

Sophia Carrazza



Resumo

IA



FÓRMULAS

Entropia:

$$\text{Entropia}(S) = - \sum_{i=1}^n P_i \cdot \log_2 (P_i)$$

→ proporção de elementos em S pertencentes à classe i.

quanto menor for, melhor

$$\text{Entropia (atributo)} = \sum_{i=1}^v \frac{P_i + n_i}{P + n} \cdot I\left(\frac{P_i}{P_i + n_i}, \frac{n_i}{P_i + n_i}\right)$$

↑ exemplos positivos ↑ exemplos negativos

$$\text{Ganho}(\text{atributo}) = \text{entropia}(S) - \text{entropia}(\text{atributo})$$

quanto maior
for, melhor

↓
(classe)

Métricos de avaliação:

$$TVP = \frac{\text{Verdadeiros } X}{\text{Verdad. } X + \text{ Falsos } X} \quad \left. \begin{array}{l} \text{(olha pela)} \\ \text{limha} \end{array} \right\} \text{e} \quad \left. \begin{array}{l} \text{(é o mesmo)} \\ \text{qui o recall} \end{array} \right.$$

TFN = Complementar de TVP

TFP = Não são X e foram classificados como X → (coluna)
Soma de todos os demais classes → (nenhuma das linhas que é + fácil)

TVN = Complementor de TFP

$$\text{Precisão} = \frac{\text{Valeadores Positivos}}{\text{Valead. Positivos} + \text{Falso Positivo}}$$

} olha pela coluna P

$$\text{Recall} = \frac{\text{Valeadores Positivos}}{\text{Valead. Positivos} + \text{Falso Negativo}}$$

} olha pela linha R

$$\text{F1 score} = \frac{2 \cdot \text{recall} \cdot \text{precisão}}{\text{recall} + \text{precisão}}$$

$$\text{Acurácia} = \frac{\text{Valeadores Positivos} + \text{Valeadores Negativos}}{\text{VP} + \text{FN} + \text{VN} + \text{FP}} \quad (\text{total de instâncias})$$

Cálculo do Naive Bayes:

5 instâncias SIM

6 instâncias NÃO

situação dos atributos:

lua cheia

céu nublado

chão seco

→ método de classificação
→ cria uma matriz de probabilidades

SIM:

$$\frac{5}{11} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} = \text{parcial 1}$$

6 instâncias
de 11 não nimm

das 5 instâncias SIM,
2 tem lua cheia

* prob. de SIM = $\frac{\text{parcial 1}}{\text{parcial 1} + \text{parcial 2}}$

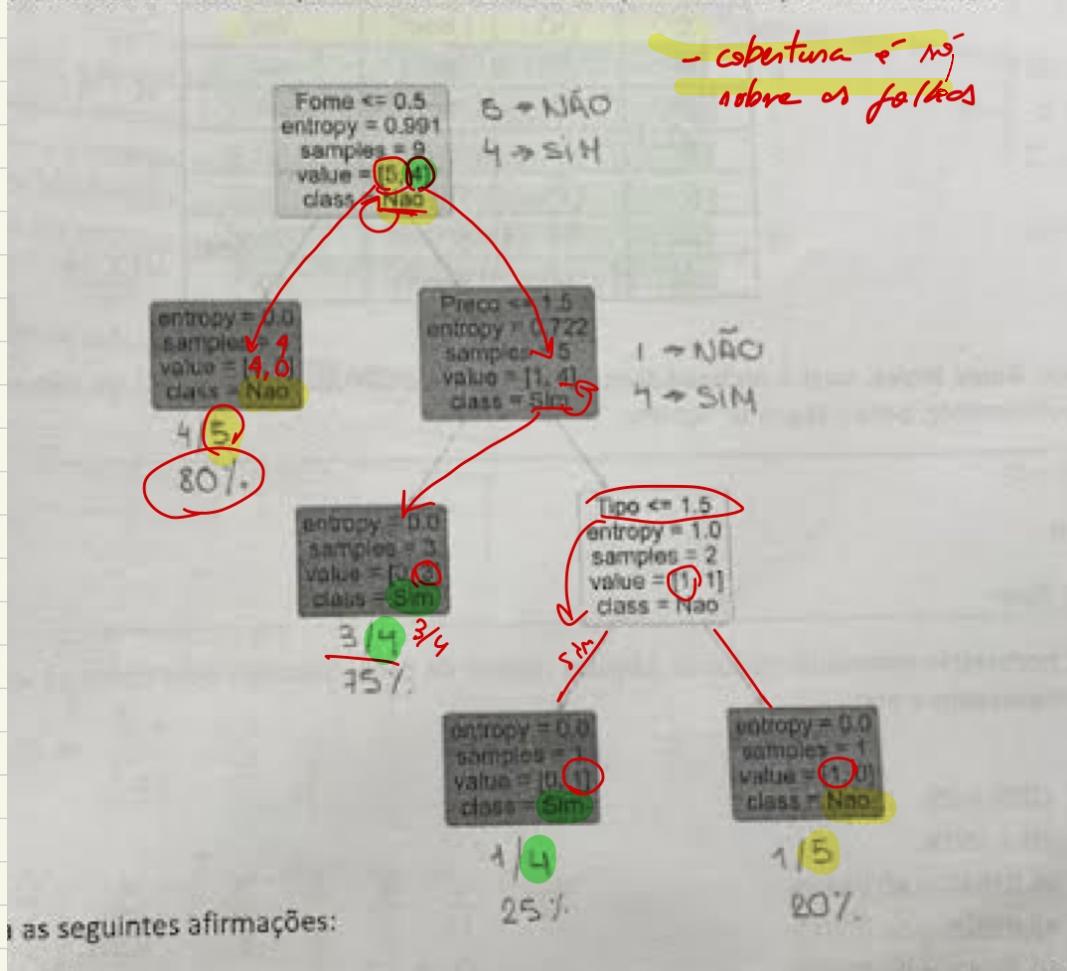
denominador p1 ou p2:
parcial 1 + parcial 2

$$\text{NÃO: } \frac{6}{11} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{2}{6} = \text{parcial 2}$$

* prob. de NÃO = $\frac{\text{parcial 2}}{\text{parcial 1} + \text{parcial 2}}$

Cálculo de cobertura por classe:

ore de decisão obtida a partir da base de dados 'Esperar ou não pelo Restaurante'.



nas seguintes afirmações:

* A qntd de folhas na árvore
é a qntd de regressão!
e a qntd de classificação!

ALGORITMOS

→ (nóção)

ID3 = é o algoritmo pioneiro em indução de árvores de decisão.

- ↳ ele escolhe, em cada etapa, o atributo que mais separa os dados usando o ganho de informação.
- ↳ começa c/ os dados na raiz da árvore e, a cada divisão, seleciona o melhor atributo para dividir os dados.
- ↳ ele só lida com valores discretos (não pode ser usado c/ atributos de valores contínuos). → númericos
- ↳ não tem mecanismos pós-poda (contra overfitting)

→ classificação nominal

CH.5 = é uma extensão do ID3 com uma série de melhorias.

- ↳ lida com atributos discretos e contínuos. Para os contínuos, cria um valor limiar e divide os dados em maior ou menor que esse valor.
- ↳ lida com atributos incompletos, com o "table" ? ou unknowns, que não faz contas c/ ele
- ↳ faz a poda de ramos depois da criação (troca ramificações que não ajudam nas decisões por nós folhas).
- ↳ utiliza o "gain ratio" ao invés do "ganho de informação".

→ gain
split info

CART = é usado tanto p/ tarefas de regressão quanto p/ de classificação

- ↳ constrói árvores binárias p/ determinar os divisões nos nós

↳ p/ classificação: índice de gini *

↳ p/ regressão: erro quadrático médio

* O índice de gini é usado p/ medir a impureza de um nó, relacionado à entropia de:

↓ gini ↑ concentração das classes

$$Gini = 1 - \sum_i (P_i)^2$$

↳ probabilidade de critério na V

retornar os classificadores da árvore

Random Forest = método de ensemble

↳ treina cada árvore c/ amostras geradas a partir do método de amostragem bootstrap

↳ p/ cada conjunto de instâncias, o RF considera n variáveis aleatórias do conjunto de dados.

↳ não faz poda

↳ p/ classificação: voto majoritário

↳ p/ regressão: média dos valores preditos

05-

PROCESSAMENTO

(etapas de pré-processamento)

PARTE 1. - balançamento da base de dados

↳ em alguns subconjuntos, seus dados aparecem com frequência maior que os dados das demais classes.

- oversampling causa:*
- o overfitting → quando o modelo é superapertado aos dados de treinamento
- undersampling causa:*
- o undersampling → quando o modelo não se ajusta aos dados de treinamento.

- ↳ quando alimentados c/ dados desbalanceados, os algoritmos tendem a favorecer a classificação de mais dados na classe majoritária.
- principais técnicas para resolver:
- ↳ redimensionar o tamanho do conjunto de dados (adiciona instâncias à classe minoritária - oversampling, quanto remover - undersampling)
 - ↳ utilizar diferentes custos de classificação para os diferentes clásses
 - ↳ induzir um novo modelo para uma classe (a classe minoritária ou majoritária não são aprendidas separadamente).

PARTE 2 - tratamento de dados ausentes

- ↳ dificuldade relacionada à qualidade dos dados. Pode ser causada por problemas nos equipamentos de coleta, transmissão e armazenamento dos dados ou no preenchimento dos dados (por humanos) → falta de atenção, distração, inexperiência, etc.
- como resolver:
- ↳ eliminar as instâncias c/ dados ausentes
 - ↳ preencher manualmente os valores faltantes → média, etc
 - ↳ método / heurística para definir valores automáticos nos campos faltantes
 - ↳ algoritmos de AM que lidam internamente c/ valores ausentes.

PARTE 3 - dados inconsistentes e redundantes

- ↳ são os dados que possuem valores conflitantes em seus atributos. Diferentes conjuntos de dados podem usar escalas diferentes, mesma medida, etc.
- ↳ podem ser tanto instâncias quanto atributos
- ex: dados idênticos com resultados opostos → ex: idade e data de nascimento

► como resolver:

- ↳ filtros que ajudam na eliminação de redundantes ou inconsistentes.

PARTES 4- conversão simbólica-métrica

↳ técnicas como Redes Neurais artificiais, SVM e outros algoritmos de agrupamento lidam apenas com dados métricos, e nesse caso precisam transformar os dados nominais/categóricos em métricos.

► como resolver:

↳ caso 1: o atributo assume apenas dois valores de presença/ausência → usamos um dígito binário.

↳ caso 2: tipo simbólico com de dois valores → ordinal

minimal ↗ ordinal

o a diferença entre todos os é a ordem dos valores que entre valores métricos deve ser a mesma.

(cada valor é uma sequência de bits → codificação 1-de-c)

bits → codificação 1-de-c → ordinal é codifica cada valor

o para a distância → distância de Hamming → de acordo com a posição mais

Hamming

o não é binário! Não é um número!

o código é armazenado em termômetro

outra alternativa: pseudoeatributos com campo formado só por zeros e uns

PARTES 5- conversão métrico-simbólica

↳ algumas técnicas de AM usam valores qualitativos.

► como resolver?

↳ discretizar o atributo em tipos de valores qualitativos

↳ o conjunto de possíveis valores é dividido em intervais, e cada intervalo de valores quantitativos é convertido em um valor qualitativo.

↳ podem ser superpostos ou não-superpostos

info sobre classe das exmaplos

PARTE 6 - Transformação de atributos numéricos

↳ seu valor mínimo pode precisar ser convertido em outro valor numérico, quando os limites inferior e superior de valores dos atributos são muito diferentes ou estão em escalas diferentes.

► como resolver:

↳ Transformação de normalização de dados:

↳ por amplitude (por rescalada ou padronização)

define uma nova escala

$$V_{\text{novo}} = \frac{\text{Valor} - \text{menor}}{\text{maior} - \text{menor}}$$

(max-min) + ou - uma medida
degraus x ou + outra.

$$V_{\text{novo}} = \frac{\text{Valor} - \text{M}}{\text{L}}$$

PARTE 7 - redução de dimensionalidade

↳ muitos problemas possuem um n.º elevado de atributos, e poucos tâmbore de AM podem lidar c/ um número tão grande, chanceando de melhoria de dimensionalidade

► como resolver:

↳ combinar ou eliminar parte dos atributos irrelevantes com:

↳ agregação (novos atributos formados pela combinação de grupos de atributos)

↳ seleção de atributos (descartam os demais)

↳ 3 abordagens não usadas: pr avaliar a qualidade / desempenho de um subconjunto

↳ embutida (seleção embutida no algoritmo de aprendizado. Ex: árvore de decisão).

↳ baseada em filtro (filtra os subconjuntos. Ex: correlação)

↳ baseada em wrappers (usa o próprio algoritmo como uma caixa-preta pra seleção).

↳ orientados a exemplos