Test seminar 1 - T1 - Rezolvat

50 minute, punctaj minim 1.5 din maxim 6

Subiectul 2

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	Bad
Medium	4	Bad
Slow	16	Good
Slow	2	Bad
Fast	16	Good
Medium	8	Bad
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.

Soluție

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

Calculăm entropia pentru întregul set de date folosind formula entropiei:

$$H(\text{Quality}) = -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.9709.$$

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

Pentru 'Speed' (1pct): Calculăm entropia pentru fiecare valoare a atributului 'Speed' și apoi calculăm câștigul de informație:

$$IG(\text{Quality}, \text{Speed}) = 0.0199.$$

Pentru 'RAM (GB)' (1pct): Pentru atributul continuu 'RAM (GB)', alegem un punct de split la RAM = 8. Se poate calcula standard, sau se observa ca datele sunt impartite perfect. Deci information gain-ul este total.

$$IG(\text{Quality}, \text{RAM}) = 0.9709.$$

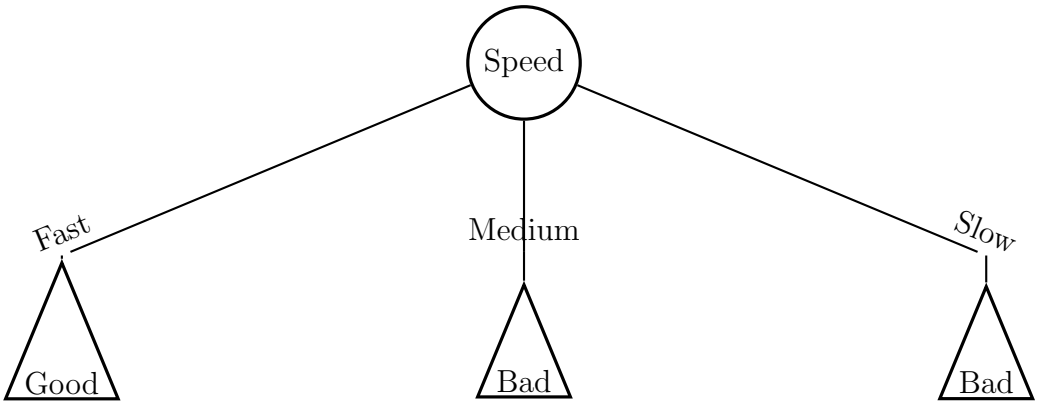
Pas 3: Determinati nodul radacina (0.5pct)

Nodul rădăcină este atributul cu câștigul de informație maxim, care, în acest caz, este 'RAM'.

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: 2 + 1 + 1 = 4.
   Fast[2 Good, 2 Bad] = Good
   Medium[1 Good, 2 Bad] = Bad
   Slow[1 Good, 2 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[1 Good, 2 Bad] = Bad
   Medium[1 Good, 2 Bad] = Bad
   Slow[1 Good, 2 Bad] = Bad

   Instanta folosita la validare are clasa Good, deci este o 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 + 1 + 1 + 1 + 1 + 1 + 1 + 1 = 9

3. Putem observa din cele 2 erori, ca algoritmul nu generalizeaza bine cand vede date "noi". Lucru care sugereaza posibilitatea prezentei fenomenului overfit.

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 = Mild, Humidity = High, Wind = Weak, 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)

Soluție

1. Decizia Clasificatorului Bayes Naiv

Pentru instanța nouă cu atributele Temperature = Mild, Humidity = High și Wind = Weak, calculăm probabilitățile condiționate și apriori pentru fiecare clasă (Yes, No).

Probabilități Condiționate și Apriori

Calculăm probabilitățile pentru fiecare atribut și clasă:

- $P(\text{Yes}) = \frac{9}{14}$
- $P(\text{Temp} = \text{Mild}|\text{Yes}) = \frac{4}{9}$
- $P(\text{Hum} = \text{High}|\text{Yes}) = \frac{3}{9}$
- $P(\text{Wind} = \text{Weak}|\text{Yes}) = \frac{6}{9}$
- $P(\text{No}) = \frac{5}{14}$
- $P(\text{Temp} = \text{Mild}|\text{No}) = \frac{2}{5}$

- $P(\text{Hum} = \text{High}|\text{No}) = \frac{4}{5}$
- $P(\text{Wind} = \text{Weak}|\text{No}) = \frac{2}{5}$

Calculul Probabilității Finale

Calculăm probabilitățile finale pentru a decide dacă se va juca tenis sau nu:

$$\begin{aligned} P(\text{Yes}|\text{Temp} = \text{Mild}, \text{Hum} = \text{High}, \text{Wind} = \text{Weak}) &= P(\text{Yes}) \times P(\text{Temp} = \text{Mild}|\text{Yes}) \\ &\quad \times P(\text{Hum} = \text{High}|\text{Yes}) \times P(\text{Wind} = \text{Weak}|\text{Yes}) \\ &= \frac{9}{14} \times \frac{4}{9} \times \frac{3}{9} \times \frac{6}{9} \\ &= 0.0635 \end{aligned}$$

$$\begin{aligned} P(\text{No}|\text{Temp} = \text{Mild}, \text{Hum} = \text{High}, \text{Wind} = \text{Weak}) &= P(\text{No}) \times P(\text{Temp} = \text{Mild}|\text{No}) \\ &\quad \times P(\text{Hum} = \text{High}|\text{No}) \times P(\text{Wind} = \text{Weak}|\text{No}) \\ &= \frac{5}{14} \times \frac{2}{5} \times \frac{4}{5} \times \frac{2}{5} \\ &= 0.0457 \end{aligned}$$

Luam decizia "Yes". (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