

Nome: santiago

Pontuação Total da Avaliação: 6.78 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".
Diretriz(es): $\{('P("problema"|"não urgente")=10/70 \sim 0.14', 0.32), ('P("imediatamente"|"urgente")=15/30=0.5', 0.32), ('P("atraso"|"não urgente")=12/70 \sim 0.17', 0.32), ('P("urgente")=30/100=0.30', 0.32), ('P("não urgente")=70/100=0.70', 0.32), ('P("problema"|"urgente")=10/30 \sim 0.33', 0.32), ('P("atraso"|"urgente")=8/30 \sim 0.27', 0.32), ('P("imediatamente"|"não urgente")=5/70 \sim 0.07', 0.32)\}$

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

1A) Para calcular a probabilidade de uma mensagem ser urgente ou não urgente, segue aqui as probabilidades:

$P(\text{Urgente}) = 30/100$ (que é o total de mensagens urgentes sobre o nosso total de mensagens especificado na questão)

$P(\text{Não-Urgente}) = 70/100$ (que é o total de mensagens não urgentes sobre o nosso total de mensagens especificado na questão)

Após isso, temos as probabilidades condicionais:

$P(\text{Imediatamente} | \text{Urgente}) = 15/30$ (Total de mensagens que contem imediatamente sobre todas as palavras que são urgentes)

$P(\text{Imediatamente} | \text{Não-Urgente}) = 5/70$ (Total de mensagens que contem imediatamente sobre todas as palavras que não são urgentes)

$P(\text{Problema} | \text{Urgente}) = 10/30$

$P(\text{Problema} | \text{Não-Urgente}) = 10/70$

$P(\text{Atraso} | \text{Urgente}) = 8/30$

$P(\text{Atraso} | \text{Não-Urgente}) = 12/70$

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante está correta no cálculo das probabilidades iniciais de uma mensagem ser "Urgente" ou "Não Urgente". No entanto, para calcular as probabilidades condicionais corretamente, é importante garantir que as contas sejam precisas e que sejam considerados todos os cenários possíveis.

=====Feedback:

Pontos fortes da resposta do estudante incluem o cálculo correto das probabilidades iniciais de $P(\text{Urgente})$ e $P(\text{Não-Urgente})$. Entretanto, os cálculos das probabilidades condicionais estão corretos, mas poderiam ser melhor apresentados. Além disso, a resposta do estudante foi quase completa e alcançou boa parte dos requisitos solicitados.

=====Pontuação:

2.360888888888893

=====

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".

Diretriz(es): $\{(P(\text{"Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = (P(\text{"urgente"}) * P(\text{"imediatamente"}|\text{"urgente"}) * P(\text{"problema"}|\text{"urgente"})) / (P(\text{"problema"}) * P(\text{"imediatamente"}))\}$, 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), $(P(\text{"Não Urgente"}|\text{"imediatamente"}\&\text{"problema"}) = (P(\text{"Não urgente"}) * P(\text{"imediatamente"}|\text{"Não urgente"}) * P(\text{"problema"}|\text{"Não urgente"})) / (P(\text{"problema"}) * P(\text{"imediatamente"}))\}$, 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)}

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

1B) Para Urgente:

$$P(\text{Urgente} | \text{Imediatamente}, \text{Problema}) = (P(\text{Urgente}) * P(\text{Imediatamente}|\text{Urgente}) * P(\text{Problema}|\text{Urgente})) / P(\text{Problema}) * P(\text{Imediatamente})$$

Vamos calcular as probabilidades da seguinte maneira:

$P(\text{Problema}) = 20/100$ (Soma da quantidade mensagens que contém a palavra "problema" sobre total de mensagens especificado na questão)

$P(\text{Imediatamente}) = 20/100$ (Soma da quantidade mensagens que contém a palavra "imediatamente" sobre o total de mensagens especificado na questão)

$P(\text{Urgente}) = 30/100$ (que é o total de mensagens urgentes sobre o nosso total de mensagens especificado na questão)

$P(\text{Não-Urgente}) = 70/100$ (que é o total de mensagens não urgentes sobre o nosso total de mensagens especificado na questão)

$P(\text{Imediatamente} | \text{Urgente}) = 15/30$ (Total de mensagens que contem imediatamente sobre todas as palavras que são urgentes)

$P(\text{Imediatamente} | \text{Não-Urgente}) = 5/70$ (Total de mensagens que contem imediatamente sobre todas as palavras que não são urgentes)

$P(\text{Problema} | \text{Urgente}) = 10/30$

$P(\text{Problema} | \text{Não-Urgente}) = 10/70$

Simplificando o denominador e simplificando os valores fracionários, ficaria assim:

$$P(\text{Urgente} | \text{Imediatamente}, \text{Problema}) = (0.3 * 0.5 * 0.33) = 0.0495$$

Agora, realizando o cálculo para não-urgente:

$$P(\text{Não-Urgente} | \text{Imediatamente}, \text{Problema}) = (P(\text{Não-Urgente}) * P(\text{Imediatamente}|\text{Não-Urgente}) * P(\text{Problema}|\text{Não-Urgente})) / P(\text{Problema}) * P(\text{Imediatamente})$$

Simplificando o denominador novamente, e substituindo os valores, ficaria da seguinte maneira:

$$P(\text{Não-Urgente} | \text{Imediatamente}, \text{Problema}) = 0.7 * 0.07 * 0.14 = 0.00686$$

Ou seja, utilizando o teorema de Bayes podemos classificar a mensagem como Urgente, já que ao comparar as duas probabilidades, vemos que:

$$P(\text{Urgente} | \text{Imediatamente}, \text{Problema}) = 0.0495$$

$$P(\text{Não-Urgente} | \text{Imediatamente}, \text{Problema}) = 0.00686$$

$$P(\text{Urgente} | \text{Imediatamente}, \text{Problema}) > P(\text{Não-Urgente} | \text{Imediatamente}, \text{Problema})$$

$$0.0495 > 0.00686$$

Então, a mensagem seria classificada como urgente.

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante está próxima do correto, mas há alguns erros e simplificações excessivas. Ele aplicou corretamente o teorema de Bayes para calcular as probabilidades de uma mensagem ser urgente ou não urgente, dado que contém as palavras "imediatamente" e "problema". No entanto, na simplificação do denominador, o estudante não aplicou corretamente o teorema, ignorando a necessidade de considerar a probabilidade conjunta das palavras "imediatamente" e "problema" ocorrerem na mensagem.

=====Feedback:

O estudante fez um bom esforço em calcular as probabilidades condicionais e aplicar o teorema de Bayes. No entanto, para melhorar, é importante considerar a probabilidade conjunta das

palavras "imediatamente" e "problema" ocorrerem na mensagem, que é necessária para calcular o denominador corretamente. Além disso, é fundamental manter a precisão nos cálculos e não simplificar demais as frações. O estudante também poderia ter explicitado melhor os passos do cálculo e justificado a escolha dos parâmetros para o teorema de Bayes.

=====Pontuação:

1.66636363636365

=====

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.

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

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

2A) Para visualizarmos isso,

Hipoteticamente, digamos que tenhamos um problema: Uma tabela de informações contendo os jogos de vôlei que um grupo jogou ao decorrer da semana, e com isso eles queriam saber se é possível prever pelas partições dessas tabelas se haverá jogo ou não no dia.

Digamos que a nossa entropia inicial, a nossa primeira ramificação seja jogar ou não jogar. Se em 14 jogos no total, 7 jogamos e 7 não jogamos, vamos calcular a entropia:

A fórmula:

$$H(x) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

Onde os p_1 e p_2 se referem a probabilidade. Para ficar mais claro, vamos utilizar a fórmula com a partição unitária jogar.

$$P(\text{Joga} = \text{sim}) = 7/14 \text{ (Quantidade de vezes jogadas sobre o total de jogos, ser o nosso } p_1)$$

$$P(\text{Joga} = \text{não}) = 7/14 \text{ (Quantidade de vezes jogadas sobre o total de jogos, ser o nosso } p_2)$$

$$H(\text{Joga}) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 0.84$$

Agora, vamos selecionar uma partição (uma ou um conjunto de colunas da nossa tabela) para compor o cálculo de entropia e ganho

Digamos que iremos utilizar a coluna de Temperatura com 90°F para a nossa partição, então separaremos para o caso de Temperatura $> 90^\circ\text{F}$ e outro para Temperatura $\leq 90^\circ\text{F}$.

Vamos dizer que de acordo com nossa tabela, nos dias em que jogaram e a temperatura estava acima de 90°F foram 3

$$P(\text{Temperatura} > 90 \mid \text{Joga} = \text{Sim}) = 3/7 \text{ (Total de dias que a temperatura estava acima de } 90 \text{ sobre a quantidade de dias que teve jogo)}$$

$$P(\text{Temperatura} > 90 \mid \text{Joga} = \text{Não}) = 4/7$$

Vamos dizer que de acordo com nossa tabela, nos dias em que jogaram e a temperatura estava acima de 90°F foram 5

$$P(\text{Temperatura} \leq 90 \mid \text{Joga} = \text{Sim}) = 4/7 \text{ (Total de dias que a temperatura estava acima de } 90 \text{ sobre a quantidade de dias que teve jogo)}$$

$$P(\text{Temperatura} \leq 90 \mid \text{Joga} = \text{Não}) = 3/7$$

$$H(\text{Temperatura} > 90) = -0.42 \log_2(0.42) - 0.57 \log_2(0.57)$$

$$H(\text{Temperatura} \leq 90) = -0.57 \log_2(0.57) - 0.57 \log_2(0.42)$$

Após o computador realizar os cálculos, faremos o cálculo de Entropia para o conjunto S:

$$H(S|\text{Temperatura}) = 7/14 \text{ (Total de Joga sim dividido pelo total de itens do conjunto)} \cdot$$

$$H(\text{Temperatura} > 90) + 7/14 \text{ (Total de Joga não dividido pelo total de itens do conjunto)} \cdot$$

$$H(\text{Temperatura} \leq 90) = -0.57 \log_2(0.57) - 0.57 \log_2(0.42)$$

Após isso, precisamos calcular o ganho, que vai se dar pela subtração da entropia inicial, menos o resultado do $H(S|\text{Temperatura})$.

$$G(S|\text{Temperatura}) = 0.84 - H(S|\text{Temperatura})$$

Esse cálculo é importante para ramificar a nossa árvore e a quantidade do ganho vai ser referente ao quanto essa partição é importante para a resolução do nosso problema.

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante demonstra uma compreensão básica do conceito de entropia e ganho de informação em árvores de decisão, mas há pontos que precisam ser corrigidos e aprofundados para atender às diretrizes fornecidas. O estudante calculou corretamente a entropia inicial e tentou aplicar o conceito de ganho de informação para decidir a melhor partição. No entanto, a explicação poderia ser mais clara e detalhada, especialmente na aplicação da fórmula de ganho de informação e no cálculo da entropia após a partição.

=====Feedback:

Um ponto forte da resposta é a tentativa de aplicar o conceito de entropia e ganho de informação em um problema hipotético, o que é essencial para entender como esses conceitos são utilizados na construção de árvores de decisão. No entanto, para melhorar, o estudante precisa ser mais preciso nos cálculos e explicar claramente como o ganho de informação é calculado e como ele é usado para decidir a melhor partição. Além disso, a aplicação da fórmula de entropia após a partição precisa ser mais cuidadosa, considerando as probabilidades condicionais e os resultados nos conjuntos após a partição.

=====Pontuação:

1.7366972477064222

=====

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?

Diretriz(es): {'Apontar a limitação de número insuficiente de amostras no nó para realizar nova partição como um possível critério de parada.', 1.25), ('Apontar a limitação da quantidade de níveis da árvore (profundidade máxima atingida) como um possível critério de parada.', 1.25), ('Apontar a limitação de entropia das classes (labels) atingir zero no nó como um possível critério de parada.', 1.25), ('Apontar qualquer outra limitação plausível para que uma árvore de decisão não continue a realizar partições com base em um número de amostras presentes em um dado nó.', 1.25)}

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

2B) Citarei vários critérios interessantes para serem usados de parada para Árvore de decisão.

Quantidade de Filhos em um Nó: Podemos utilizar isso principalmente como ponto de parada, porque a quantidade de itens em um nó, além de gerar um gasto computacional alto, pode gerar uma quantidade de partições que podem prejudicar o nosso problema. Uma versão alternativa a essa seria também parar em quantidade x de nós em uma árvore, os motivos são os mesmos. Pureza dos Nós: Se em minha árvore, a entropia dos itens do nó está muito próxima, é um critério de parada porque cada item está muito similar, então as divisões dentro da árvore vão ficando redundante.

=====

Correção proposta pela IA:

=====

=====Considerações:

A resposta do estudante apresenta alguns critérios de parada interessantes para a construção de uma árvore de decisão, como a quantidade de filhos em um nó e a pureza dos nós. No entanto, esses critérios não estão alinhados com as diretrizes fornecidas.

=====Feedback:

Os principais pontos fortes da resposta são a menção à limitação da quantidade de filhos em um nó e a pureza dos nós. No entanto, para atingir a nota máxima, o estudante deveria ter mencionado critérios como a limitação do número insuficiente de amostras no nó, a profundidade máxima atingida e a limitação da entropia das classes atingir zero no nó. Além disso, o estudante poderia ter mencionado outras limitações plausíveis para evitar a continuação de partições com base em um número de amostras presentes em um nó.

=====Pontuação:

1.0138888888888888

=====