Test seminar 1 - T1 - Rezolvat

50 minute, punctaj minim 1.5 din maxim 6

Subiectul 1

09.11.2023

Arbori de decizie ID3 - 3pct

Considerați setul de date de mai jos, care reprezintă diferite caracteristici ale calculatoarelor și atributul țintă 'Quality', care le clasifică fie ca 'Good', fie ca 'Bad'. Calculați entropia pentru întregul set de date și câștigul de informație pentru împărțirea pe fiecare atribut. Determinați nodul rădăcină al Arborelui de Decizie ID3 pe baza acestor calcule.

Speed	RAM (GB)	Quality
Fast	16	Good
Fast	8	Good
Medium	4	Bad
Slow	16	Bad
Slow	2	Bad
Fast	16	Good
Medium	8	Good
Medium	16	Good
Fast	4	Bad
Slow	8	Bad

Parcurgeti toti pasii calculului, folositi 4 zecimale.
Observatie: Rezolvati tratand atributul "RAM (GB)" ca fiind continuu.

Solutie

Pas 1: Calculati entropia setului de date (0.5pct)

$H(\text{Quality}) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1.$

Step 2: Calcul castig informatie pentru fiecare atribut (2pct)

Pentru 'Speed' (1pct):

$$\text{H}_{\text{Fast}} = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811$$

$$\text{H}_{\text{Medium}} = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9182$$

$$\text{H}_{\text{Slow}} = 0$$

$$\text{IG}(\text{Quality}, \text{Speed}) = H(\text{Quality}) - \left(\frac{4}{10} \times \text{H}_{\text{Fast}} + \frac{3}{10} \times \text{H}_{\text{Medium}} + \frac{3}{10} \times \text{H}_{\text{Slow}} \right) = 0.4001$$

Pentru 'RAM (GB)' (1pct): Atributul fiind continuu trebuie gasite puncte de split. Observam ca pentru RAM=4, avem doar clasificari "Bad" si pentru RAM=8 incepem sa avem atat clasificari "Good" cat si "Bad" si continua sa fie impur si pentru RAM=16. Am putea alege sa folosim pentru split valoarea RAM=6, fie am putea folosi chiar RAM=8, lucuri unde este posibil sa aiba loc tranzitia de la "Bad" la "Good". Vom alege RAM=8.

$$H[RAM \leq 8] = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.9183$$

$$H[RAM > 8] = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

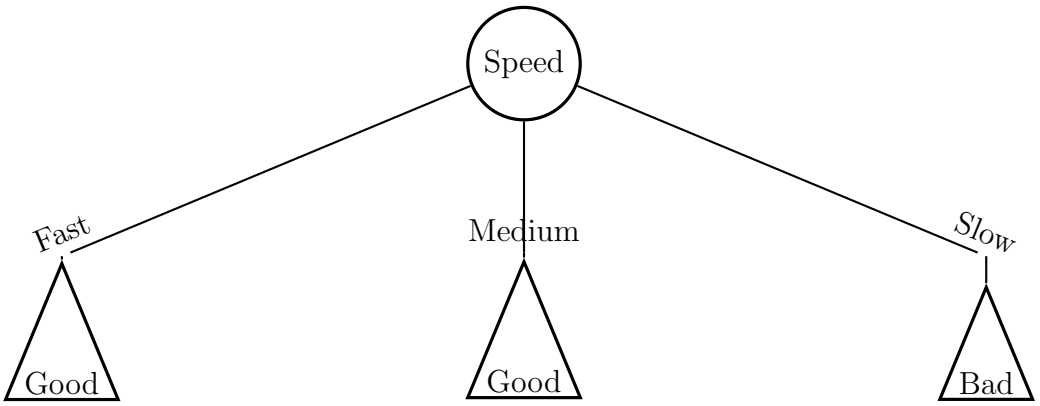
$$IG(\text{Quality}, \text{RAM}) = H(\text{Quality}) - \left(\frac{6}{10} \times H[RAM \leq 8] + \frac{4}{10} \times H[RAM > 8]\right) = 0.1245$$

Step 3: Determinati nodul radacina (0.5pct)
Nodul radacina este dat de atributul cu castig de informatie maxim: Speed

Leave-One-Out Cross Validation (LOOCV) - 1pct

Folosind "decision stump"-ul de mai jos:

1. Calculati cate erori va produce pe setul de date de mai sus. (0.25pct)
2. Aplicati LOOCV cu acest clasificator, cate erori produce? (0.5pct)
3. Ce concluzii puteti trage din cele doua erori? (0.25pct)



Solutie

1. Numaram numarul de instante unde clasificatorul de mai sus greseste. Ne putem uita la impartirea clasificarilori in nodurile de decizie: 1 + 1 + 0 = 2.
   Fast[3 Good, 1 Bad] = Good
   Medium[2 Good, 1 Bad] = Good
   Slow[0 Good, 3 Bad] = Bad

2. Folosim acelasi model (acelasi arbore, unde avem doar nodul radacina "Speed"), aplicam LOOCV si recalculam impartirea in nodurile de decizie, pentru a vedea cum se schimba decizia nodurilor. Ex: eliminand din setul de antrenament prima inregistrare, nodurile de decizie vor deveni:
   Fast[2 Good, 1 Bad] = Good
   Medium[2 Good, 1 Bad] = Good
   Slow[0 Good, 3 Bad] = Bad

   Instanta folosita la validare are clasa Good, deci nu produce eroare.

   In caz de egalitate, vom considera clasificarea sa fie "Good" (sau puteati numara eroarea ca fiind 0.5 in loc de 1) In total avem: 1 + 1 = 2.

3. Putem observa din cele 2 erori, ca algoritmul nu pare a suferii de overfitting (generalieaza ok), eroarea ramanand neschimbata pentru date "noi".

Clasificator Bayes Naiv - 2pct

Considerati urmatorul set de date folosit pentru a prezice daca o persoana va juca tenis sau nu:

Temperature	Humidity	Wind	PlayTennis
Hot	High	Weak	No
Hot	High	Strong	No
Hot	High	Weak	Yes
Mild	High	Weak	Yes
Cool	Normal	Weak	Yes
Cool	Normal	Strong	No
Cool	Normal	Strong	Yes
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Weak	Yes
Mild	Normal	Strong	Yes
Mild	High	Strong	Yes
Hot	Normal	Weak	Yes
Mild	High	Strong	No

1. Pentru o instanta noua cu atributele: Temperature = Cool, Hum = High, Wind = Strong, care va fi decizia luata de clasificatorul Bayes Naiv? Parcurgeti toti pasii calculului, folositi 4 zecimale. (1.5pct)

2. Pentru instanta noua de mai sus (pct. 1), cati parametrii a trebuit sa calculezi? (0.25pct)

3. Cati parametrii trebuiesc calculati pentru a putea clasifica orice instanta viitoare folosind Bayes Naiv? Justificati. (0.25pct)

Solutie

1. Decizia Clasificatorului Bayes Naiv

Calculăm probabilitățile condiționate și apriori pentru fiecare clasă (Yes, No) și atributul corespunzător.

Probabilități Condiționate și Apriori (1pct)

Pentru Yes:

- $P(\text{Temp} = \text{Cool}|\text{Yes}) = \frac{3}{9}$
- $P(\text{Hum} = \text{h}|\text{Yes}) = \frac{3}{9}$
- $P(\text{Wind} = \text{s}|\text{Yes}) = \frac{3}{9}$
- $P(\text{Yes}) = \frac{9}{14}$

Pentru No:

- $P(\text{Temp} = \text{c}|\text{No}) = \frac{1}{5}$
- $P(\text{Hum} = \text{h}|\text{No}) = \frac{4}{5}$
- $P(\text{Wind} = \text{s}|\text{No}) = \frac{3}{5}$
- $P(\text{No}) = \frac{5}{14}$

Calculul Probabilității Finale

$$\begin{aligned} P(\text{Yes}|\text{Temp} = \text{c}, \text{Hum} = \text{h}, \text{Wind} = \text{s}) &= P(\text{Yes}) \times P(\text{Temp} = \text{c}|\text{Yes}) \\ &\quad \times P(\text{Hum} = \text{h}|\text{Yes}) \times P(\text{Wind} = \text{s}|\text{Yes}) \\ &= \frac{9}{14} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \\ &= 0.0237 \end{aligned}$$

$$\begin{aligned} P(\text{No}|\text{Temp} = \text{c}, \text{Hum} = \text{h}, \text{Wind} = \text{s}) &= P(\text{No}) \times P(\text{Temp} = \text{c}|\text{No}) \\ &\quad \times P(\text{Hum} = \text{h}|\text{No}) \times P(\text{Wind} = \text{s}|\text{No}) \\ &= \frac{5}{14} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \\ &= 0.0342 \end{aligned}$$

Luam decizia "No". (0.5pct)

2. Numărul de Parametri Calculați pentru Instanța Nouă

Pentru instanța nouă, am calculat 7 parametri.

3. Numărul Total de Parametri pentru Clasificarea Viitoare

Pentru fiecare atribut, trebuie sa calculam probabilitatile conditionale Yes/No pentru unique(atribute)-1 valori + probabilitatea pentru PlayTennis=Yes. In total avem $(3-1+2-1+2-1)*2+1 = 9$ parametrii