

45.

(O aproximare a numărului de instanțe greșit clasificate care au fost asignate la un nod frunză dintr-un arbore ID3)

• CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 9.b

Învățăm un arbore de decizie folosind un set de date de antrenament cu atributul de ieșire (*class*) având valorile 0 sau 1.

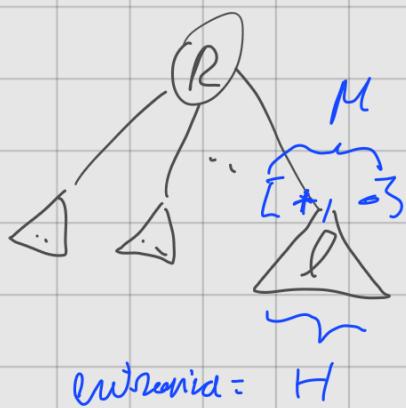
Presupunând că pentru un nod frunză l din acest arbore,

- există M instanțe de antrenament asignate la acel nod, iar
- entropia sa este H ,

scrieți un algoritm simplu care ia ca valori de intrare M și H și furnizează la ieșire numărul de exemple de antrenament clasificate greșit de către nodul frunză l .

Sugestie: Folosiți o aproximare simplă (polinomială) pentru funcția entropie $H(p)$.

584



$$H(a, b) = \frac{a}{a+b} \log_2 \frac{a}{a+b} + \frac{b}{a+b} \log_2 \frac{b}{a+b}$$

Pie p - abz. coresp. clasi majoritari:

$$H(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

dară d p.n majoritar $\Rightarrow \frac{b}{a+b} = 1 - \frac{a}{a+b}$

INPUT:

M, H

OUTPUT

K (nr. of wrong classifications)

fem (M, H):

1: if $H < 0$:

return "invalid H"

2: if $H = 1$:

$$P = \frac{1}{2}$$

elif $H = 0$:

$$P = 1$$

else:

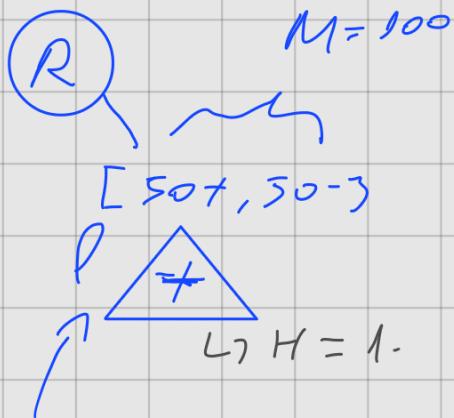
$$P = 2^{-H}$$

3:

rezervă $(1-P) \cdot M$ numărul de crisi.

Etape rezervare;

Ex:



$$R = 100 \cdot \frac{1}{2} = 50\%$$

\rightarrow jum. rez. rez. și gresit clasific.

rez. maj. neadele lă solvare.

Căutăm să rezolvăm P în funcție de $H(P)$:

$$P \log_2 \frac{1}{P} + (1-P) \log_2 \frac{1}{(1-P)}$$

$$-P \log_2 P - (1-P) \log_2 (1-P) = M$$

$$-P \log_2 P - \log_2 (1-P) + P \cdot \log_2 (1-P) = M$$

$$-\log_2 \frac{P^P}{(1-P)^{1-P}} + \log_2 \frac{(1-P)^P}{P^P} = M$$

$$\log_2 \frac{(1-P)^P}{P^P} - \log_2 \frac{(1-P)}{P} = M$$

$$\log_2 \frac{(1-P)^{P-1}}{P^P} = M$$

$$\frac{(1-P)^{P-1}}{P^P} = 2^M$$

$$P \approx \frac{1}{2} - M$$

Obr. Dacă $H=1$, eroarea nu poate fi de 50%.

Dacă $H=0$, eroarea nu poate fi 0%.

Stim că $H \geq 0$.

Probleme propuse

ARBORI de DECIZIE

46.

(Algoritmul ID3: eroarea la antrenare)

• CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 2.1

Un student mi-a spus următoarele:

- el poate să construiască un set de instanțe cu atributele de intrare discrete și atributul de ieșire binar;
- mie îmi dă voie să aleg o parte din acest set de instanțe (dar nu toate!) pentru a antrena un arbore de decizie;
- indiferent de modul cum mi-ă aleger datele de antrenament din setul construit de el, eroarea de clasificare pe care arborele de decizie (obținut în urma antrenării cu algoritm ID3) o va face pe instanțele care nu au fost incluse în setul de antrenament va fi de cel puțin 50%.

Credeti că studentul are dreptate? Explicați de ce sau dați un exemplu.

Nu gîndim

le XOR

XOR este cel

corect

Ești că nu are dreptate, deoarece dacă aleg

un set de instanțe consistent, în care nă obțin eroarea

de antrenare 0%, conform prop. ID3, \Rightarrow voi putea

obține pe setul de validare un procent de eroare

mai mic de 50%. Deși suntem să avem un nr de

instanțe nu foarte mare și să nu întrevină fenomenul

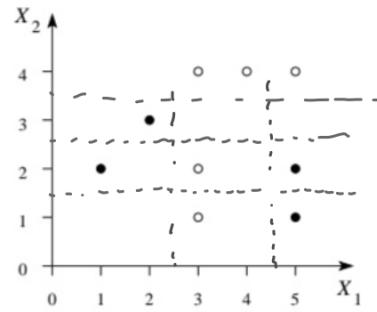
de overfitting.

(ID3 cu atribut continue:
zone de decizie și separatori decizionali)

■ □ • Liviu Ciortuz, 2017,
folosind datele de la problema 24

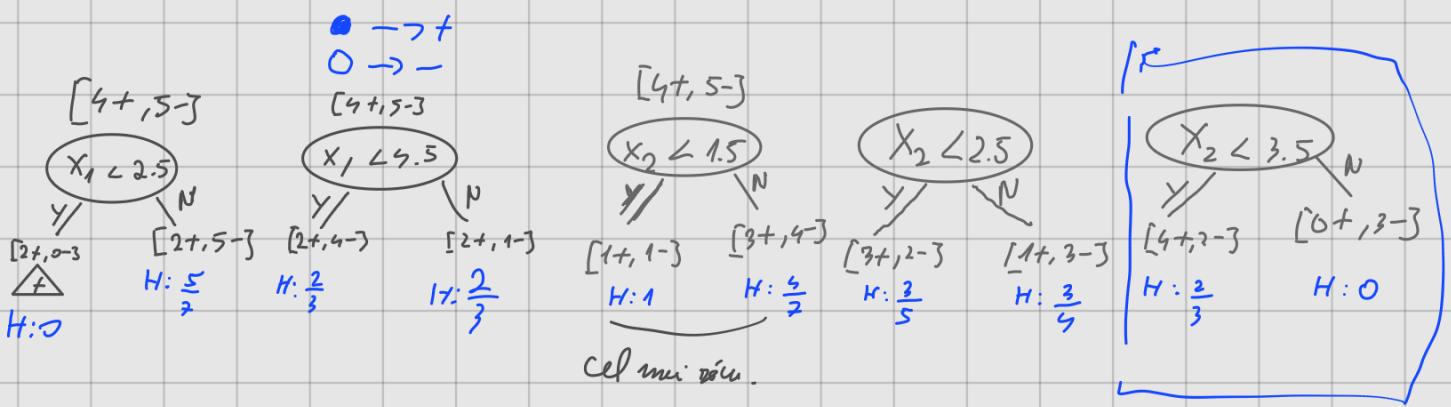
Fie setul de date de antrenament din figura de mai jos (partea dreaptă). X_1 și X_2 sunt considerate atributuri numerice continue. Vă readucem aminte convenția noastră de notare: simbolul • desemnează instanțe pozitive, iar simbolul ○ instanțe negative.

Aplicați algoritmul ID3 pe acest set de date.
(Faceți toate calculele necesare, în mod detaliat; precizați la fiecare pas care sunt pragurile de splitare pentru cele două atributuri.) Desenați arborele de decizie rezultat. La final, reprezentați grafic zonele de decizie și separatorii decizionali, marcând clar zona pozitivă (sau zonele pozitive) și zona negativă (sau zonele negative).



585

Nivel 1



cel mai bine

 $X_1 < 2.5$

$$H_{\text{mild}}|X_1 = \frac{4}{9} \cdot H[2+, 0-] + \frac{5}{9} \cdot H[2+, 5-] = \frac{4}{9} \cdot \left(\frac{2}{7} \cdot \log \frac{2}{7} + \frac{5}{7} \cdot \log \frac{5}{7} \right) = \frac{4}{9} \cdot 0.4010 + \frac{5}{9} \cdot 0.4010 = 0.4020 = 0.4795$$

 $X_1 < 2.5$

$$H_{\text{mild}}|X_1 = \frac{4}{9} \cdot H[2+, 4-] + \frac{5}{9} \cdot H[2+, 1-] = H[2+, 1-] \left(\frac{4}{9} + \frac{5}{9} \right) = H[2+, 1-] =$$

$$= \frac{1}{3} \log \frac{2}{3} + \frac{2}{3} \log \frac{3}{2} = 0.5283 + 0.5283 = 1.0566$$

 $X_2 < 1.5$

$$H_{\text{mild}}|X_2 = \frac{4}{9} \cdot 1 + \frac{5}{9} \cdot H[3+, 4-] = \frac{4}{9} + \frac{5}{9} \left(\frac{3}{7} \log \frac{3}{7} + \frac{4}{7} \log \frac{4}{7} \right) =$$

$$= 0.8900$$

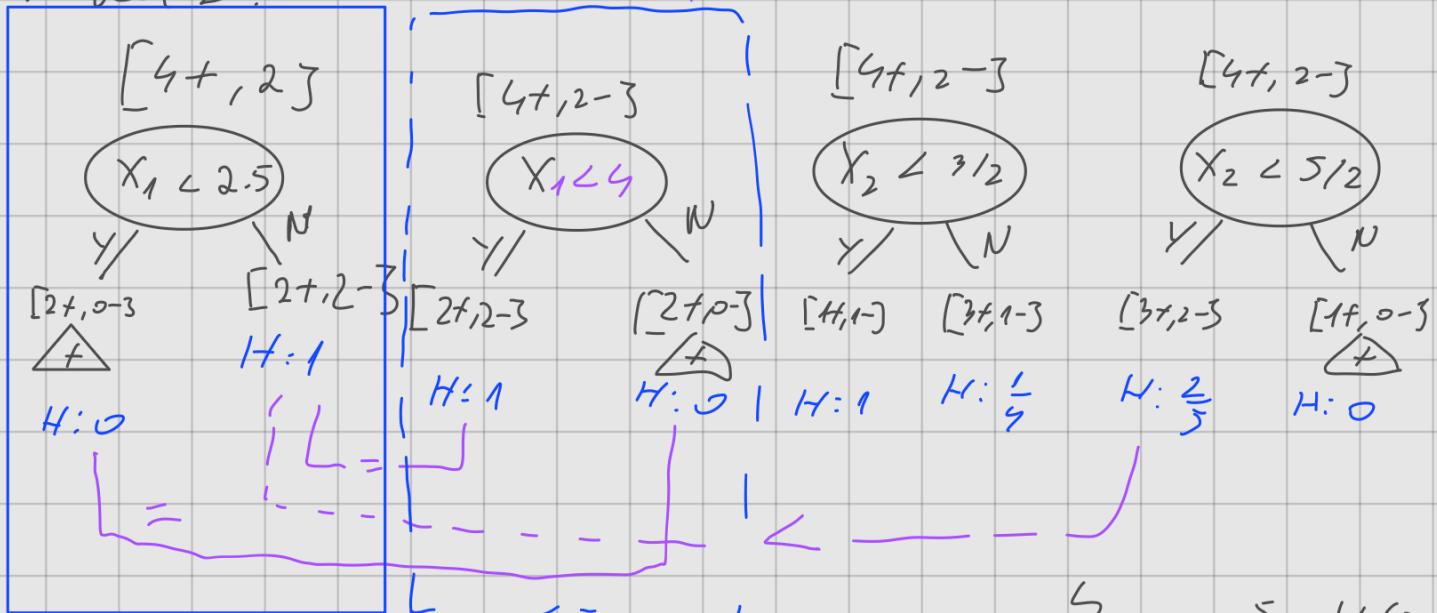
$$H_{\text{node } 1 | X_2} = \frac{5}{9} \cdot H[3+, 2-3] + \frac{5}{9} \cdot H[1+, 3-3] =$$

$\bullet H[3+, 2-3] = \frac{2}{5} \log \frac{5}{2} + \frac{3}{5} \log \frac{5}{3} = 0.4643 + 0.4643 = 0.9286$

$$X_2 < 3.5$$

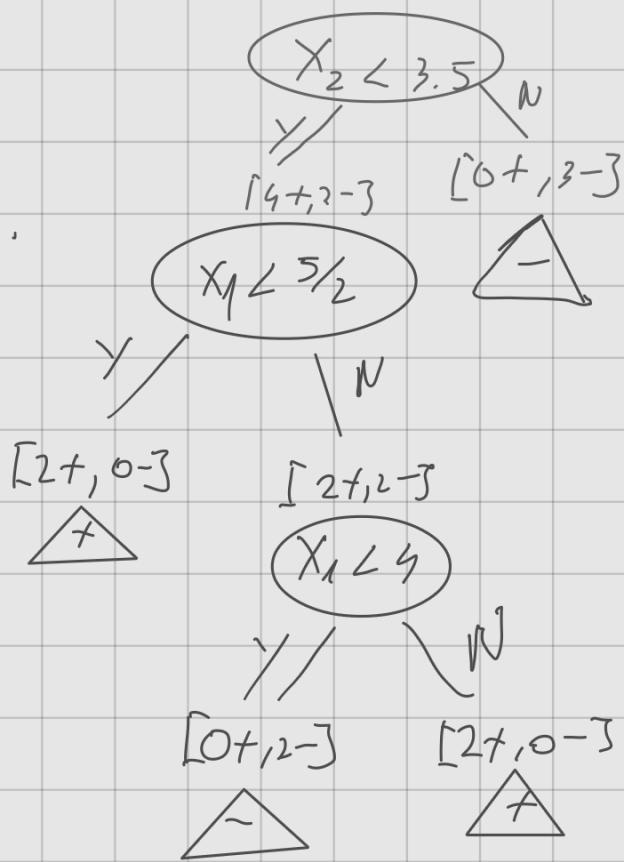
$$H_{\text{node } 1 | X_2} = \frac{5}{9} \cdot H[4+, 2-3] = \frac{5}{9} \cdot 0.8616 = 0.4820$$

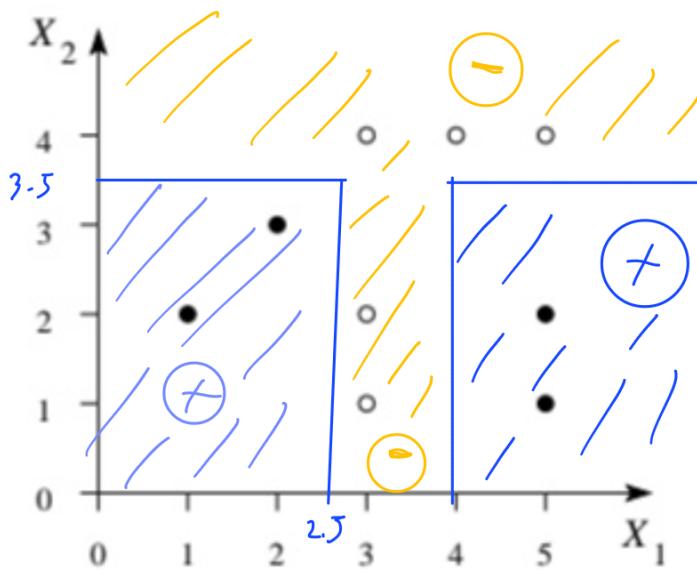
Nivel 2:



The final decisions:

$$\frac{5}{6} < \frac{5}{6} \cdot 4 \left(\frac{2}{5} \right) \\ 0.6666666666666666 < 0.9286 \\ 0.7776$$



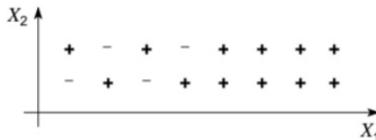


49.

(Extinderea algoritmului ID3 cu atrbute continue; eroarea la antrenare, eroarea la CVLOO; overfitting)

* CMU, 2005 fall, T. Mitchell, A. Moore, midterm exam, pr. 2

Figura alăturată prezintă un set de date cu două intrări X_1 și X_2 , variabile cu valori reale, și o ieșire Y care poate lua valori pozitive (+) sau negative (-).

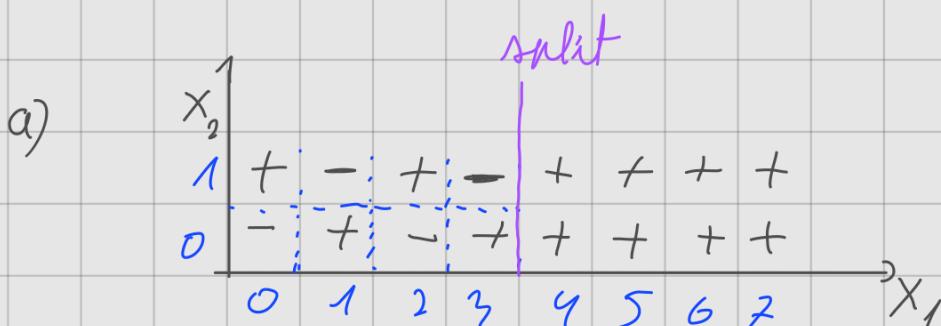


Testăm doi algoritmi extremi de învățare de arbori de decizie. Algoritmul **OVERFIT** construiește un arbore de decizie în maniera standard a algoritmului ID3, fără a face pruning. Algoritmul **UNDERFIT** refuză complet să-și asume riscul splitării intervalelor de valori pentru X_1 și X_2 și construiește un arbore de decizie alcătuit doar dintr-un singur nod (care va fi simultan și rădăcină și nod frunză, deci și nod de decizie).

- Câte noduri frunză vor fi în arborele de decizie învățat de **OVERFIT** pe aceste date?
- Care va fi eroarea de clasificare la cross-validation cu metoda "Leave-One-Out" pe acest set de date atunci când folosim algoritmul **OVERFIT**? (Indicați datele incorect clasificate.)
- Similar punctului anterior, pentru cazul în care se folosește algoritmul **UNDERFIT**.

ID3 → face overfit
↓
combatență cu pruning

De Refacere.



Avem circa 9 frunze, decareea avem circa cîte o frunză pentru fiecare zece majoritar, în partea stângă se ob. singurul mod posibil pentru delimitarea setului, astfel vom avea cîte o frunză pentru fiecare interval. iar după split vom avea cîte o frunză pentru fiecare interval.

majoritar, deci înăoș o frunză. $\Rightarrow 8+1=9$ frunze

b)

vom considera intervalul

X_2

X_1

X_0

X_0

X_1

X_2

X_3

X_4

X_5

X_6

X_7

X_8

X_9

X_10

X_11

X_12

X_13

X_14

X_15

X_16

X_17

X_18

X_19

X_20

X_21

X_22

X_23

X_24

X_25

X_26

X_27

X_28

X_29

X_30

X_31

X_32

X_33

X_34

X_35

X_36

X_37

X_38

X_39

X_40

X_41

X_42

X_43

X_44

X_45

X_46

X_47

X_48

X_49

X_50

X_51

X_52

X_53

X_54

X_55

X_56

X_57

X_58

X_59

X_60

X_61

X_62

X_63

X_64

X_65

X_66

X_67

X_68

X_69

X_70

X_71

X_72

X_73

X_74

X_75

X_76

X_77

X_78

X_79

X_80

X_81

X_82

X_83

X_84

X_85

X_86

X_87

X_88

X_89

X_90

X_91

X_92

X_93

X_94

X_95

X_96

X_97

X_98

X_99

X_100

Conform algoritmului, vom reține
nu treind către un segment din datele noile
de antrenament, vom antrena alături de acel segment
apoi vom verifica cu segmentul de date noile inițial.

- Dacă scoatem nu treind X_4, X_5, X_6, X_7 de la linia X_1 ,

pe nivelurile $X_2 = X_0$ și $X_2 = X_1$, vom obține că urmărește să validăm

trei segmente din stânga splitului nostru ca cum valoarea
alternativă, creând practic „segment” în datele noile. Astfel dacă vom elimiune
nu treind către una din acetele date, vom obține o creare. $\Rightarrow 8$ eroare ✓

Dacă, urmărește să clasificăm pt. OVERFIT va fi de $\frac{8}{16} = \frac{1}{2}$

c) Pentru algoritm Underfit vom avea singur mod să validăm?

frunză.

→ în felul decisiei majoritare =

\Rightarrow atât datele „decidute” nu vor fi mereu clasificate corect,

decare Underfit indifferent de datele de intrare și ieșire

$\frac{1}{16} + \frac{9}{16}$ (semnul magazionator) \Rightarrow avem 3 semne, - "din 16" \Rightarrow litera mărfă

$$\frac{Y}{16} = \frac{1}{4}$$

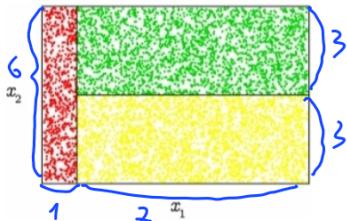
50.

(Arbore de decizie cu variabile continue:
ID3 ca algoritm "greedy")

• CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 8.abcf

Fie următoarea problemă de clasificare ternară.

Considerăm că în figura alăturată regiunea dreptunghiulară este populată în mod dens cu puncte caracterizate de două atrbute numerice (continue), x_1 și x_2 . Cele trei subdreptunghiuri (roșu, verde și galben) reprezintă trei clase de puncte, C_1, C_2 , și C_3 .



Dimensiunile $x_1 \times x_2$ ale dreptunghiurilor roșu, verde și galben sunt 1×6 , 7×3 și respectiv 7×3 . Dreptunghiul roșu este populat în mod uniform cu 6000 de puncte din clasa C_1 . Dreptunghiul verde este populat în mod uniform cu 42000 de puncte din clasa C_2 . Dreptunghiul galben este populat în mod uniform cu 42000 de puncte din clasa C_3 . Pentru simplitate, nu vom considera alte puncte decât acestea.

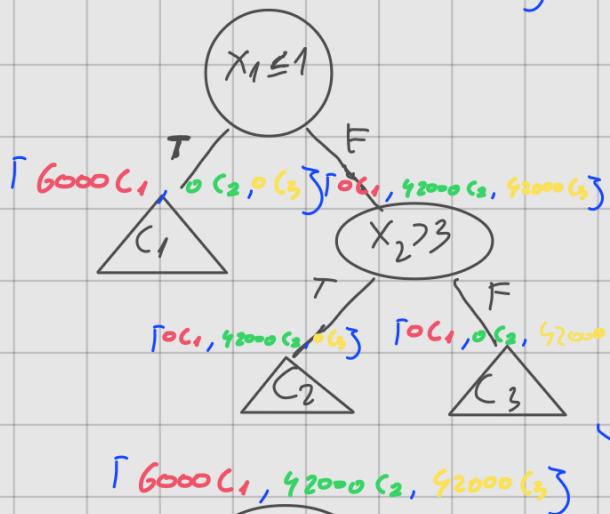
- a. Care este numărul minim de noduri de test pe care trebuie să le aibă un arbore de decizie pentru a clasifica în mod corect acest set de date?

→ arbore optim

d)

$$[6000C_1, 42000C_2, 42000C_3]$$

Caz 1:



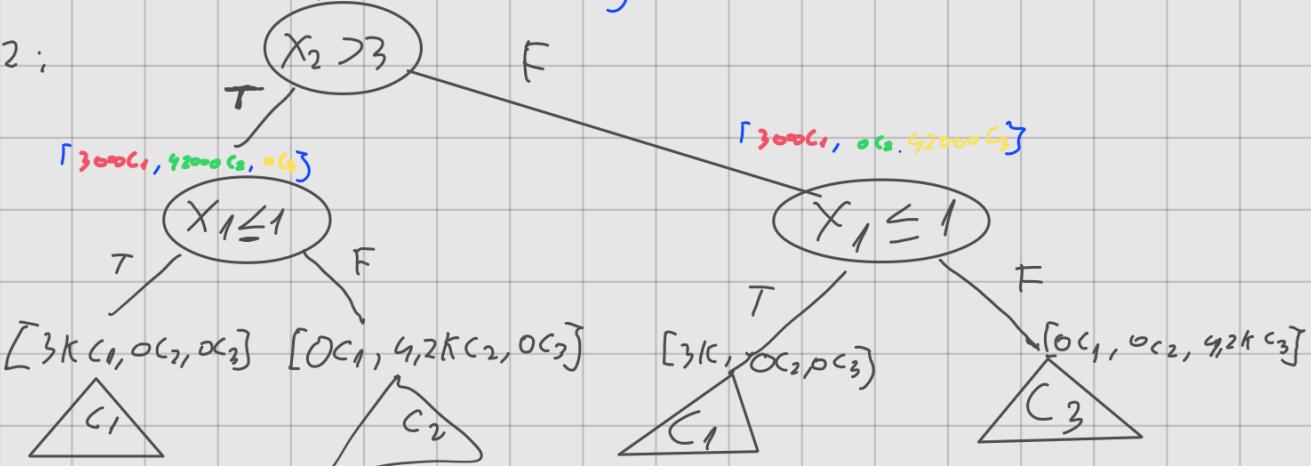
- Având în vedere că avem 3 clase \Rightarrow avem nevoie de cel puțin 3 frunze.

- Având valori continue, vom avea un arbore binar cu doar 2 semne (True și False).

Deci, vom avea un arbore mai mic de 2 nodule.

→ optim

Caz 2:



b. Câte noduri de test are arborele de decizie obținut în urma antrenării algoritmului ID3 pe acest set de date, folosind criteriul maximizării câștigului de informație?

Indicație: Pentru a determina entropiile condiționale minime, puteți folosi proprietatea A3 de la problema 62 de la capitolul de *Fundamente*.

c. Avem același număr de noduri în cele două cazuri de mai sus, sau nu? Care credeți că este explicația?

d. Un arbore de decizie poate să clasifice setul de date din figura de mai sus cu 100% acuratețe (presupunând că nu există zgomote la nivel de etichete). Ce condiții trebuie să satisfacă în general un set de date de acest gen astfel încât arborele de decizie rezultat în urma antrenării să fie cât mai compact și să producă o acuratețe de 100%?

Indicație: Fiecare nod intern al arborelui de decizie corespunde unui test bazat pe o singură trăsătură. Gândiți-vă ce fel de clase de funcții / granițe de separare corespund unui astfel de arbore de decizie.

b) Vom avea ^{minim} 3 noduri de test. Avem un arbore compact datorită clasificării corecte și extrem declare a datelor din setul de date.

Nu erau corecte

✓) Răspunsul 103 este de tip *Greedily*.

c) Nu avem, deoarece cazul 1 prima dată se face delimitarea

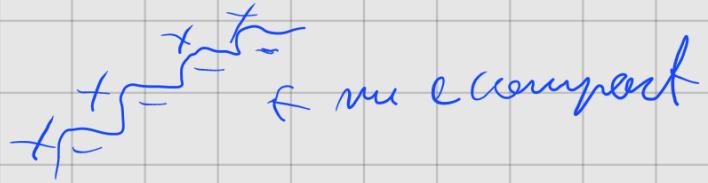
(X) optimă, în care rezultă să delimităm o clasă de restul), într-

în cazul 2 nodul rădăcină nu face o delimitare optimă (practic

împarte datele între (rosu, verde) și (rosu, galben),

astfel picind nevoie de încă 2 noduri de test pentru a separa

corect aceste clase.



- datele ar trebui să aibă o clasificare clară, în ceea ce

nodurile din acestă clasă să fie cât mai bine grupate și că nu

lucră separat față de celelalte clase.

- Frontiera de separare ar trebui să fie cât mai rectangulară, deoarece fiecare nod va avea testa o singură trănsătură (oricare o liniă x_1 sau x_2). Acum aspect face dificilă clasificarea dacă datele sunt amestecate.

- Densitatea datelor ar trebui să fie cât mai uniformă pe fiecare

Regiune de clasă (drecțională)

51.

(Un exemplu de aplicare a algoritmului ID3:
cazul când se folosesc atât variabile discrete
cât și variabile continue)

□ • * CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 1

Cursul de Învățare Automată pe care l-am urmat în acest semestrul, îl-a insuflat dorința ca după absolvirea facultății să fondezi pe cont propriu o firmă (engl., start-up company). Pentru a-ți evalua șansele de succes — adică, mai exact, dacă vei deveni milionar sau nu —, ai colectat date de la absolvenții de la CMU, referitoare la foști studenți care și-au înființat propriile lor companii start-up. Pentru fiecare start-up, știi acum ce anume produce compania respectivă, ce fonduri de capital de investiție a obținut (exprimat în milioane de dolari), dacă directorul companiei a studiat la CMU o disciplină din domeniul științelor exacte și, în final, dacă directorul a devenit milionar. Tabelul de mai jos centralizează datele pe care le-ai cules.

Produs	Capital atras	Științe exacte	Milionar
SiteDeSocializare	2.9	Da	Nu
SiteDeSocializare	1.7	Da	Nu
SiteDeSocializare	3.4	Da	Nu
SiteDeSocializare	2.3	Nu	Da
MașinaAlimentatăCuCombustibilBio	3.4	Da	Da) +
MașinaAlimentatăCuCombustibilBio	6.1	Da	Da
MașinaAlimentatăCuCombustibilBio	5.6	Nu	Nu
MașinaAlimentatăCuCombustibilBio	0.6	Nu	Nu
NanoVaccin	1.9	Nu	Nu
NanoVaccin	2.9	Nu	Da) +
NanoVaccin	3.1	Da	Da
NanoVaccin	0.3	Da	Nu

X Grej! re tot schimbat
off

separatör

Acum ai vrea să construiești un arbore de decizie pornind de la aceste date. Referitor la atributul cu valori continue *CapitalAtras*, știm că arborele de decizie poate conține teste (partiționări binare) de forma $\text{CapitalAtras} \leq v$ și $\text{CapitalAtras} > v$ și că pot exista mai multe teste de acest fel în arbore.

- a. Câte „praguri” distințe vă trebuie să considerăm pentru CapitalAtras atunci când căutăm atributul (optim) care trebuie pus în nodul rădăcină?

- b. Desenați arborele de decizie care va fi învățat de către algoritmul ID3 extins cu atribute cu valori continue, aşa cum a fost prezentat la curs. Ne vom referi ulterior la acest arbore ca fiind *arborele original*. Adnotați fiecare nod intern (i.e., nod de test) din arbore cu câștigul de informație obținut în urma aplicării testului respectiv.

Indicație: Pentru punctele c și d de mai jos, deși nu este neapărat necesar, este recomandabil să folosiți o implementare a algoritmului ID3 (extins cu atributuri numerice continue). Vă sugerăm să abordați mai întâi problemele 35 și 58, care vă ghidează cum să construiți propria dumneavoastră implementare.

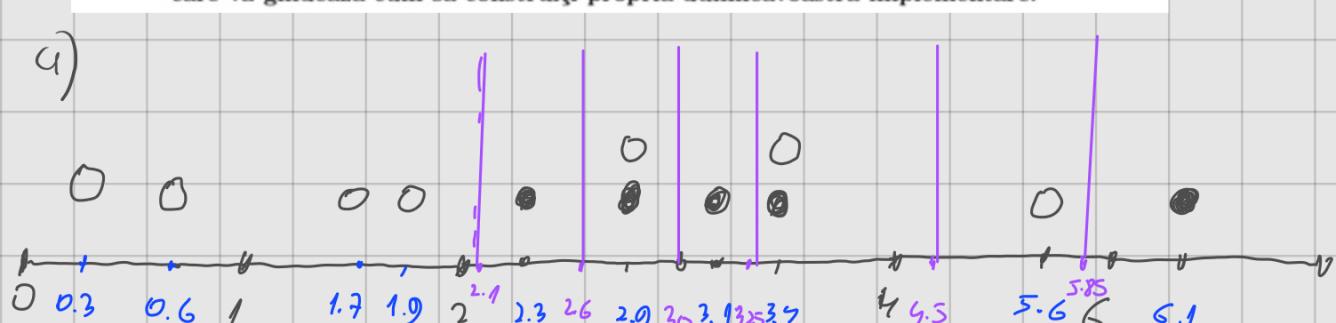
Text 3

Deepa Piccare

Nas, actuatoraz

Osa ventru at.

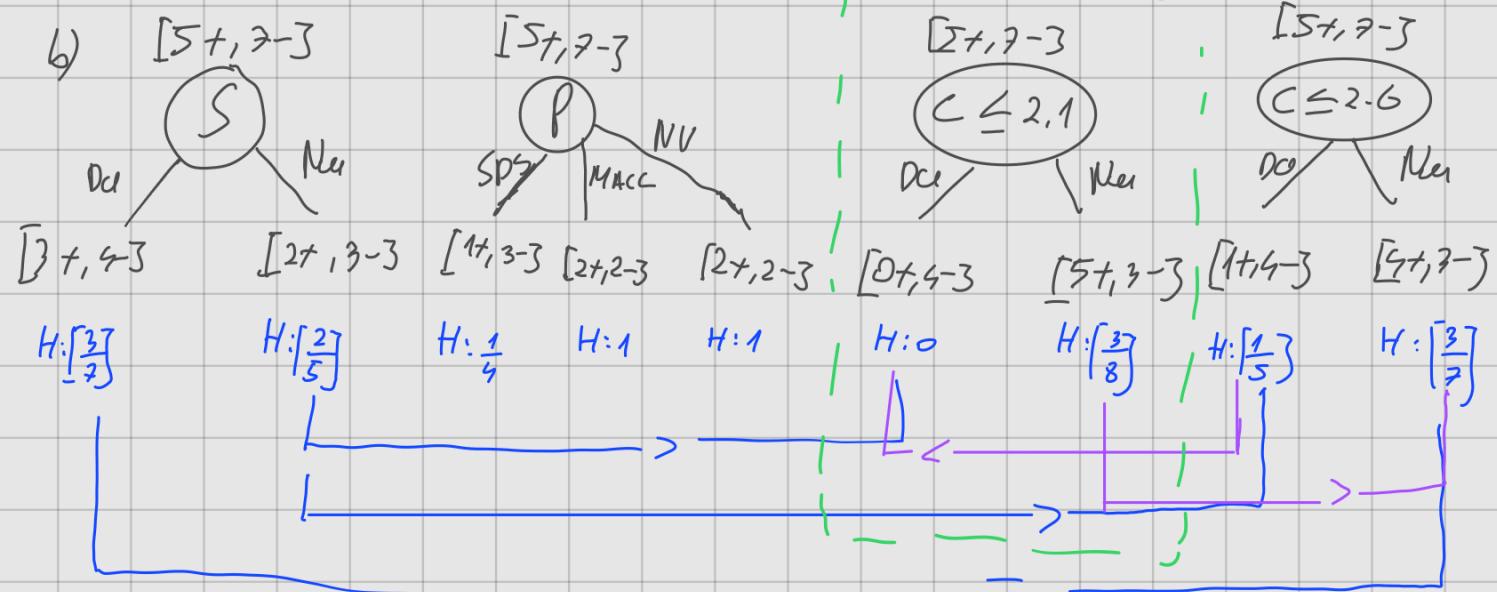
continued



$$\begin{array}{r}
 1.07 + & 2.3 & 2.0 \\
 2.3 & \underline{2.0} & \underline{3.1} \\
 \hline
 2.1 & \underline{\underline{5.212}} & \underline{\underline{6.012}} \\
 & 2.6 & 3.0
 \end{array}
 \quad
 \begin{array}{r}
 41.712 \\
 \hline
 5.85
 \end{array}$$

areum G praguevi: $\{2.1, 2.6, 3.0, 3.25, 4.5, 5.85\}$

Panel 1:



$$[5+, 7-3]$$

$$C \leq 3.0$$

Da / Ne

$$[5+, 7-3]$$

$$C \leq 3.25$$

Ne / Ne

$$[5+, 7-3]$$

$$C \leq 4.5$$

Da / Ne

$$[5+, 7-3]$$

$$C \leq 5.85$$

Da / Ne

$$[2+, 5-3]$$

$$[3+, 2-3]$$

$$[3+, 5-3]$$

$$[2+, 2-3]$$

$$[5+, 6-3]$$

$$[1+, 1-3]$$

$$[4+, 7-3]$$

$$[4+, 0-3]$$

$$H: \left[\frac{2}{7} \right]$$

$$H: \left[\frac{2}{5} \right]$$

$$H: \left[\frac{3}{8} \right]$$

$$H: 0$$

$$H: \left[\frac{2}{5} \right]$$

$$H: 1$$

$$H: \left[\frac{5}{8} \right]$$

$$H: 0$$

$$H: 0$$

$$\text{Calc: } {}^{\text{mod}} H = \frac{5}{12} \cdot H[0+, 4-3] + \frac{8}{12} H[5+, 3-3] = \frac{8}{12} \left(\frac{3}{8} \log \frac{3}{8} + \right.$$

$$+ \left. \frac{5}{8} \log \frac{5}{8} \right) = 0.6362$$

$${}^{\text{mod}} H = \frac{5}{12} \cdot H[1+, 4-3] + \frac{7}{12} H[4+, 3-3] = \frac{5}{12} \left(\frac{1}{5} \log \frac{5}{7} + \frac{2}{5} \log \frac{2}{7} \right) +$$

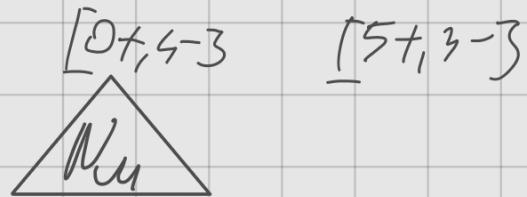
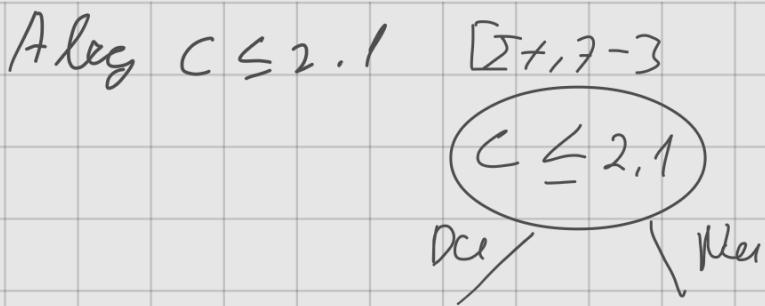
$$+ \frac{7}{12} \left(\frac{3}{7} \log \frac{7}{3} + \frac{4}{7} \log \frac{7}{4} \right) = \frac{5}{12} (0.4643 + 0.2575) + \frac{7}{12} (0.5238 + 0.4613)$$

$$= \frac{5}{12} \cdot 0.7218 + \frac{7}{12} \cdot 0.9851 = 0.87539$$

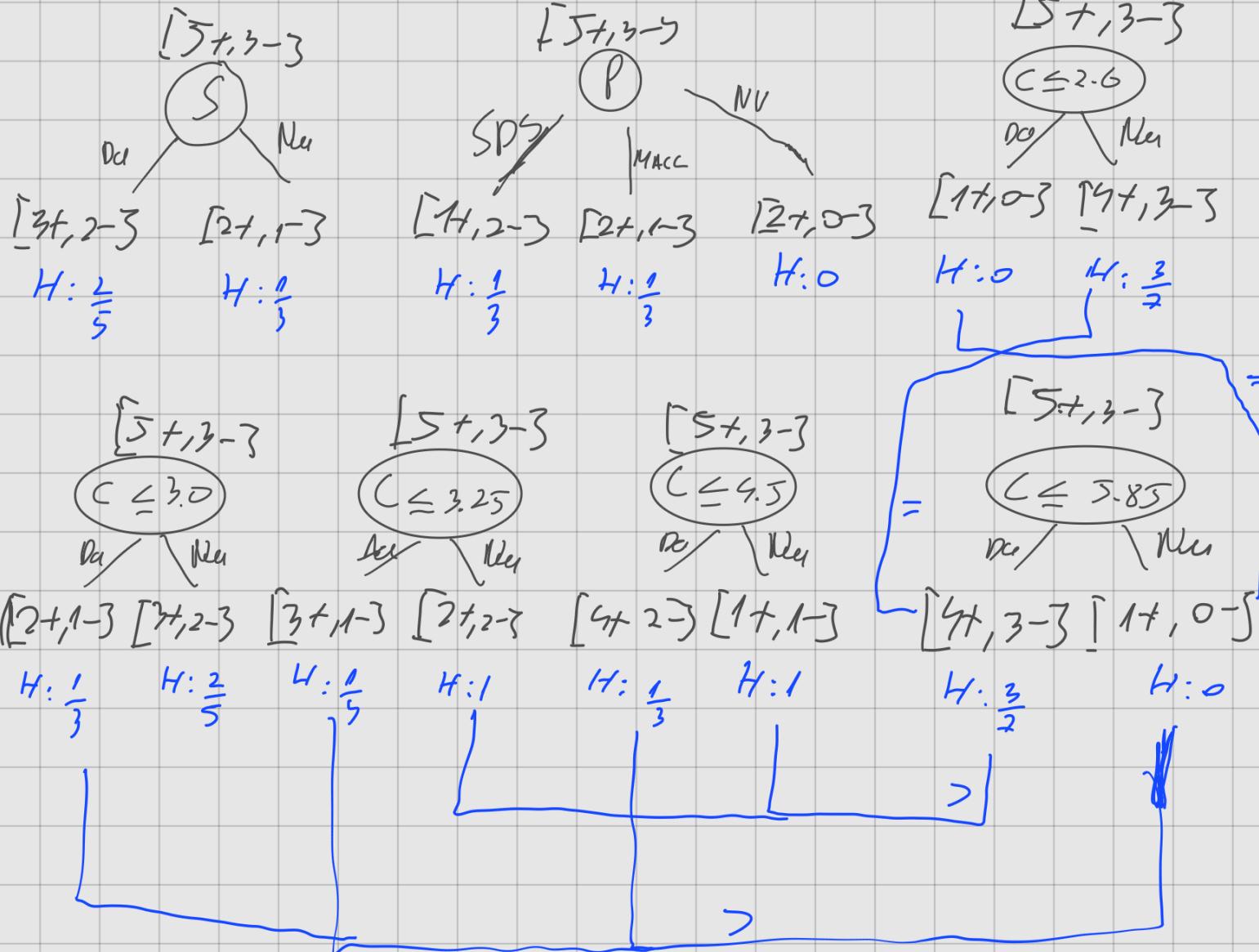
$${}^{\text{mod}} H = \frac{5}{12} \cdot H[4+, 7-3] + \frac{7}{12} H[1+, 0-3] =$$

$$= \frac{11}{12} \cdot \left(\frac{5}{11} \log \frac{11}{5} + \frac{7}{11} \log \frac{11}{7} \right) - 0.8668$$

$$H|_{\text{mod} / C \leq 2, 1} = 0.6262 \quad \text{and} \quad H|_{\text{mod} / C \leq 5.85} = 0.8753$$



Panel 2:



⑩ H million/p = $\frac{3}{8} H [1+, 2-3] + \frac{3}{8} H [2+, 1-3] + 0 = 2 \cdot \frac{3}{8} H [1+, 2-3] =$

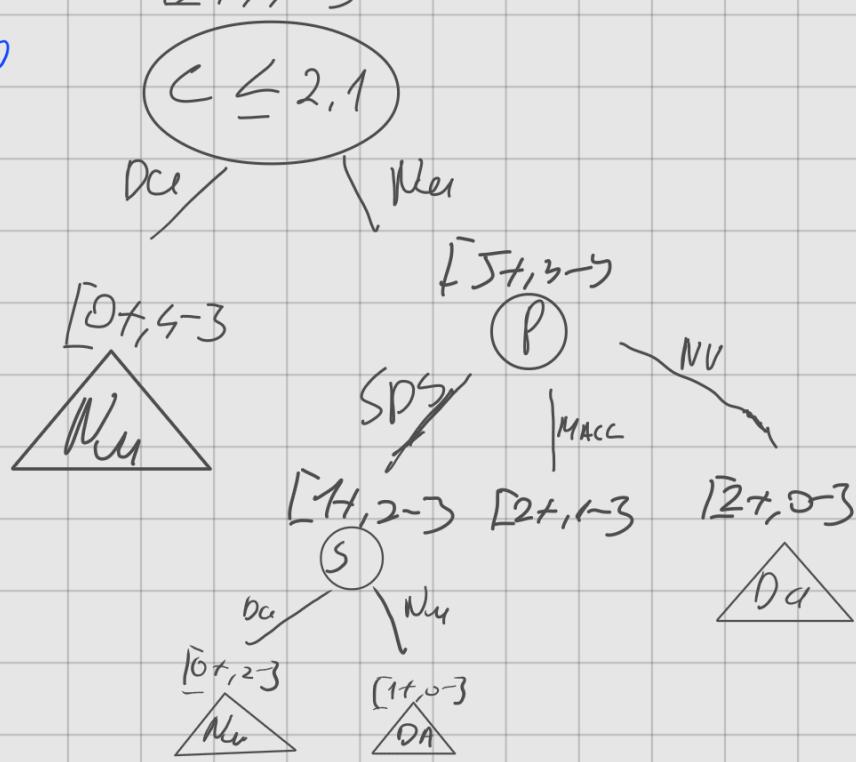
 $= \frac{3}{8} \cdot 0.0182 = 0.1365$

0.9182

$$\textcircled{8} \quad H_{\text{Mitsamw} | C \leq 2.6} = \frac{2}{8} \cdot H[\Sigma+3-] = 0.8619$$

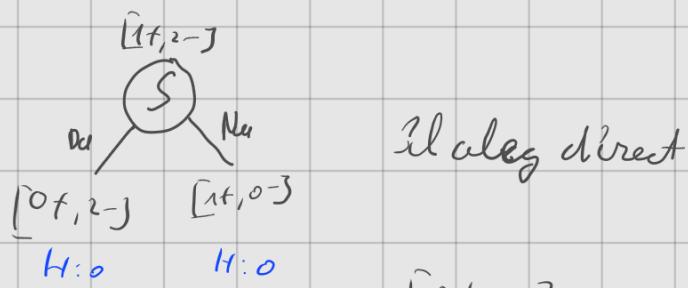
$\Sigma+7-3$

Alegem P



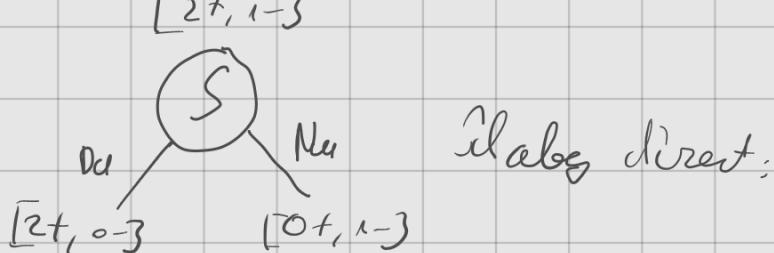
Pentru

SPS



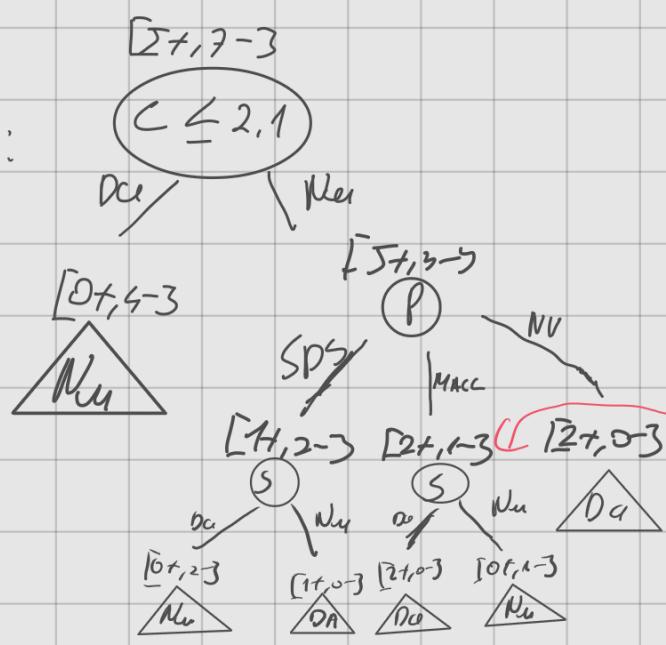
il aleg direct

Pentru MACC



il aleg direct:

Dexi, altim:



corect.



$C \leq 2.6$

alt si candidat

Bayes- Naiv

28.

(Algoritmul Bayes Naiv: aplicare)

* CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, midterm, pr. 6.a

Se dă setul de date din tabelul alăturat, în care A, B, C sunt attribute (de intrare) binare, iar Y este atribut de ieșire.
Care va fi răspunsul algoritmului de clasificare Bayes Naiv pentru intrarea $A = 0, B = 0, C = 1$?

A	B	C	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

406

$$P(A=0, B=0, C=1) = P(A=0, B=0, C=1 | Y=0) \cdot P(Y=0) + P(A=0, B=0, C=1 | Y=1) \cdot P(Y=1) = \\ = \frac{1}{7} \cdot \frac{3}{7} + \frac{1}{7} \cdot \frac{5}{7} = \frac{2}{7} = 1$$

$(A=0, B=0, C=1)$

$$\begin{aligned} y_{MAP} &= \underset{y \in \{0, 1\}}{\text{argmax}} P(Y=y | A=0, B=0, C=1) = \\ &= \underset{y \in \{0, 1\}}{\text{argmax}} \frac{P(A=0, B=0, C=1 | Y=y) \cdot P(Y=y)}{P(A=0, B=0, C=1)} = \\ &\quad P(A=0, B=0, C=1) = 1 \end{aligned}$$

✓ P.P. cod. M. ind. cond.

$$= \underset{y \in \{0, 1\}}{\text{argmax}} P(A=0 | Y=y) \cdot P(B=0 | Y=y) \cdot P(C=1 | Y=y) \cdot P(Y=y)$$

Amen:

$$P_0 = P(A=0 | Y=0) \cdot P(B=0 | Y=0) \cdot P(C=1 | Y=0) \cdot P(Y=0)$$

$$P_0 = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{63}$$

$$P_1 = P(A=0 | Y=1) \cdot P(B=0 | Y=1) \cdot P(C=1 | Y=1) \cdot P(Y=1)$$

$$P_1 = \frac{1}{4} \cdot \cancel{\frac{2}{5}} \cdot \cancel{\frac{2}{5}} \cdot \cancel{\frac{4}{2}} = \frac{1}{28}$$

$$\text{Comparim } P_0 \text{ și } P_1: \frac{2}{63} < \frac{1}{28} \quad \text{Deci: } P_0 < P_1$$

$$\frac{56}{\dots} < \frac{63}{\dots}$$

Ajadar clasificatorul Bayes Naïve nu prezice $Y=1$ pentru instanță ($A=0, B=0, C=1$) cu probabilitatea:

$$P(Y=1 | A=0, B=0, C=1)$$

$$= \frac{P(A=0, B=0, C=1 | Y=1) \cdot P(Y=1)}{P(A=0, B=0, C=1)}$$

$P(A=0, B=0, C=1) \rightarrow$ form. nrak. totală

$$\underline{\underline{L}}_1$$

$$= \frac{P_1}{P_0} =$$

$$= \frac{P(A=0, B=0, C=1 | Y=1) \cdot P(Y=1) + P(A=0, B=0, C=1 | Y=0) \cdot P(Y=0)}{P_1 + P_0}$$

$$= \frac{\frac{2}{63}}{\frac{2}{63} + \frac{1}{28}} = \frac{2}{63} \cdot \frac{28}{56+63}$$

29.

(Algoritmul Bayes Naiv și algoritmul Bayes Optimal: aplicare)

prelucrare de Liviu Ciortuz, după

• * CMU, 2002 fall, Andrew Moore, final exam, pr. 4.b-e

Se dă setul de date alăturat, cu A și B variabile de intrare, iar C variabilă de ieșire.

A	B	C	nr. apariții
0	0	1	3
0	1	0	1
0	1	1	4
1	0	0	5
1	1	0	2
1	1	1	1
			16

- Care este numărul minim de probabilități ce trebuie estimate pentru a putea construi după aceea (pe acest set de date) un clasificator de tip Bayes Naiv? Justificați.
- Similar, pentru clasificatorul Bayes Optimal. Justificați.
- Care este decizia clasificatorului Bayes Naiv pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.
- Care este decizia clasificatorului Bayes Optimal pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.
- Dacă rezultatele obținute la punctele c și d diferă (fie și numai în privința probabilităților cu care sunt luate deciziile), care este explicația? Justificați în mod riguros.

a)

$$u_{NB} = \arg\max_{u_j \in V} \prod_i P(a_i | u_j) P(r_j)$$

Naive Bayes

$$\begin{aligned} P(C=0) & \text{ sau } P(C=1) = 1 - P(C=0) \\ P(A=0|C=0) & \Rightarrow P(A=1|C=0) = 1 - P(A=0|C=0) \\ P(A=0|C=1) & \Rightarrow P(A=1|C=1) = 1 - P(A=0|C=1) \\ P(B=0|C=0) & \Rightarrow P(B=1|C=0) = 1 - P(B=0|C=0) \\ P(B=0|C=1) & \Rightarrow P(B=1|C=1) = 1 - P(B=0|C=1) \end{aligned}$$

Deci, anumit mecanism decorează 5 valori pentru a putea calcula clasificatorul Bayes Naiv complet.

b) Bayes Optimal: $\arg\max_{u_j \in V} \sum_{h_i \in H} P(u_j | h_i) \cdot P(h_i | D)$

$$\begin{aligned} P(C=0) & \quad P(C=1) = 1 - P(C=0) \\ P(A=0, B=0 | C=0) & \quad \} \\ P(A=0, B=1 | C=0) & \quad \} \end{aligned}$$

Se poate determina $P(A=1, B=1 | C=0)$

$$P(A=1, B=0 | C=0)$$

fiecare dintre cele 3 valori

$$P(A=0, B=0 | C=1)$$

Se poate determina $P(A=1, B=1 | C=1)$

$$P(A=0, B=1 | C=1)$$

fiecare dintre cele 3 valori

$$P(A=1, B=0 | C=1)$$

Deci, avem nevoie doar de 7 valori pentru a putea calcula clasificatorul Bayes continuu complet.

c. Care este decizia clasificatorului Bayes Naiv pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.

c)

$$\hat{c}_{NB} = \underset{c \in \{0,1\}}{\operatorname{argmax}} P(C=c | A=0, B=1) = \underset{c \in \{0,1\}}{\operatorname{argmax}} \frac{P(A=0, B=1 | C=c) P(C=c)}{P(A=0, B=1)}$$

Aneam:

$$P(A=0, B=1 | C=c) \cdot P(C=c) \xrightarrow{\text{P.P. de indep. cond.}} P(A=0 | C=c) \cdot P(B=1 | C=c) \cdot P(C=c)$$

$$P_0 : P(A=0 | C=0) \cdot P(B=1 | C=0) \cdot P(C=0) = \frac{1}{8} \cdot \frac{3}{8} \cdot \frac{8}{16} = \frac{3}{128}$$

$$P_1 : P(A=0 | C=1) \cdot P(B=1 | C=1) \cdot P(C=1) = \frac{7}{8} \cdot \frac{5}{8} \cdot \frac{8}{16} = \frac{35}{128}$$

Obo. $P_0 < P_1$

$$\frac{3}{128} < \frac{35}{128} \Rightarrow \text{alegem } c=1 \text{ cu prob.}$$

Deci: $P(C=1 | A=0, B=1) =$

$$P_1 = \frac{P(A=0, B=1 | C=0) \cdot P(C=0)}{P(A=0, B=1 | C=0) + P(A=0, B=1 | C=1)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$$

d. Care este decizia clasificatorului Bayes Optimal pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.

d)

Bayes optimal

$$\hat{C}_B = \underset{c \in \{0,1\}}{\operatorname{argmax}} P(A=0, B=1 | C=c) \cdot P(C=c)$$

$$P_0 = P(A=0, B=1 | C=0) \cdot P(C=0) = \frac{1}{8} \cdot \frac{8}{16} = \frac{1}{16}$$

$$P_1 = P(A=0, B=1 | C=1) \cdot P(C=1) = \frac{4}{8} \cdot \frac{8}{16} = \frac{4}{16}$$

$$P_1 > P_0 \Rightarrow \text{vom alege } C=1$$

e) Ele 2 rezultate diferenții din cursul capituluș că la calcularea alg. Bayes Naïve am presupus **independență condițională**, înțelesă de Bayes certim, iar de aci îndezem că **cel puțin o dependență** dintre atributele de **intrare** și cele de **iesire**.

• * Edinburgh, 2009 fall, C. Williams, V. Lavrenko, tutorial 2, pr. 2

Firma Whizzco decide să implementeze un clasificator de texte. Pentru început, ei vor să clasifice documente aparținând fie clasei *sport* fie clasei *politică*. Ei decid să reprezinte fiecare document ca un vector de atrbute descriind prezența ori absența unor cuvinte-cheie:

0 1 2 3 4 5 6 7
goal, football, golf, defence, offence, wicket, office, strategy.

Datele de antrenament sunt reprezentate folosind o matrice în care fiecare linie este un vector de valori (0 sau 1) pentru cele 8 atrbute.

$xP = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$ % Politica

$xS = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$ % Sport

Folosind algoritmul Bayes Naiv, care este probabilitatea cu care documentul $x = (1, 0, 0, 1, 1, 1, 1, 0)$ va fi clasificat ca aparținând clasei *politică*?

$$X = (1, 0, 0, 1, 1, 1, 1, 0)$$

$$u_{NB} = \arg\max_{u_j \in V} \prod_{i=1}^8 P(a_i | u_j) \cdot P(u_j)$$

$$\hat{x}_{NB} = \arg\max_{x \in \{P, S\}} P(X_0=1 | X=x) \cdot P(X_1=0 | X=x) \cdot P(X_2=0 | X=x) \cdot \\ \cdot P(X_3=1 | X=x) \cdot P(X_4=1 | X=x) \cdot P(X_5=1 | X=x) \cdot P(X_6=1 | X=x) \cdot \\ \cdot P(X_7=0 | X=x) \cdot P(X=x)$$

* P. men. iniț. cern. *

$$P_P(\text{Politico} | (1, 0, 0, 1, 1, 1, 1, 0)) = P(1, 0, 0, 1, 1, 1, 1, 0 | \text{Politico}) \cdot P(\text{Politico})$$

$$P(X_0=1 | X=P) \cdot P(X_1=0 | X=P) \cdot P(X_2=0 | X=P) \\ \cdot P(X_3=1 | X=P) \cdot P(X_4=1 | X=P) \cdot P(X_5=1 | X=P) \cdot P(X_6=1 | X=P) \\ \cdot P(X_7=0 | X=P) \cdot P(X=P) = \frac{2}{3,6} \cdot \left(\frac{5}{6}\right)^4 \cdot \left(\frac{1}{6}\right)^2 \cdot \frac{52}{6} \cdot \frac{6}{13} = 0,0013$$

$$P_S = P(Sport | (1, 0, 0, 1, 1, 1, 1, 0)) = P(1, 0, 0, 1, 1, 1, 1, 0 | Sport) \cdot P(Sport)$$

$$= P(X_0=1 | X=S) \cdot P(X_1=0 | X=S) \cdot P(X_2=0 | X=S) \cdot$$

$$\cdot P(X_3=1 | X=S) \cdot P(X_4=1 | X=S) \cdot P(X_5=1 | X=S) \cdot P(X_6=1 | X=S)$$

$$\cdot P(X_7=0 | X=S) \cdot P(X=S) = \left(\frac{5}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^3 \left(\frac{6}{7}\right)^2 \cdot \frac{1}{2} \cdot \frac{7}{13} = 0.0006$$

Deci, $P_P > P_S \Rightarrow$ documentul X va fi clasificat

către Politică cu probabilitatea:

$$P(Politică | (1, 0, 0, 1, 1, 1, 1, 0))$$

$$P(1, 0, 0, 1, 1, 1, 1, 0 | Politică) \cdot P(Politică) + P(1, 0, 0, 1, 1, 1, 1, 0 | Sport) \cdot P(Sport)$$

$$\frac{P_P}{P_P + P_S} = \frac{0.0013}{0.0006 + 0.0013} = 0.6842$$

68%.