

Nome: ALVARO MAIA CHAVES

Pontuação Total da Avaliação: 7.60 pontos

=====

Correção da Questão 1:

Pergunta: 1a) Uma empresa está desenvolvendo um sistema para classificar mensagens recebidas como "Urgente" ou "Não Urgente" com base nas palavras presentes na mensagem. Foi analisado um conjunto de 100 mensagens, e os dados a seguir foram coletados: Mensagens Urgentes: 30; Mensagens Não Urgentes: 70; Palavra Presente "imediatamente": 15 (Mensagens Urgentes) e 5 (Mensagens Não Urgentes); Palavra Presente "problema": 10 (Mensagens Urgentes) e 10 (Mensagens Não Urgentes); Palavra Presente "atraso": 8 (Mensagens Urgentes) e 12 (Mensagens Não Urgentes). Calcule as probabilidades de uma mensagem ser "Urgente" e "Não Urgente" com base no conjunto de dados fornecido. Adicionalmente, determine as probabilidades condicionais para cada palavra ("imediatamente", "problema" e "atraso") em relação às mensagens "Urgentes" e "Não Urgentes".

=====Rubrica(s)

('Acertar a resposta aproximada de  $P(\text{"imediatamente"}|\text{"urgente"})=15/30=0.5'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"não urgente"})=70/100=0.70'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"problema"}|\text{"urgente"})=10/30\sim0.33'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"atraso"}|\text{"não urgente"})=12/70\sim0.17'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"atraso"}|\text{"urgente"})=8/30\sim0.27'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"urgente"})=30/100=0.30'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"imediatamente"}|\text{"não urgente"})=5/70\sim0.07'$ , 0.32)

('Acertar a resposta aproximada de  $P(\text{"problema"}|\text{"não urgente"})=10/70\sim0.14'$ , 0.32)

=====Resposta do estudante da questão 1:

1a)  $p(U) = 30/100 = 0,3$

$p(\bar{U}) = 70/100 = 0,7$

$P(I|U) = 15/0,3 = 50\%$

$P(I|\bar{U}) = 5/0,7 = 7,14\%$

$P(P|U) = 10/0,3 = 33,3\%$

$P(P|\bar{U}) = 10/0,7 = 14,2\%$

$P(A|U) = 8/0,3 = 26,6\%$

$P(A|\bar{U}) = 12/0,7 = 17,14\%$

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante está próxima da solução correta, mas apresenta alguns erros de cálculo. As probabilidades de uma mensagem ser "Urgente" ( $p(U)$ ) e "Não Urgente" ( $p(NU)$ ) foram calculadas corretamente como 0,3 e 0,7, respectivamente. No entanto, ao calcular as probabilidades condicionais para cada palavra, o estudante cometeu erros. Por exemplo, a probabilidade condicional de uma mensagem conter a palavra "imediatamente" dado que é "Urgente" ( $P(I|U)$ ) foi calculada como  $15/0,3$ , o que é incorreto because a probabilidade deve ser calculada com base no número total de mensagens urgentes (30), não na probabilidade de uma mensagem ser urgente (0,3). Portanto,  $P(I|U) = 15/30 = 0,5$ .

=====Feedback:

Para melhorar a resposta, o estudante deve calcular as probabilidades condicionais corretas. Além disso, é fundamental entender que as probabilidades condicionais devem ser calculadas com base nos dados fornecidos e não nas probabilidades anteriores. Por exemplo,  $P(I|U) = 15/30 = 0,5$ ,  $P(P|U) = 10/30 = 0,33$ ,  $P(A|U) = 8/30 = 0,27$ ,  $P(I|NU) = 5/70 = 0,07$ ,  $P(P|NU) = 10/70 = 0,14$  e  $P(A|NU) = 12/70 = 0,17$ . O estudante também deve verificar se as respostas estão de acordo com as rubricas fornecidas para garantir que estão alcançando os critérios de avaliação.

=====Pontuação:

2.1

=====

Correção da Questão 2:

Pergunta: 1b) Uma empresa está desenvolvendo um sistema para classificar mensagens recebidas como "Urgente" ou "Não Urgente" com base nas palavras presentes na mensagem. Foi analisado um conjunto de 100 mensagens, e os dados a seguir foram coletados: Mensagens Urgentes: 30; Mensagens Não Urgentes: 70; Palavra Presente "imediatamente": 15 (Mensagens Urgentes) e 5

(Mensagens Não Urgentes); Palavra Presente "problema": 10 (Mensagens Urgentes) e 10 (Mensagens Não Urgentes); Palavra Presente "atraso": 8 (Mensagens Urgentes) e 12 (Mensagens Não Urgentes). Suponha que uma nova mensagem contenha as palavras "imediatamente" e "problema". Calcule a probabilidade de ser uma mensagem "Urgente" e de ser "Não Urgente" utilizando o teorema de Bayes e classifique a mensagem como "Urgente" ou "Não Urgente".

=====Rubrica(s)

('Calcular a probabilidade  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = (P(\text{"Não urgente"}) * P(\text{"imediatamente"}|\text{"Não urgente"}) * P(\text{"problema"}|\text{"Não urgente"})) / (P(\text{"problema"}) * P(\text{"imediatamente"}))$ , podendo simplificar para  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = P(\text{"Não urgente"}) * P(\text{"imediatamente"}|\text{"Não urgente"}) * P(\text{"problema"}|\text{"Não urgente"}) \sim 0.70 * 0.07 * 0.014 \sim 0.001$ ', 1.0)

('Calcular a probabilidade  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = (P(\text{"urgente"}) * P(\text{"imediatamente"}|\text{"urgente"}) * P(\text{"problema"}|\text{"urgente"})) / (P(\text{"problema"}) * P(\text{"imediatamente"}))$ , podendo simplificar para  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = P(\text{"urgente"}) * P(\text{"imediatamente"}|\text{"urgente"}) * P(\text{"problema"}|\text{"urgente"}) \sim 0.30 * 0.5 * 0.33 \sim 0.05$ ', 1.0)

('Classificar o e-mail como "URGENTE" baseado no cálculo do máximo argumento entre  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"})$  e  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"})$ ', 0.5)

=====Resposta do estudante da questão 2:

1b)  $P(U|I\&P) = [P(U) * P(I|U)] + [P(U) * P(P|U)] = 0,3 * 50 + 0,3 * 33,3 = 24,99\%$

$P(\bar{A}|\bar{U}|I\&P) = [P(\bar{A}|\bar{U}) * P(I|\bar{A}|\bar{U})] + [P(U) * P(P|\bar{A}|\bar{U})] = 0,7 * 7,14 + 0,7 * 14,2 = 14,93\%$

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante não seguiu corretamente o teorema de Bayes para calcular as probabilidades. De acordo com as rubricas fornecidas, o cálculo para  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"})$  e  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"})$  deve considerar a probabilidade a priori de cada categoria (Urgente ou Não Urgente) e as probabilidades condicionais das palavras "imediatamente" e "problema" dadas essas categorias.

=====Feedback:

Para calcular  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"})$ , primeiro encontre as probabilidades condicionais:  $P(\text{"imediatamente"}|\text{"Urgente"}) = 15/30 = 0,5$  e  $P(\text{"problema"}|\text{"Urgente"}) = 10/30 = 0,33$ . Para  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"})$ , as probabilidades condicionais são:  $P(\text{"imediatamente"}|\text{"Não Urgente"}) = 5/70 \sim 0,07$  e  $P(\text{"problema"}|\text{"Não Urgente"}) = 10/70 \sim 0,14$ . O cálculo correto seria  $P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = [P(\text{"Urgente"}) * P(\text{"imediatamente"}|\text{"Urgente"}) * P(\text{"problema"}|\text{"Urgente"})] / [P(\text{"imediatamente"}) * P(\text{"problema"})]$  e  $P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = [P(\text{"Não Urgente"}) * P(\text{"imediatamente"}|\text{"Não Urgente"}) * P(\text{"problema"}|\text{"Não Urgente"})] / [P(\text{"imediatamente"}) * P(\text{"problema"})]$ , considerando a normalização pela probabilidade conjunta das palavras "imediatamente" e "problema".

=====Pontuação:

1.2

=====

Correção da Questão 3:

Pergunta: 2a) Árvores de decisão são modelos amplamente utilizados para classificação e regressão. Explique como o conceito de ganho de informação é utilizado na construção de uma árvore de decisão. Detalhe mostrando o uso do cálculo de entropia e ganho de informação em um problema hipotético.

=====Rubrica(s)

('Demonstrar o cálculo de entropia final como sendo  $H(\text{"depois da partição"}) = -P(\text{"amostras irem para conjunto 1"}) * \log_2(P(\text{"conjunto 1 após partição"})) - P(\text{"amostras irem para conjunto 2"}) * \log_2(P(\text{"conjunto 2 após partição"}))$ ', 1.0)

('Demonstrar o cálculo de entropia inicial como sendo  $H(\text{"antes da partição"}) = -p(\text{"classe1"}) * \log_2(P(\text{"classe1"})) - \dots - p(\text{"classeN"}) * \log_2(P(\text{"classeN"}))$ ', 1.0)

('Explicar que o conceito de Information Gain (IG) como sendo  $IG = H(\text{"antes da partição"}) - H(\text{"depois da partição"})$  é essencial para decidir qual atributo/partição escolher a cada nível da árvore', 1.0)

=====Resposta do estudante da questão 3:

2a) O ganho de informa

seja, será escolhido de acordo com o cálculo a variável que naquele momento mostra um ganho de informação, como exemplo simples considere,

Uma tabela de disponibilidade de dias para se jogar vôlei onde nas colunas contem os dados:

-tempo: nublado, chuvoso, limpo, nublado, limpo, limpo, nublado, chuvoso

-joga: joga, não joga, joga, joga, não joga, não joga, joga, joga

Agora temos que calcular a entropia principal ou seja da classe joga ou não joga:

$$h(joga) = -p(joga) \cdot \log_2(p(joga)) - p(\text{não joga}) \cdot \log_2(p(\text{não joga}))$$

$$h(\text{não joga}) = -p(\text{não joga}) \cdot \log_2(p(\text{não joga})) - p(joga) \cdot \log_2(p(joga))$$

contando que temos uma proporção para quem joga = 5/3 e para quem não joga = 3/5

Substituindo,

$$h(joga) = -5/3 \cdot \log_2(5/3) - 3/3 \cdot \log_2(3/3)$$

$$h(joga) = -1,666 \cdot 0,736 - 1,66 \cdot 0,736$$

$$h(joga) = -2,452$$

$$h(\text{não joga}) = -3/5 \cdot \log_2(3/5) - 2/5 \cdot \log_2(2/5)$$

$$h(\text{não joga}) = -0,6 \cdot 0,736 - 0,6 \cdot 0,736$$

$$h(\text{não joga}) = 0,8832$$

$$h(joga|\text{não joga}) = -h(joga) - h(\text{não joga})$$

$$h(joga|\text{não joga}) = 2,452 - 0,8832$$

Assim a entropia de jogar fica,

$$h(joga|\text{não joga}) = 1,568$$

Agora para ver qual atributo mostra um ganho de informação, nesse caso tempo:

$$h(\text{nublado}|joga) = -p(\text{nublado}|joga) \cdot \log_2(p(\text{nublado}|joga))$$

$$-p(\text{nublado}|\text{não joga}) \cdot \log_2(p(\text{nublado}|\text{não joga}))$$

$$h(\text{nublado}|joga) = -3/5 \cdot \log_2(3/5) - 0/3 \cdot \log_2(0/3)$$

$$h(\text{nublado}|joga) = -0,6 \cdot 0,736$$

$$h(\text{nublado}|joga) = 0,4416$$

$$h(\text{chuvoso}|joga) = -p(\text{chuvoso}|joga) \cdot \log_2(p(\text{chuvoso}|joga))$$

$$-p(\text{chuvoso}|\text{não joga}) \cdot \log_2(p(\text{chuvoso}|\text{não joga}))$$

$$h(\text{chuvoso}|joga) = -1/5 \cdot \log_2(1/5) - 1/3 \cdot \log_2(1/3)$$

$$h(\text{chuvoso}|joga) = -0,2 \cdot 2,32 - 0,33 \cdot 1,58$$

$$h(\text{chuvoso}|joga) = 0,464 + 0,52$$

$$h(\text{chuvoso}|joga) = 0,984$$

$$h(\text{limpo}|joga) = -p(\text{limpo}|joga) \cdot \log_2(p(\text{limpo}|joga)) - p(\text{limpo}|\text{não joga}) \cdot \log_2(p(\text{limpo}|\text{não joga}))$$

$$h(\text{limpo}|joga) = -1/5 \cdot \log_2(1/4) - 2/3 \cdot \log_2(2/3)$$

$$h(\text{limpo}|joga) = -0,2 \cdot 2 - 0,66 \cdot 0,58$$

$$h(\text{limpo}|joga) = 0,4 + 1,24$$

$$h(\text{limpo}|joga) = 1,64$$

Agora comparamos com a entropia de jogar

$$h(joga) = 1,568 - h(\text{nublado}|joga) +$$

$$1,568 - h(\text{chuvoso}|joga)$$

$$1,568 - h(\text{limpo}|joga)$$

$$h(joga) = 1,568 - 0,4416 + 1,568 - 0,984 + 1,568 - 1,64$$

$$h(joga) = 1,126 + 0,584 - 0,072$$

No caso a coluna tempo tem uma influência naquela na escolha de jogar ou não

$$h(joga) = 1,638$$

Correção proposta pela IA:

Considerações:

A resposta do estudante demonstra uma compreensão básica do conceito de ganho de informação e entropia em árvores de decisão. No entanto, há alguns erros e falta de clareza em alguns cálculos e explicações. O estudante começa bem calculando a entropia inicial, mas em seguida, apresenta cálculos que não seguem a fórmula correta para o ganho de informação. Além disso, a explicação sobre como o conceito de Information Gain é essencial para decidir qual atributo/partição escolher a cada nível da árvore pode ser mais clara.

Feedback:

Para melhorar, o estudante deve revisar a fórmula correta para o cálculo de entropia e ganho de informação. A entropia inicial deve ser calculada como  $H(\text{"antes da partição"}) =$

$-p(\text{"classe1"}) \cdot \log_2(P(\text{"classe1"})) - \dots - p(\text{"classeN"}) \cdot \log_2(P(\text{"classeN"}))$ ). Em seguida, o estudante deve calcular a entropia final após a partição como  $H(\text{"depois da partição"}) = P(\text{"amostras irem para conjunto 1"}) \cdot H(\text{"conjunto 1 após partição"}) + P(\text{"amostras irem para conjunto 2"}) \cdot H(\text{"conjunto 2 após partição"})$ . O ganho de informação é então calculado como  $IG = H(\text{"antes da partição"}) - H(\text{"depois da partição"})$ . Além disso, o estudante deve explicar claramente como o conceito de Information Gain é utilizado para decidir qual atributo/partição escolher a cada nível da árvore.

=====Pontuação:

1.8

=====

Correção da Questão 4:

Pergunta: 2b) Árvores de decisão são modelos amplamente utilizados para classificação e regressão. Suponha que você está desenvolvendo um sistema de suporte à decisão, aponte dois critérios de parada que você utilizaria na construção da árvore de decisão para garantir um modelo eficiente?

=====Rubrica(s)

('Mencionar e/ou explicar dois apenas um critério de parada válido na construção de árvores de decisão.', 1.25)

('Mencionar e/ou explicar dois ou mais critérios de parada válidos na construção de árvores de decisão.', 2.5)

('Não mencionar e/ou explicar qualquer critério de parada na construção de árvores de decisão.', 0.0)

=====Resposta do estudante da questão 4:

2b) O tamanho máximo da árvore para que não exceda e comece a gerar dados errôneos, e limitar o número de atributos por nó para que se alcance um tamanho uniforme em cada nó.

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante menciona dois critérios de parada para a construção de árvores de decisão, que são o tamanho máximo da árvore e a limitação do número de atributos por nó. No entanto, a explicação fornecida é básica e poderia ser mais detalhada.

=====Feedback:

Para atingir a nota máxima, o estudante poderia ter explicado melhor como esses critérios de parada contribuem para evitar o overfitting e garantir um modelo eficiente. Além disso, poderia ter mencionado outros critérios de parada importantes, como a profundidade máxima da árvore ou o número mínimo de amostras por nó. Uma explicação mais detalhada sobre como esses critérios são implementados e ajustados seria benéfica.

=====Pontuação:

2.5

=====