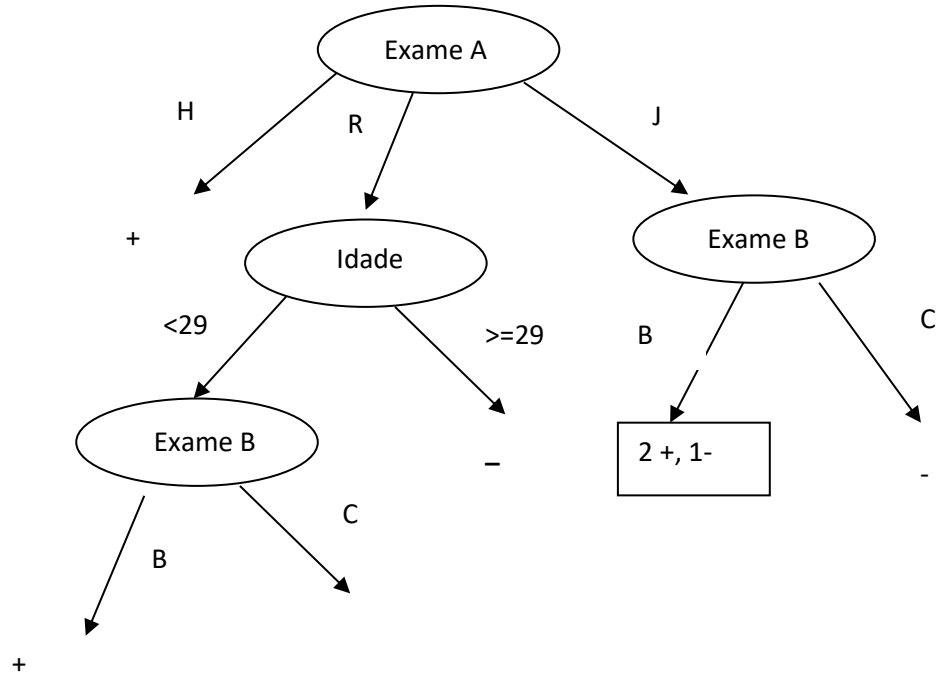
Usando atributo Exame B

B		C	
+	-	+	-
1	0	1	1

Cálculo da informação usando o Atributo Exame B

$$Info_{Exame\ B}(D) = -\frac{1}{3}\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) - \frac{2}{3}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 0{,}66$$

$$Gain(Exame\ B) = 0{,}9183 - 0.6667 = 0.2516$$



+	-
2	1

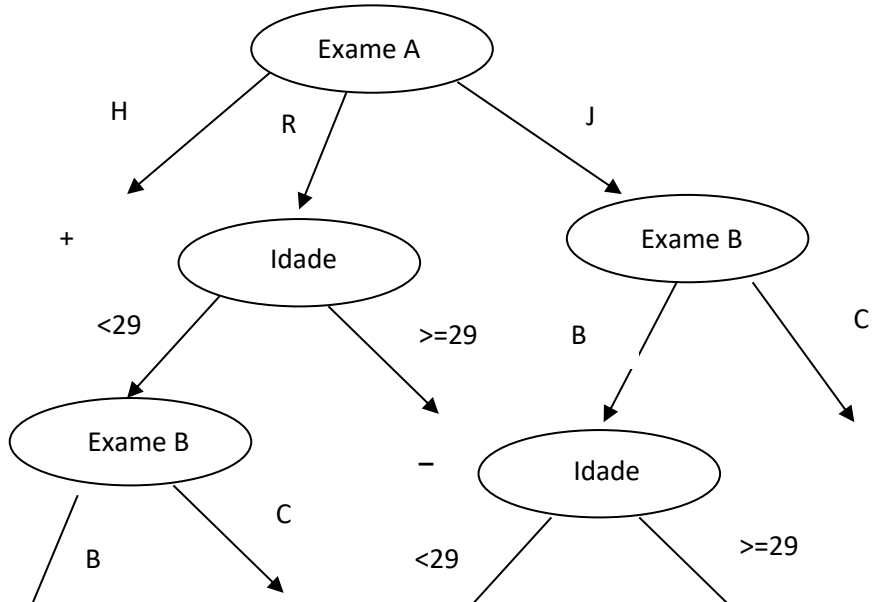
$$Info(D) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0{,}9183\ bits$$

Idade

<29		>=29	
+	-	+	-
2	0	0	1

$$Info_{Exame\ B}(D) = -\frac{2}{3}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) - \frac{1}{3}\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$

$$Gain(Exame\ B) = 0{,}9183 - 0 = 0.9183$$



-

+ +

b) Análise de desempenho: use a árvore de decisão produzida em (a) para classificar os exemplos do conjunto de teste. Informe a precisão e matriz de confusão da árvore para esses exemplos. Discuta resumidamente os resultados.

c) Conjunto de Teste					Classificação
14	[19-29]	R	B	+	+
15	[19-29]	J	B	-	+
16	[29-40]	J	C	-	-
17	[29-40]	R	C	+	-
18	[29-40]	H	C	+	+

Matriz de confusão

Real\Predita	C+	C-
C+	2 – TP	1– FN
C-	1 – FP	1 – TN

Precision e Acurracy

$$Precision = \frac{TP}{TP + FP} = \frac{2}{2 + 1} = 0.66$$

$$Accuray = \frac{TP + TN}{TP + FN + FP + TN} = \frac{3}{5} = 0.6$$

Análise

Observe que a árvore de decisão não obteve uma alta acurácia para o conjunto de teste. Comparando o conjunto de treinamento com o de teste, pode-se observar que as duas instâncias do conjunto de teste classificadas incorretamente apresentam rótulos contrários ao usado no conjunto de treinamento (conforme destaque realizado nas tabelas de treinamento e teste). Em função disso, uma análise deve ser realizada para verificar se o conjunto de treinamento ou teste foram rotulados corretamente.

d) Aplique duas critérios de poda visto em sala de aula a árvore gerada no item b). Em seguida apresente a matriz de confusão para o conjunto de teste.

9ª Questão) Considerando Máquinas de Vetores Suporte, responda os itens abaixo.

e) Descreva detalhadamente o SVM e suas variações (pelo menos 2). Apresente a formulação do primal e do dual.
f) Descreva detalhadamente como SVM pode ser empregado para problemas com múltiplas classes. Apresente três estratégias e explique detalhadamente cada uma.
g) Explique detalhadamente os três métodos alternativos para treinamento de SVM (Chunking, Algoritmo de Osuna, SMO).
h) Explique o que significa o Truque do Kernel (Kernel Trick).
i) Explique detalhadamente como podemos aplicar dimensão VC para selecionar o tipo de kernel.

10ª Questão) Considerando uma Rede Neural Artificial (RNA), pede-se

a) Desenhe uma RNA, com número mínimo de neurônios na camada escondida capaz de modelar o problema do Ou-Exclusivo.
b) Considerando o item c), inicialize todos os pesos iguais 0.1 e calcule a saída da RNA. Em seguida, mostre os cálculos que o algoritmo Levenberg–Marquardt de atualização dos pesos executará na primeira época para apenas uma instância
c) Apresente duas vantagens e duas desvantagens dos algoritmos de segunda e primeira ordem para atualização dos pesos
d) Alguns autores sugerem inserir um termo de penalidade na função objetiva para controlar a suavidade do mapeamento produzido pela RNA. Descreva detalhadamente esta abordagem e apresente a função objetiva a ser minimizada. Qual o significado do parâmetro λ?

11ª Questão) Considerando Comitê de Máquinas
a) Faça um comparativo entre a abordagem proposta por Rosen (Descorrelated Neural Networks) e por Liu (Learning via Negative Correlation)
b) Explique detalhadamente as 3 fases de construção do ensemble. Para cada fase apresente 2 estratégias
Fases: Geração, Seleção, Combinação
c) Apresente três vantagens e duas desvantagens do Ensemble e da Mistura de Especialista.
d) Explique detalhadamente o papel gating em Mistura de Especialista. Apresente estratégias de incialização para gating e especialistas baseada em aprendizado não supervisionado.

12ª Questão) Considerando Máquinas de Vetores Suporte

a) Faça uma breve discussão sobre as variações (pelo menos 3) de SVM. Apresente a formulação primal e dual de cada abordagem.
b) Descreva detalhadamente como SVM pode ser empregado para problemas com múltiplas classes.
c) Explique o que significa o Truque do Kernel (Kernel Trick).
d) Explique detalhadamente os três métodos alternativos para treinamento de SVM (Chunking, Algoritmo de Osuna, SMO).

e) Descreva detalhadamente Support Vector Clustering (SVC). Qual o problema dessa abordagem.Explique detalhadamente como podemos aplicar dimensão VC e SVC para selecionar o tipo de kernel.

f) Explique detalhadamente Extreme Support Vector Machine

12ª Questão) Apresente uma arquitetura de uma MLP que pode ser empregada para modelar componentes principais. Em seguida apresente todos passos e as derivadas para treinamento da rede neural especificada e o cálculo das componentes.

Veja a questão do NLPCA

13ª Questão) Suponha que uma MLP com 3 neurônios na camada escondida taxa de aprendizado seja 0.5, seja utilizada para modelar o problema do Ou-Exclusivo. Utilize o menor número de saídas possível para o problema.

a) Inicialize todos pesos com 0.1. Calcule a saída da rede neural e monte a matriz de confusão

b) Mostre os cálculos que o algoritmo backpropagation executará na primeira época para o peso a_{22} da camada de entrada e b_{11} para camada intermediária, considerando apenas a primeira instância

14ª Questão) Discuta a função da camada escondida de uma rede neural artificial multicamadas. Forneça um exemplo didático para ilustrar sua discussão. Inclua em sua discussão uma relação entre o papel de uma camada escondida e problemas não linearmente separáveis.

16ª Questão) Assinale verdadeiro (V) ou falso (F). Lembre-se que um item assinalado incorretamente anula um item corretamente

(..V...) A escolha adequada da taxa de aprendizado em Redes Neurais Artificiais é muito importante para assegurar a estabilidade da convergência do processo de aprendizado iterativo, pois taxas pequenas permitem um aprendizado mais lento porém mais consistente, mas com o perigo de cair em mínimos locais, enquanto taxas maiores permitem um aprendizado mais rápido a custo muitas vezes de desestabilização (oscilação)

(..V...) A ativação de um neurônio na rede MLP se dá pelo produto interno entre seu vetor de pesos e o vetor de entradas, seguida pela aplicação da função de ativação, geralmente do tipo sigmoidal. Por outro lado, a ativação de um neurônio na rede RBF se dá pelo cálculo da distância euclidiana entre o vetor de pesos do neurônio e o vetor de entradas. Quanto mais distante o vetor de entrada do vetor de pesos, menor a ativação do neurônio.

(..V....) as redes MLP tendem a se dar melhor no caso de número elevado de entradas, quando comparado às redes RBF. Conforme aumenta o número de entradas, o número de funções de base radial tende a crescer exponencialmente, caso se queira manter o mesmo nível de desempenho. Essa lei é conhecida como “maldição da dimensionalidade”.

(...V...) motivado pelas decisões de projeto, o treinamento de uma rede neural RBF se dá em um único passo de cálculo (usando uma fórmula matemática fechada), representado pela pseudo-inversão de uma matriz e por produtos entre matrizes e vetores. Por outro lado, o treinamento de uma rede neural MLP envolve aplicações iterativas de um processo de ajuste incremental do vetor de pesos, sendo necessário definir a cada iteração um passo e uma direção de ajuste.

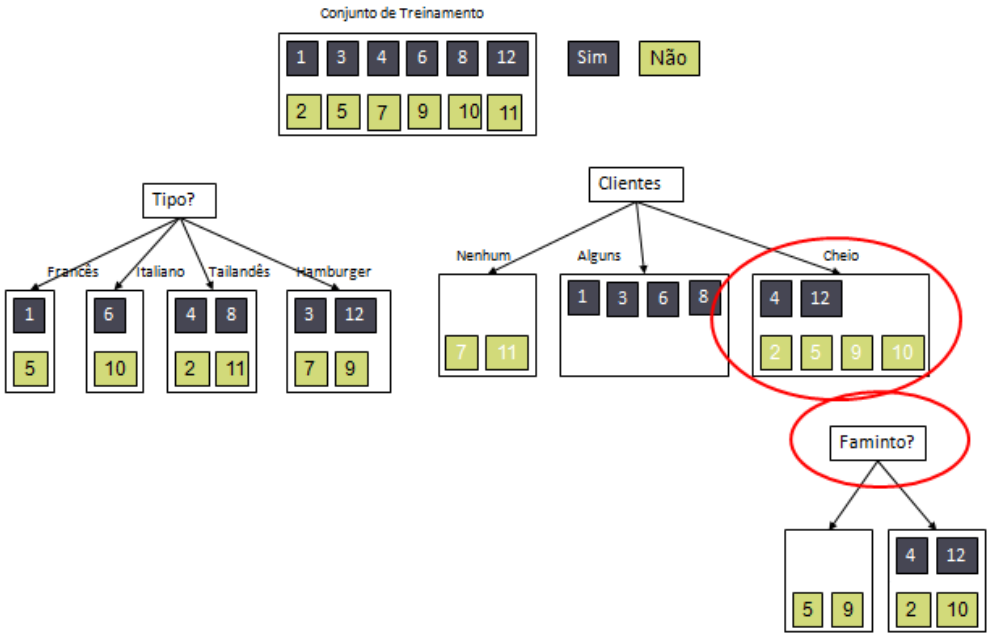
(..V...) os únicos parâmetros ajustáveis na rede RBF são os pesos da camada de saída, enquanto que na rede MLP os pesos da camada intermediária são também ajustáveis. Com isso, o projetista da rede RBF deve definir o número de neurônios, o centro das funções de base radial e parâmetros de dispersão desta função de base radial.

(F) Holdout reserva uma certa quantidade de dados para treinamento e o restante para teste (podendo ainda usar parte para validação). Comumente esta estratégia uma 1/3 dos dados para teste e o restante para treinamento, escolhido randomicamente. Para conjunto de dados randômicos é interessante assegurar que a amostragem randômica seja feita de tal maneira que garanta que cada classe é apropriadamente representada tanto no conjunto de treinamento quanto no conjunto de teste. Este procedimento é chamado de **holdout repetitivo**. Se for realizado apenas uma divisão do conjunto de dados, a estimativa da taxa de erro vai ser enganosa se acontecer de termos uma divisão ruim. Visando amenizar tendências, emprega-se **holdout estratificado**, o qual consiste em repetir todo o processo de treino e teste várias vezes com diferentes amostragens randômicas.

(V) O bootstrap é em um procedimento estatístico de amostragem com reposição. Considerando um conjunto de dados com n instâncias, n instâncias são escolhidas aleatoriamente. Uma instância não é retirada do conjunto de dados original quando ela é escolhida para compor o conjunto de treinamento, ou seja, a mesma instância pode ser selecionada várias vezes durante o procedimento de amostragem.

(V) No n fold cross validation o conjunto de dado é dividido em n partições de tamanhos aproximadamente iguais e, de maneira rotativa, cada uma delas é usada para teste enquanto as restantes são usadas para treinamento. Este procedimento é repetido n vezes. Para conjunto de dados pequeno, geralmente n é escolhido igual ao número de instâncias no conjunto de dados. Este é também conhecido como leave one out.

16ª Questão) Analise o gráfico abaixo e responda as questões na sequência:



a) Entre os atributos TIPO e CLIENTES, qual tem maior entropia? Qual deles deve ser escolhido para

iniciar a construção de uma árvore? Apresentes os cálculos que justifique sua resposta.

$$Inf(D) = \frac{6}{12} log \frac{6}{12} + \frac{6}{12} log \frac{6}{12} = 1$$

$$Info_{Tipo}(D) = \frac{2}{12} E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{12} E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} E\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{4}{12} E\left(\frac{2}{4}, \frac{2}{4}\right) = 1$$

$$Gain_{Tipo}(D) = 1 - 1 = 0$$

$$Info_{Cliente}(D) = \frac{2}{12} E(0,1) + \frac{4}{12} E(1,0) + \frac{6}{12} E\left(\frac{2}{6}, \frac{4}{6}\right) = 0.4591$$

$$Gain_{Tipo}(D) = 1 - 0.4591 = 0.5409$$

Logo, o atributo que tem maior entropia é Tipo, pois este atributo não consegue definir nenhuma classe. Se analisarmos a informação necessária para separar os dados após utilizarmos o atributo Tipo, ainda precisamos de informação igual a 1. Já para o atributo Clientes precisamos apenas 0.54. Logo o melhor atributo é Cliente.

b) Se o atributo CLIENTES é usado na árvore, o que ocorre com os nós no fim dos ramos "NENHUM" e "ALGUNS"?

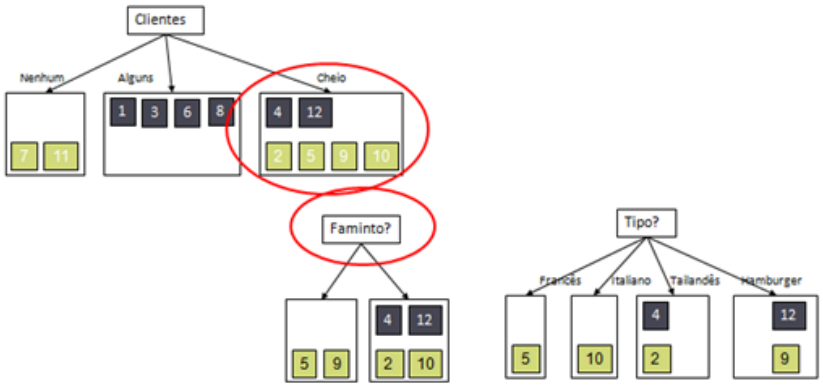
O nó Nenhum receberá rótulo NÃO e o nó Alguns receberá rótulo SIM.

c) Se pararmos o processo de construção da árvore após a inserção dos nós com os atributos CLIENTE e FAMINTO, o que podemos dizer sobre o desempenho de classificação desta árvore?

A árvore cometerá 2 erro, referente as instâncias 4 e 12 que serão rotuladas como Não.

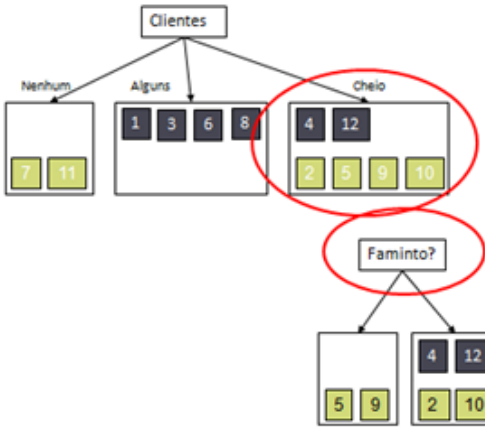
d) Considerando os três atributos (tipo, clientes e faminto), qual seria a árvore de decisão final (apresente todos os cálculos)? Rotule todos os nós folhas, em caso de empate escolha a classe final como SIM. Mostre a matriz de confusão

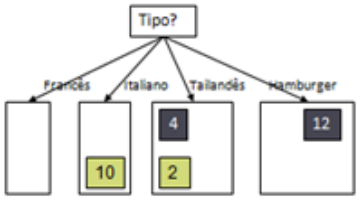
Sabemos que o atributo raiz deve ser Clientes. Temos que definir qual atributo a ser utilizado na sequencia.



A árvove acima ilustra o resultado obtido usando o atributo Faminto ou Tipo. Se calcularmos o ganho obtido com Faminto ou Tipo podemos obervar um empate.

Se escolhermos o atributo Faminto e em sequencia utilizarmos o atributo Tipo pode-se observar que temos um ganho pequeno. A figura abaixo ilustra a arvore gerada.





Se escolhermos o atributo Tipo e em sequencia utilizarmos o atributo Faminto pode-se observar que temos um ganho pequeno. A figura abaixo ilustra a arvore gerada. Em ambos os casos não é possível separar as instâncias 4 e 2

17ª Questão) Considere o seguinte conjunto de exemplos de treinamento, onde a variável GripeA é a classe:

Tosse	Febre	Viagem	GripeA
V	F	F	F
F	F	V	F
V	V	F	F

a) Utilizando o algoritmo de indução de árvores de decisão, construa a árvore correspondente (sem poda e sem número mínimo de exemplos em cada folha), utilizando o critério de ganho de informação para selecionar

V	V	V	V
F	V	F	F
F	F	F	F
V	F	V	V
F	V	V	V

atributos para este conjunto de exemplos. Anote, para cada nível da árvore de decisão, o valor do ganho de informação calculado para cada atributo, bem como aquele escolhido para particionar os exemplos. Se houver empate entre valores do ganho de informação, escolha o primeiro (na ordem da tabela acima).

b) Por que uma árvore de decisão podada que não se ajusta perfeitamente aos dados de treinamento pode ser melhor do que uma árvore não podada?

18ª Questão) Considere os conjuntos de dados descritos nos quadros abaixo:

i)

Seeds Data Set (adaptado de UCI – Machine Learning Repository / doado pela John Paul II Catholic University of Lublin e Cracow University of Technology)

Características do conjunto de dados:	Multivariado	**Número de instâncias**	210
Características dos atributos:	Numérico / categórico	**Número de atributos**	7
Tarefas associadas:	Classificação / agrupamento	**Valores faltantes**	Não

No conjunto de dados estão armazenadas as seguintes informações sobre sementes:
1. área (A): numérico
2. perímetro (P): numérico
3. compacidade (C = 4*pi*A/P^2): numérico
4. comprimento do núcleo: numérico
5. largura do núcleo: numérico
6. coeficiente de assimetria: numérico
7. comprimento do núcleo do sulco: numérico
8. variedade: kama, rosa, canadense (atributo classe)

ii)

Bank Marketing Data Set

Resumo: Os dados estão relacionados com campanhas de marketing direto (via chamadas telefônicas) de uma instituição financeira portuguesa. A meta do problema de classificação é decidir se o cliente aceitará o produto. Frequentemente, é necessária a realização de mais de um contato com o mesmo cliente, a fim de verificar se o produto seria aceito ou não.

Data Set Characteristics:	Multivariate	Number of Instances:	45211
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	17

Associated Tasks:	Classification	Missing Values?	N/A

Fonte: [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS (http://hdl.handle.net/1822/14838)

Informações sobre o conjunto de dados:
Existem dois conjuntos de dados:
1) bank-full.csv com todos os dados (instâncias) disponíveis, ordenadores por data (de maio de 2009 a novembro de 2010)
2) bank.csv com 10% dos dados (4521 instâncias), randomicamente selecionados de bank-full.csv.

Informações sobre os atributos:
(dados sobre os clientes)

- Idade (numérico)
- Tipo de emprego (categórico: “admin”, “desconhecido”, “desempregado”, “gerente”, “dona de casa”, “empresário”, “estudante”, “operário”, “autônomo”, “aposentado”, “técnico”, “prestador de serviço”)
- Estado civil (categórico: “casado”, “divorciado”, “solteiro”)
- Grau de instrução (categórico: “desconhecido”, “primário”, “secundário”, “terciário”)
- Possui crédito? (binário: sim / não)
- Renda anual (numérico)
- Possui hipoteca imobiliária? (binário: sim / não)
- Possui empréstimo? (binário: sim / não)

(dados sobre o último contato da campanha atual)

- Tipo de comunicação do contato (categórico: “desconhecido”, “telefone”, “celular”)
- Dia do mês do contato (numérico)
- Duração do contato em segundos (numérico)

(outros atributos)

- Número de contatos executados durante a campanha para este cliente (numérico)
- Número de dias que se passaram desde o último contato com o cliente considerando campanha anterior (numérico / -1 significa que o cliente não foi contatado antes)
- Número de contatos executados com este cliente antes desta campanha (numérico)
- Resultado da campanha anterior com relação ao cliente (categórico: “desconhecido”, “outros”, “falha”, “sucesso”)

(variável de classe)

- O cliente contatado aceitou o produto? (binário: sim (5288 instâncias) / não (39923 instâncias))

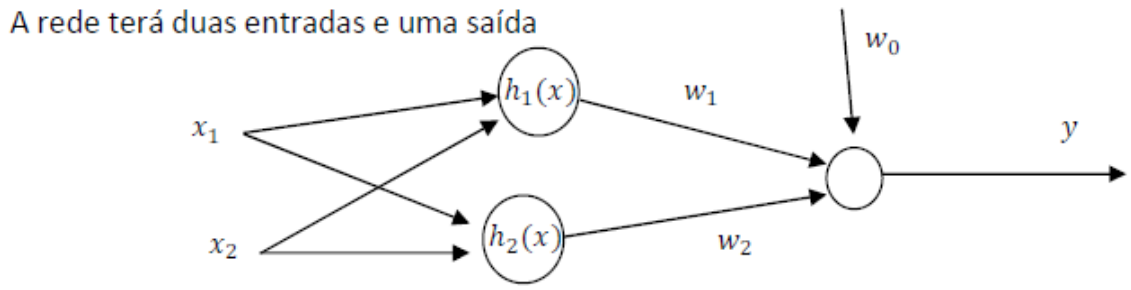
a) Para resolver os problemas de classificação descritos nos quadros i) e ii), você considera que seria mais adequado usar uma rede neural artificial (com aprendizado supervisionado) ou uma árvore de decisão? Justifique sua resposta. Em sua justificativa, considere a teoria que você estudou sobre redes neurais artificiais e árvores de decisão, e argumente em favor de sua escolha usando informações presentes na descrição do conjunto de dados.

Como os atributos do quadro i) são números, para aplicarmos uma Árvore de Decisão a este problema teríamos que discretizá-los. Se decidirmos aplicar Rede Neural este pré-processamento não será necessário. Visando eliminar este processo, a técnica mais recomendada seria uma Rede Neural. Por outro lado, o quadro ii) apresenta vários categóricos e alguns números, para aplicarmos Rede Neural teriamos que codificar os atributos categóricos. Se decidirmos aplicar Árvore de decisão precisariamos discretizalos alguns atributos números. Sobre este ponto de vista, a técnica mais recomendada seria Árvore de Decisão

b) Considerando que alguns atributos descritivos dos problemas acima não possuem um poder de expressividade muito bom, e que há algumas classes mal representadas e difíceis de aprender, que tipo de alteração você poderia fazer para atender melhor ao problema? Por que essa alteração vai ajudar na construção de um modelo de classificação melhor?

Poderia ser empregado métodos de seleção/extração de características para reduzir a dimensionalidade dos dados, como por exemplo, feature selection ou análise de componentes principais. Ao reduzir a dimensionalidade teremos um número menor de características

18a Questão) (1.0 pontos) Considere o problema do Ou-Exlusivo, definido em sala de aula. Suponha que uma rede neural do tipo RBF, com dois na camada interna, tenha sido empregada para este problema
a) (0.5 ponto) Apresente a representação gráfica da rede para este problema, definindo o número de entradas e saída.



b) (0.5 ponto) Considere $h_1(x), h_2(x)$ descrito abaixo, como a função de ativação dos neurônios 1 e 2 da camada escondida da rede RBF. Supondo que o vetor de pesos da camada de saída seja $w = [w_0\ w_1\ w_2] = [0.1\ 0.4\ 0.3]$. Calcule a saída da rede para o problema do Ou-Exclusivo

$$h_1(x) = exp\left(-\frac{(x_1-c_{11})^2}{2\sigma_1^2} - \frac{(x_2-c_{12})^2}{2\sigma_1^2}\right), h_2(x) = exp\left(-\frac{(x_1-c_{21})^2}{2\sigma_2^2} - \frac{(x_2-c_{22})^2}{2\sigma_2^2}\right)$$

$$\text{onde } c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ e } \sigma_1 = 1, \sigma_2 = 1$$

x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	0

Cálculo da saída para primeira entrada

$$h_1(x) = exp\left(-\frac{(0-1)^2}{2} - \frac{(0-0)^2}{2}\right) = 0.606$$
$$h_2(x) = exp\left(-\frac{(0-0)^2}{2} - \frac{(0-1)^2}{2}\right) = 0.606$$
$$f(x) = 0.1 * 1 + 0.4 * 0.606 + 0.3 * 0.606 = 0.524$$

$$f(x) = 1$$

Cálculo da saída para segunda entrada

$$h_1(x) = exp\left(-\frac{(1-1)^2}{2} - \frac{(0-0)^2}{2}\right) = 1$$

$$h_2(x) = exp\left(-\frac{(1-0)^2}{2} - \frac{(0-1)^2}{2}\right) = 0.367$$
$$f(x) = 0.1 * 1 + 0.4 * 1 + 0.3 * 0.367 = 0.81$$

$$f(x) = 1$$

Cálculo da saída para terceira entrada

$$h_1(x) = exp\left(-\frac{(0-1)^2}{2} - \frac{(1-0)^2}{2}\right) = 0.367$$
$$h_2(x) = exp\left(-\frac{(0-0)^2}{2} - \frac{(1-1)^2}{2}\right) = 1$$
$$f(x) = 0.1 * 1 + 0.4 * 0.367 + 0.3 * 1 = 0.546$$

$$f(x) = 1$$

Cálculo da saída para quarta entrada

$$h_1(x) = exp\left(-\frac{(1-1)^2}{2} - \frac{(1-0)^2}{2}\right) = 0.606$$
$$h_2(x) = exp\left(-\frac{(1-0)^2}{2} - \frac{(1-1)^2}{2}\right) = 0.606$$
$$f(x) = 0.1 * 1 + 0.4 * 0.606 + 0.3 * 0.606 = 0.524$$

$$f(x) = 1$$