

Nome: FRANCISCO LUCAS BENVINDO DA SILVA

Pontuação Total da Avaliação: 7.72 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) $P(\text{"Urgente"}) = 30/100 = 0.3$

$P(\text{"NÃO urgente"}) = 70/100 = 0.7$

$P(\text{"imediatamente"} | \text{"Urgente"}) = 15/30 = 0.5$

$P(\text{"problema"} | \text{"Urgente"}) = 10/30 = 0.33$

$P(\text{"atraso"} | \text{"Urgente"}) = 8/30 = 0.27$

$P(\text{"imediatamente"} | \text{"NÃO urgente"}) = 5/70 = 0.07$

$P(\text{"problema"} | \text{"NÃO urgente"}) = 10/70 = 0.14$

$P(\text{"atraso"} | \text{"NÃO urgente"}) = 12/70 = 0.17$

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante apresenta várias incorreções em relação às probabilidades condicionais. Além disso, o cálculo das probabilidades anteriores de "Urgente" e "Não Urgente" está correto.

=====Feedback:

Foram identificados pontos fortes em relação às probabilidades anteriores e pontos fracos nas probabilidades condicionais para as palavras presentes nas mensagens. Para atingir a nota máxima, é necessário rever os cálculos das probabilidades condicionais: $P(\text{"imediatamente"} | \text{"Urgente"}) = 15/30 = 0.5$, $P(\text{"problema"} | \text{"Urgente"}) = 10/30 = 0.33$, $P(\text{"atraso"} | \text{"Urgente"}) = 8/30 = 0.27$, $P(\text{"imediatamente"} | \text{"Não Urgente"}) = 5/70 = 0.07$, $P(\text{"problema"} | \text{"Não Urgente"}) = 10/70 = 0.14$ e $P(\text{"atraso"} | \text{"Não Urgente"}) = 12/70 = 0.17$.

=====Pontuação:

1.92

=====

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) $P(\text{"Urgente"} | \text{"imediatamente"} \& \text{"problema"}) = P(\text{"Urgente"}) * P(\text{"imediatamente"} | \text{"Urgente"}) * P(\text{"problema"} | \text{"Urgente"})$

Substituindo: $0.3 * 0.75 * 0.5 = 0.1125$

$P(\text{"Não Urgente"} | \text{"imediatamente"} \& \text{"problema"}) =$

$P(\text{"Não Urgente"}) * P(\text{"imediatamente"} | \text{"Não Urgente"}) * P(\text{"problema"} | \text{"Não Urgente"})$

Substituindo: $0.7 * 0.25 * 0.5 = 0.0875$

A mensagem pode ser classificada como urgente.

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante apresenta cálculos para as probabilidades de uma mensagem ser "Urgente" ou "Não Urgente" dado que contém as palavras "imediatamente" e "problema". No entanto, o estudante não aplicou corretamente o teorema de Bayes para calcular essas probabilidades condicionais. O teorema de Bayes exige a divisão pelo termo de evidência, que é a probabilidade de observar as palavras "imediatamente" e "problema" em qualquer mensagem, que pode ser calculada considerando todas as ocorrências das palavras em mensagens urgentes e não urgentes.

=====Feedback:

Um ponto forte da resposta é a intenção de classificar a mensagem com base nas probabilidades calculadas. No entanto, para melhorar, o estudante deve aplicar corretamente o teorema de Bayes, incluindo o cálculo do termo de evidência para normalizar as probabilidades. Além disso, é importante usar dados precisos para calcular as probabilidades anteriores e condicionais das palavras "imediatamente" e "problema" tanto em mensagens urgentes quanto em não urgentes.

=====Pontuação:

1.5

=====

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 $H(\text{"depois da partição"}) = P(\text{"amostras irem para conjunto 1"}) * H(\text{"conjunto 1 após partição"}) + P(\text{"amostras irem para conjunto 2"}) * H(\text{"conjunto 2 após partição"})$, 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 conceito de ganho de informação é utilizado em uma árvore de decisão para reduzir a entropia, ou grau de confusão nos dados. Ao selecionar determinada variável e realizar uma partição com nela, o ideal é que haja um ganho de informação e redução da entropia. Exemplo:

Dado um conjunto de dados tabular que registra a ocorrência de idas a praia com 12 linhas, e

com as variáveis independentes de temperatura, clima, força do vento, estado da maré entre outras e uma variável alvo "ida" que pode ser classificada como "foi" ou "não foi". Considere que dos 12 registros, 7 são classificados como "foi" e 5 como "não foi". O cálculo de entropia antes da tomada de qualquer decisão se da por:

Legenda:

f = foi

nf = não foi

$$-P(f) \cdot \log_2(P(f)) - P(nf) \cdot \log_2(P(nf)) = -7/12 \cdot \log_2(7/12) - 5/12 \cdot \log_2(5/12) = 0.455 + 0.527 = 0.982$$

Dado que por exemplo seleciona-se a variável de temperatura e cria-se uma partição com a condição temperatura > 28. Realiza-se novamente o cálculo de entropia, onde a temperatura foi maior que 28 e a ida foi classificada como "foi" e também como "não foi" o mesmo ocorre para quando a temperatura foi menor que 28. Ao final, somam-se as duas entropias e subtrai-se da entropia antes da tomada de qualquer decisão, calculando assim o ganho de informação

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. O estudante explica que o objetivo é reduzir a entropia ou o grau de confusão nos dados ao selecionar uma variável e realizar uma partição. Além disso, o estudante fornece um exemplo hipotético e calcula a entropia inicial com base nas probabilidades das classes "foi" e "não foi".

Feedback:

No entanto, a resposta do estudante carece de clareza e precisão em alguns pontos.

Primeiramente, o estudante não explicita claramente a fórmula para o cálculo da entropia final após a partição, que deve ser $H(\text{"depois da partição"}) = P(\text{"amostras irem para conjunto 1"}) \cdot H(\text{"conjunto 1 após partição"}) + P(\text{"amostras irem para conjunto 2"}) \cdot H(\text{"conjunto 2 após partição"})$. Além disso, o estudante não fornece uma explicação clara sobre como o conceito de Information Gain (IG) é utilizado para decidir qual atributo ou partição escolher a cada nível da árvore, que é essencial para a construção da árvore de decisão.

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) Utilizaria os critérios de parada de nível da árvore, determinando um nível máximo onde a árvore deve parar de expandir, evitando com que a árvore de decisão fique muito extensa e o critério da pureza ou homogeneidade dos dados, onde os dados de determinado não são da mesma classe.

Correção proposta pela IA:

Considerações:

A resposta do estudante menciona dois critérios de parada na construção da árvore de decisão: o nível da árvore e a pureza ou homogeneidade dos dados. Esses critérios são válidos e relevantes para garantir um modelo eficiente. No entanto, a explicação poderia ser mais detalhada e clara.

=====Feedback:

Um feedback detalhado seria que o estudante poderia ter expandido a explicação sobre como esses critérios funcionam na prática e como eles contribuem para a eficiência do modelo. Além disso, poderia ter mencionado outros critérios de parada, como o tamanho mínimo da amostra ou a profundidade máxima da árvore, para demonstrar uma compreensão mais ampla do tema.

=====Pontuação:

2.5

=====