MBA em Ciência de Dados

CeMEAI - ICMC/USP - São Carlos

Avaliação Final

Aprendizado de Máquina

Docente: Prof. Dr. André C. P. L. F. de Carvalho

Tutores: Edesio, Eugênio, Gustavo, Moisés, Saulo e Victor

Aluno: Benicio Ramos Magalhães

Instruções e Avisos:

- A avaliação vale 10 pontos
- As questões de 1 a 4, caso respondidas da forma correta, já totalizam 10 pontos
- A questão 5 é extra e vale 1 ponto
- Responda as questões nas células reservadas para respostas, demarcadas com *"Resposta: "*
- Seja sucinto e objetivo nas respostas

Questão 1. (2,5)

O algorimto k-vizinhos mais próximos (kNN) é um algoritmo de aprendizado de máquina baseado em distância. Pré-processamento e escolhas adequadas de k devem ser feitas para evitar comportamentos indesejados. O que pode-se dizer sobre:

a) O que pode acontecer de indesejado quando k tende a ser muito pequeno, tendendo a 1 vizinho. (1,0)

Resposta:

Esse algoritmo usa do cálculo da distância entre a amostra que desejamos classificar e as K amostras mais próximas (já classificadas). Dependendo do valor de K, podemos ter resultados diferentes em cada classificação. Sendo assim, um K muito pequeno pode gerar um problema de overfitting pois poucas amostras podem deixar o modelo muito sensível as regiões mais próximas. Além disso, temos também mais chances de sofrer influência de ruídos.

b) O que pode acontecer de indesejado quando k tende a ser muito grande, tendendo ao número de exemplos do conjunto de dados. (1,0)

Resposta:

Para a situação em que temos um K muito grande, podemos ter um problema de underfitting. A algoritmo não consegue encontrar relações fortes e significativas entre as amostras para atribuir uma classe para uma nova amostra. Como temos distâncias maiores consideradas, a incerteza também aumenta. Além disso, os pontos podem sofrer mais influência das classes majoritárias, aumentando também o viés.

c) Por que o algoritmo kNN é considerado lazy (preguiçoso). (0,5)

Resposta:

No geral, ele não gera um modelo propriamente dito para realizar a classificação, ele segue um processo com passos bem definidos que acontecem no momento em que é requisitado a classificação de uma nova amostra. Por este motivo ele é denominado lazy.

Questão 2 (2,5)

O algoritmo k-médias é um método de agrupamento particional. Uma das principais escolhas que devem ser feitas é o número k de agrupamentos que serão formados. Sobre o valor de k, o que é possível afirmar quando:

a) O valor de k é grande, tendendo ao número de exemplos do conjunto de dados? (1,0)

Resposta:

O algoritmo do K-médias escolhe K objetos como centros de agrupamento e vai classificando os demais a partir desses centróides. Para um K tão grande quanto o número de conjunto de dados, teremos praticamente uma partição por amostra, diminuindo assim a semelhança entre os objetos e gerando um número muito elevado de partições.

b) O valor de k é pequeno, tendendo a $k = 1$? (1,0)

Resposta:

Todos os elementos serão agrupados na mesma partição, ou seja, não haverá distinção entre os dados, todos serão considerados semelhantes entre si, perdendo um pouco o conceito de classificar os objetos em grupos.

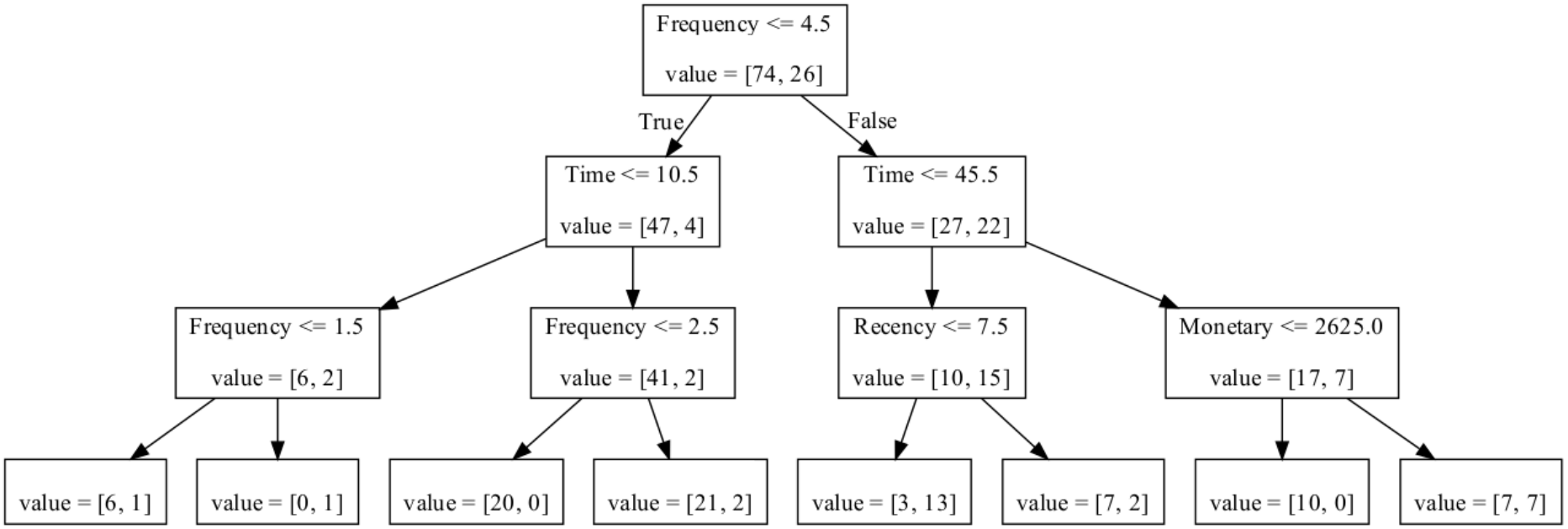
c) Cite um método que pode ser utilizado para escolha do valor de k visto em aula. (0,5)

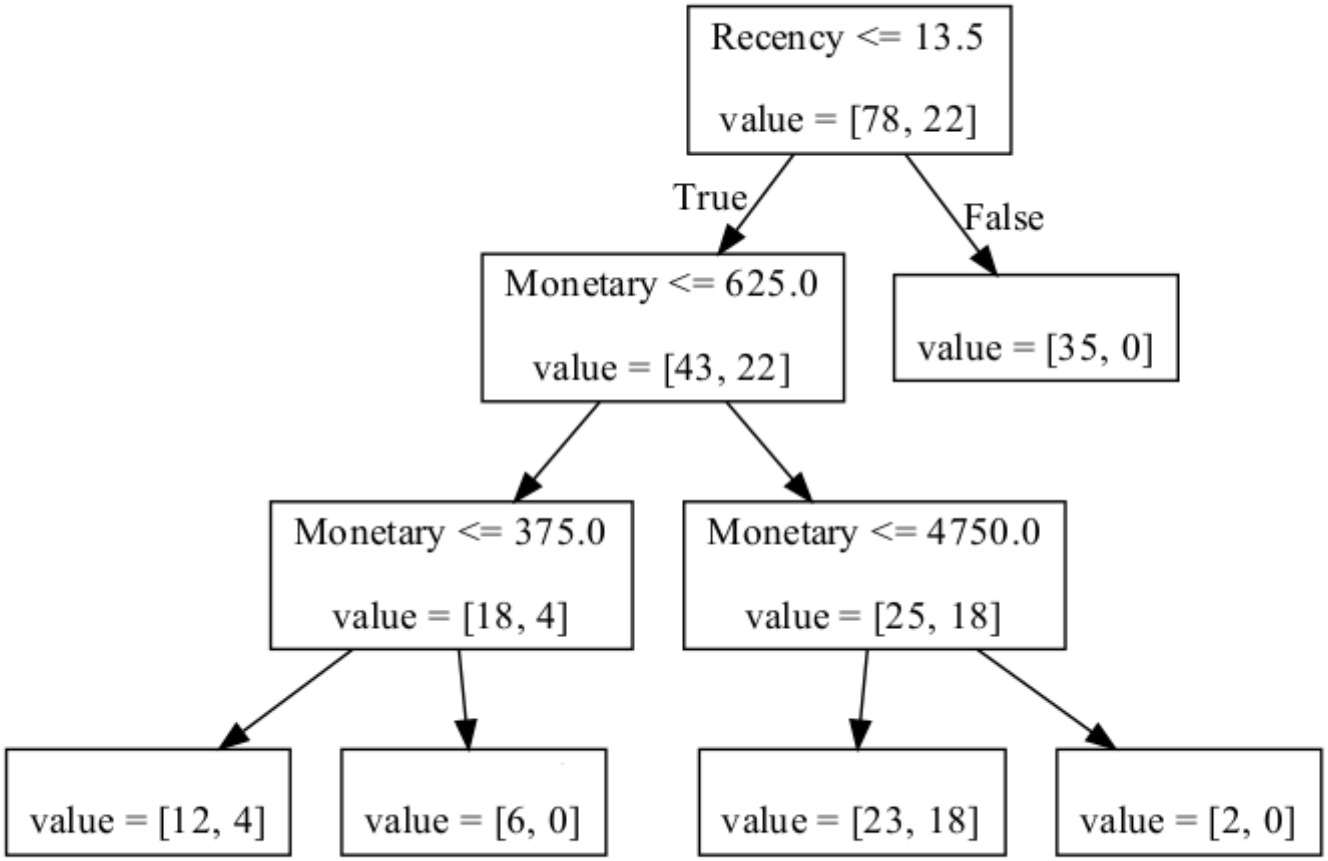
Resposta:

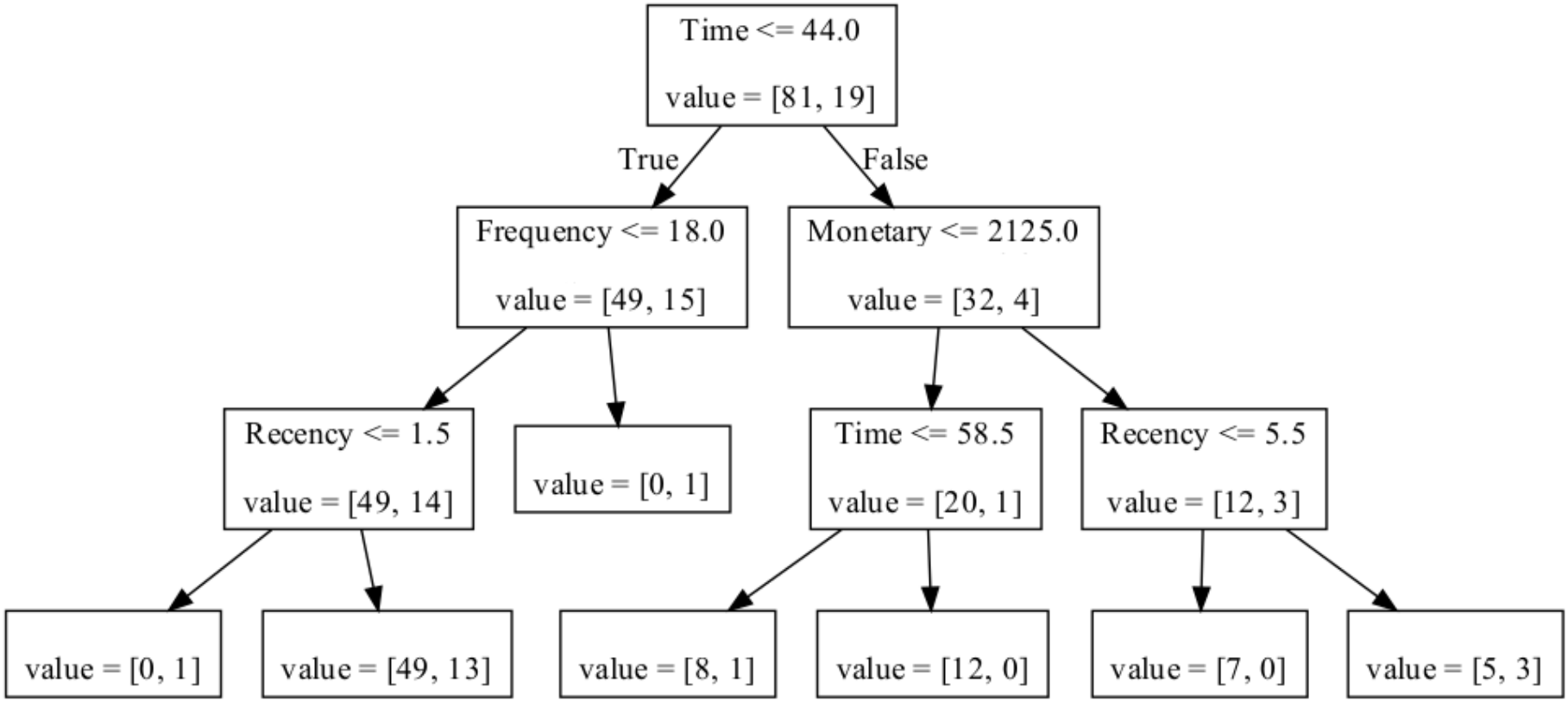
Método de cotovelo (Elbow method). Na prática é possível traçar um gráfico do percentual de variância explicada em função no número de clusters. Haverá um ponto ótimo nesta curva, onde teremos um maior ganho nessa relação de percentual de variação e número de clusters, achando assim o "joelho da curva" e em consequência uma sugestão otimizada para o valor de K.

Questão 3. (2,5)

Considere as representações gráficas de três modelos de árvore de classificação para um problema de classificação binária a seguir, respectivamente as árvores **A**, **B** e **C**. Em cada nó, observamos o *array* no formato `value = [<n_classe_1>, <n_classe_2>]`, nos quais `n_classe_1` representa o número de exemplos pertencentes a classe 1 naquele nó e `n_classe_2` o número de exemplos pertencentes à classe 2 naquele nó. Em nós não-folha, um campo extra é apresentando, simbolizando uma condição encontrada pelo algoritmo de árvore, no formato `<atributo_preditivo> <operador> <valor>`, no qual `atributo_preditivo` representa um atributo preditivo da base de dados, `operador` representa um possível operador como `<, <=, >, >=, =` e `valor` representa um valor a ser comparado com o `atributo_preditivo` usando `operador`. Caso a condição seja **verdadeira**, segue-se o caminho à **esquerda**, e caso seja **falsa**, segue-se o caminho da **direita**. Considere que cada árvore classifica um exemplo de acordo com a classe com maior quantidade de exemplos no nó folha.




Recency <= 13.5
value = [78, 22]
True
False
Monetary <= 625.0
value = [43, 22]
value = [35, 0]
Monetary <= 375.0
value = [18, 4]
Monetary <= 4750.0
value = [25, 18]
value = [12, 4]
value = [6, 0]
value = [23, 18]
value = [2, 0]



a) Considere que as árvores **A**, **B** e **C** formam um *ensemble* cuja combinação é feita por **voto majoritário**. Use o ensemble para classificar os exemplos apresentados na tabela a seguir. Qual a predição para cada exemplo? Qual o valor da acurácia? Qual o valor da acurácia balanceada? (1,0)

Recency	Frequency	Monetary	Time	Class
0.0	13.0	3250.0	28.0	2
1.0	24.0	6000.0	77.0	1
23.0	2.0	500.0	38.0	1
21.0	2.0	500.0	52.0	1
2.0	20.0	5000.0	45.0	2

Resposta:

Recency	Frequency	Monetary	Time	Class	A	B	C	Resultado
0	13	3250	28	2	2	1	2	2
1	24	6000	77	1	X	1	1	1
23	2	500	38	1	1	1	2	1
21	2	500	52	1	1	1	1	1
2	20	0	45	2	2	1	1	1

Matriz confusão:

		Valor Predito	
		1	2
Valor Atual	1	3 TP	0 FN
	2	1 FP	1 TN

Predição:

resultado = [2, 1, 1, 1, 1]

Acurácia:

ac = (TP + TN) / (TF + FP + TN + FN) = predições corretas / total de predições

ac = (3+1)/(3+1+1) = 4/5 = 0.8 = 80%

Acurácia balanceada:

acb = 0.5 * [(TP/(TP+FN)) + (TN/(TN+FP))]

acb = 0.5 * [(3/3+0) + (1/1+1)] = 0.5 * 1.5 = 0.75 = 75%

b) Escolha uma das árvores (**A**, **B** ou **C**). Indique a árvore escolhida e calcule a impureza de **todos** os nós-folhas usando *Gini*. (1,0)

In [65]:
```
#criando função para calcular o Gini:
def gini(c1,c2):
    pc1 = c1/(c1+c2)
    pc2 = c2/(c1+c2)
    gini = 1 - pc1**2 - pc2**2
    return '%.2f' % gini, '%.2f' % pc1, '%.2f' % pc2

print('ÁRVORE C')
print('Folha 1:',gini(0,1))
print('Folha 2:',gini(49,13))
print('Folha 3:',gini(0,1))
print('Folha 4:',gini(8,1))
print('Folha 5:',gini(12,0))
print('Folha 6:',gini(7,0))
print('Folha 7:',gini(5,3))
```

```
ÁRVORE C
Folha 1: ('0.00', '0.00', '1.00')
Folha 2: ('0.33', '0.79', '0.21')
Folha 3: ('0.00', '0.00', '1.00')
Folha 4: ('0.20', '0.89', '0.11')
Folha 5: ('0.00', '1.00', '0.00')
Folha 6: ('0.00', '1.00', '0.00')
Folha 7: ('0.47', '0.62', '0.38')
```

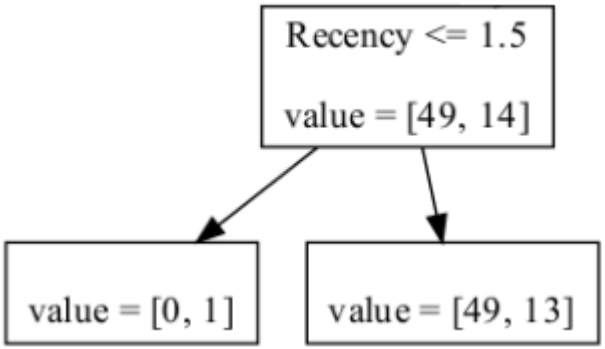
Resposta:

Gini = 1 - SOMA { [P(i/v)]^2 }

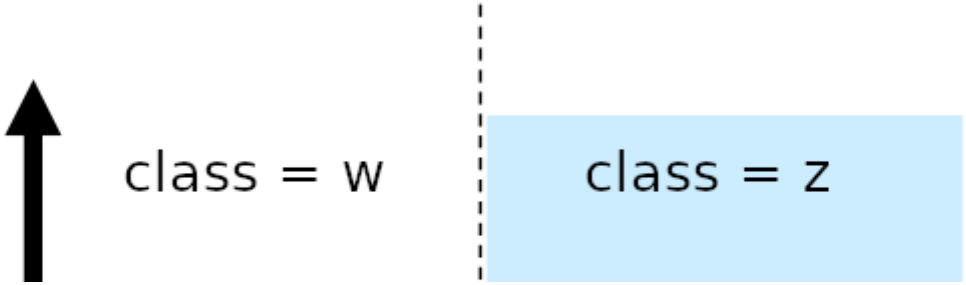
Considerando a árvore C:

Árvore C	Folha1	Folha2	Folha3	Folha4	Folha5	Folha6	Folha7
C1	0	49	0	8	12	7	5
C2	1	13	1	1	0	0	3
P(C1)	0.00	0.79	0.00	0.89	1.00	1.00	0.62
P(C2)	1.00	0.21	1.00	0.11	0.00	0.00	0.38
Gini	0.00	0.33	0.00	0.20	0.00	0.00	0.47

c) Considere a sub-árvore apresentada a seguir. (0,5)



Considere a figura abaixo que representa o corte no espaço de atributos que apenas a sub-árvore apresentada faz. Quais os valores de **X**, **Y**, **z** e **w** para que a figura esteja correta?



Resposta: X = ; Y = ; z = ; w =

```
X = Recency (atributo preditivo)
Y = 1.5     (valor da condição do atributo preditivo)
z = 1       (classe 1)
w = 2       (classe 2)
```

Questão 4 (2,5)

Dadas as características do conjunto de dados a seguir, as próximas questões se referem a escolha de um projeto de experimento adequado. Considere que os conjuntos de dados possuem dados suficientes.

Características do conjunto de dados:

- Tamanho: 3 atributos preditivos e 500 instâncias
- Atributo 1: numérico, com valores entre (-2,2)
- Atributo 2: numérico, com valores entre (0, 10000)
- Atributo 3: categórico, com valores pertencentes ao conjunto {'A', 'B', 'C'}
- Tarefa: Classificação
- Número de classes: 2
- Proporção entre as classes: 1:49

Responda as seguintes questões:

a) Quais dos seguintes particionamentos de dados você consideraria para avaliar um algoritmo de classificação qualquer: hold-out, validação cruzada estratificada, ou leave-one-out? Justifique. (1,0)

Resposta:

```
Considerando as características dos dados, temos classes desbalanceadas. Neste caso, a validação cruzada estratificada é ma
is adequada, pois ela permite criarmos partições que compensem o desbalanceamento das classes, diferentemente dos   outros
 métodos.
```

b) Quais das seguintes medidas de avaliação você consideraria: acurácia ou acurácia balanceada? Justifique. (0,5)

Resposta:

Acurácia balanceada, pois apenas a acurácia não é adequada para dados desbalanceados, pois pode prejudicar o modelo pri
vilegiando a classe majoritária.

c) Considerando apenas as características apresentadas da base, você consideraria algum tipo de pré-processamento quando avaliando *k-nn*? Se sim, qual(is)? Justifique. (1,0):

Resposta:

Como o KNN é um método baseado em distâncias, devemos tomar cuidado com as escalas entre os atributos. Verificando que temo
s escalas muito diferentes entre os atributos 1 e 2, consideraria realizar uma normalização (intervalo entre 0 e 1) e padro
nização. Para a variável categórica, consideraria aplicar uma transformação dos dados para classificação binária como o one
-hot enconding para podermos trabalhar com dados numéricos nos cálculos de distâncias sem maiores impactos no modelo.

Questão 5 (1,0)

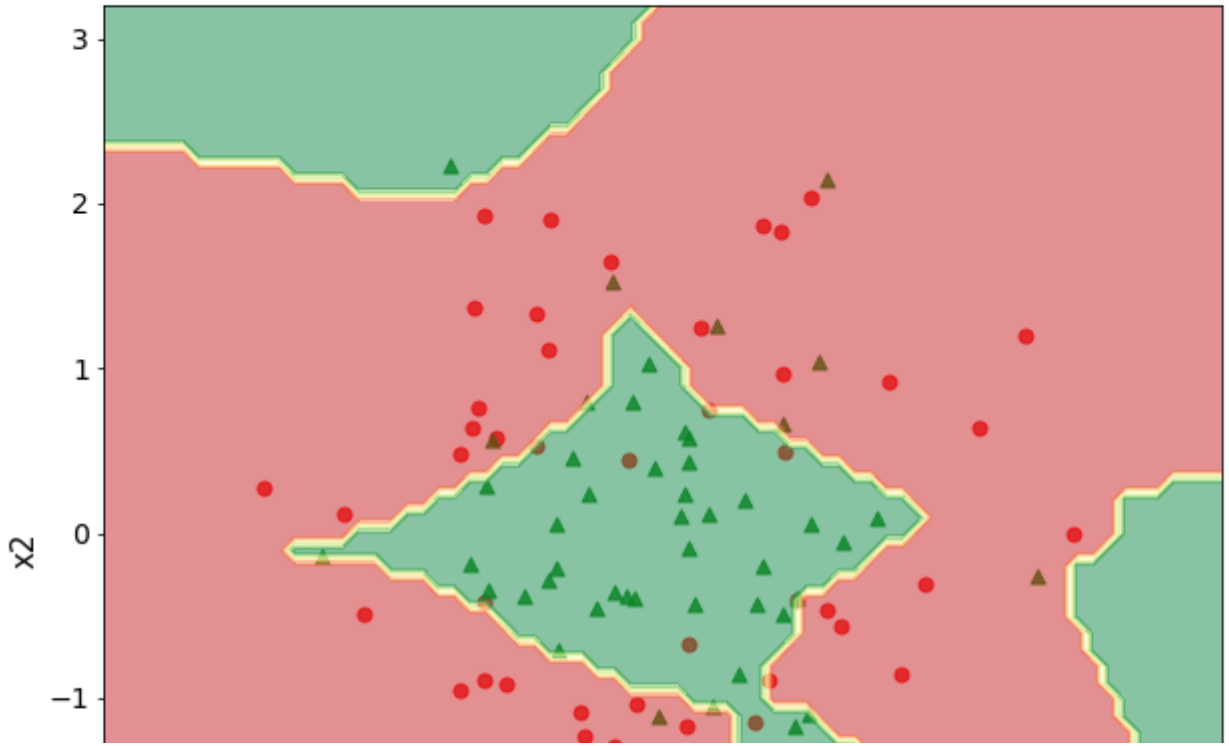
As imagens a seguir são resultados da aplicação das redes neurais perceptron com múltiplas camadas (MLP) em um conjunto de dados de classificação. Sobre os conceitos de underfitting e overfitting, diga o que está ocorrendo intuitivamente em cada imagem.

a) Underfitting ou overfitting? (0,5)

Resposta:

Underfitting

b) Underfitting ou overfitting? (0,5)



Resposta:

Overfitting