Învăţare automată

— Licenţă, anul III, 2020-2021, examenul parţial I —

Nume student: **Grupa:**

1. (2.7p) **(Extensii ale algoritmului ID3: cazul atributelor care au multe valori)**

● ○ ***prelucrare de Liviu Ciortuz, după***
CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 4.2

Una dintre limitările algoritmului ID3 [LC: ne referim la varianta de bază, fără extensiile pe care le-am discutat ulterior la curs] este faptul că el este prea sensibil la prezenţa atributelor care au un număr mare de valori. De exemplu, dacă fiecare instanţă de antrenament are un ID unic, atunci câştigul de informaţie va fi maxim atunci când folosim acest ID ca atribut de intrare, ceea ce nu este deloc bine pentru faza de generalizare / predicţie, în care putem întâlni adeseori instanţe având ID-uri care nu se regăsesc în datele de antrenament. Algoritmul C4.5 remediază acest aspect din funcţionarea lui ID3 folosind în locul câştigului de informaţie, pentru a evalua atributele, *raportul câştigului de informaţie* (engl., information gain ratio).

Vom nota un atribut de intrare oarecare cu X, iar eticheta sa cu Y. Vă reducem aminte că în algoritmul ID3, alegem pentru nodul curent acel atribut X care maximizează câştigul de informaţie, $IG(X)$:

$$IG(X) = H(Y) - H(Y|X).$$

Acum vom defini noţiunea de [LC: cantitate de] *informaţie la separare* (engl., split information) după cum urmează. Presupunem că avem $|D|$ instanţe ataşate la nodul curent, iar după ce se face testul pe baza [valorii] atributului X, aceste instanţe sunt repartizate la V noduri descendente [din nodul curent]. Presupunând că numărul de instanţe care sunt asociate la aceste noduri-fii sunt respectiv $|D_1|$, $|D_2|$, $\cdots$, $|D_V|$, vom defini *informaţia la separarea* valorilor atributului X astfel:[1]

$$\mathit{SplitInfo}\,(X) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}.$$

Ca şi în cazul entropiei, se va considera, prin convenţie, că $0 \cdot \log_2 0 = 0$.

***Raportul câştigului de informaţie* se defineşte în felul următor:**[2]

$$\mathit{GainRatio}\,(X) = \frac{IG(X)}{\mathit{SplitInfo}\,(X)}.$$

Algoritmul C4.5 (succesorul lui ID3) foloseşte *raportul câştigului de informaţie* pentru a determina „cel mai bun“ atribut de pus în nodul curent. Intuiţia spune că $\mathit{SplitInfo}\,(X)$ actionează ca un „normalizator“ (engl., normalizer), care penalizează atributele care au un număr mare de valori. De exemplu, dacă pentru $\forall i, j$ avem

[1] Se poate observa din formula de definiţie că $\mathit{SplitInfo}\,(X)$ este *entropia* atributului X calculată pentru cele D instanţe asociate la nodul curent. (Ea *nu* este însă nicidecum acelaşi lucru cu *entropia condiţională medie* a atributului X în raport cu variabila de ieşire Y, pe care o folosim la calcularea câştigului de informaţie.)

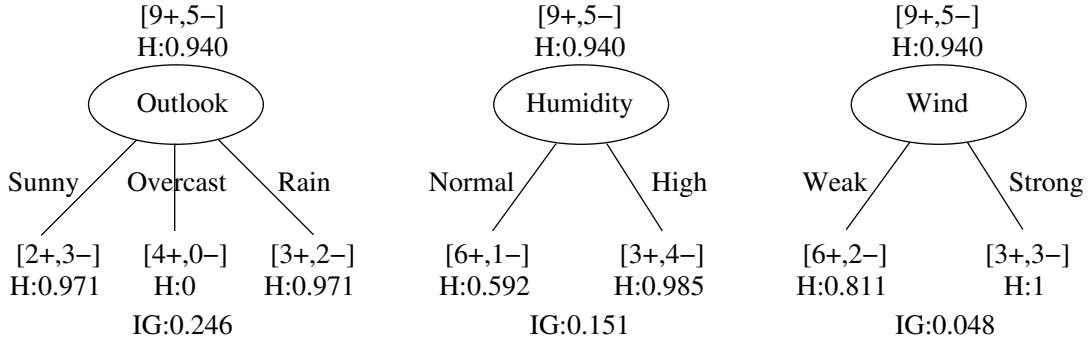
[2] Cazul când $\mathit{SplitInfo}\,(X) = 0$ corespunde situaţiei când X are o singură valoare (pentru instanţele din nodul în care se calculează acest *SplitInfo*), deci nu are putere discriminativă în raport cu variabila de ieşire. În consecinţă, astfel de atribute nu vor fi luate în considerare atunci când, pentru nodul curent, se va pune problema să alegem atributul cu *GainRatio* maxim.

$|D_i| = |D_j|$, **atunci** $SplitInfo(X) = \log_2 V$ **(adică, maximum posibil), aşadar atributele care au mulţimea de valori** (V) **mai restrânsă vor fi preferate.**

În acest exerciţiu veţi elabora arborele de decizie corespunzător setului de date din tabelul de mai jos, folosind pe de o parte algoritmul ID3 şi pe de altă parte algoritmul C4.5. Atributele de intrare sunt *Outlook*, *Temperature*, *Humidity* şi *Wind*, iar eticheta este asociată cu variabila de ieşire *EnjoyTennis*. Veţi trata *Temperature* ca atribut discret.

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	26	High	Weak	No
D2	Sunny	25	High	Strong	No
D3	Overcast	25	High	Weak	Yes
D4	Rain	24	High	Weak	Yes
D5	Rain	19	Normal	Weak	Yes
D6	Rain	20	Normal	Strong	No
D7	Overcast	20	Normal	Strong	Yes
D8	Sunny	23	High	Weak	No
D9	Sunny	20	Normal	Weak	Yes
D10	Rain	25	Normal	Weak	Yes
D11	Sunny	24	Normal	Strong	Yes
D12	Overcast	22	High	Strong	Yes
D13	Overcast	23	Normal	Weak	Yes
D14	Rain	23	High	Strong	No

***Comentariu*:**
Pentru a vă ajuta, vă punem la dispoziţie următoarele reprezentări grafice care să vă ajute la a determina atributul care trebuie pus în nodul rădăcină de către algoritmul ID3:



a. (0.5p)
Vă cerem să desenaţi [în mod similar] un arbore de decizie de adâncime 1, corespunzător atributului *Temperature*. Asociaţi mai întâi la nodul lui rădăcină *partiţia* de instanţe corespunzătoare, iar după aceea procedaţi similar pentru fiecare dintre descendenţii lui direcţi (adică nodurile-fii). Calculaţi apoi entropia condiţională medie a acestui atribut şi, în fine, câştigul de informaţie în raport cu atributul de ieşire *EnjoyTennis*.

***Indicaţie*: Următoarele valori pentru entropia distribuţiei Bernoulli vă pot fi de folos:**

p	0	1/3	2/5	3/7	1/2	1
$H(p)$	0	0.918	0.970	0.985	1	0

b. (0.2p)
Ce atribut va selecta algoritmul ID3 pentru nodul rădăcină al arborelui pe care-l „învaţă“?

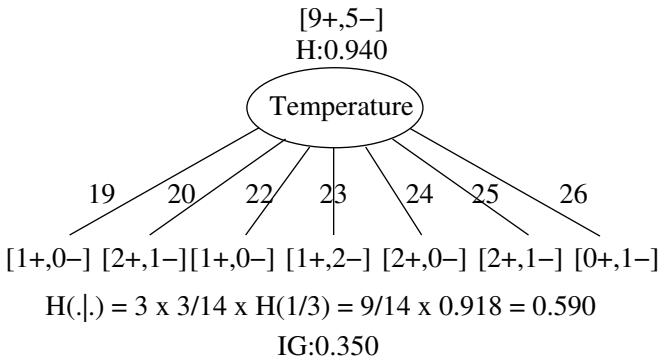
c. (1p, defalcat cf. cu ceea ce urmează)
Calculaţi valorile $SplitInfo(\cdot)$ **pentru toate atributele de intrare, relativ la întregul set de date de antrenament.**
(0.2p)
***Atenţie*!** $SplitInfo(Humidity)$ **şi** $SplitInfo(Wind)$ **se calculează foarte uşor folosind informaţiile din tabelul dat la *Indicaţia* de mai sus.**
(0.5p)
Pentru $SplitInfo(Outlook)$ **şi** $SplitInfo(Temperature)$**, vă cerem elaboraţi calculul cât mai complet posibil.**
(La calcule, puteţi folosi aproximaţiile $\log_2 3 = 1.584$**,** $\log_2 5 = 2.322$ **şi** $\log_2 7 = 2.807$**.)**
(0.3p)
Presupunând că $SplitInfo(Outlook) = 1.577$**, şi** $SplitInfo(Temperature) = 2.646$**, ce atribut va selecta algoritmul C4.5 pentru nodul rădăcină al arborelui pe care-l „învaţă“?**

d. (0.3p)
Elaboraţi complet arborele de decizie produs de algoritmul ID3. (Atenţie! Dacă aţi procedat corect la punctele a **şi** b**, atunci aici nu veţi avea de făcut calcule aproape deloc!)**

e. (0.7p)
Elaboraţi complet arborele de decizie produs de algoritmul C4.5. (Atenţie! Dacă aţi procedat corect la punctul c**, atunci aici nu veţi avea de făcut calcule aproape deloc!)**
Ce observaţi comparând arborele obţinut aici cu arborele care a fost obţinut la punctul d**?**

Răspuns:

a. Pornind de la setul de date de antrenament obţinem:



b. Din rezultatul de la punctul a **şi din informatiile din Comentariul din enunţ, rezultă că**

$$IG(EnjoyTennis, Temperature) > IG(EnjoyTennis, Outlook) > IG(EnjoyTennis, Humidity) > IG(EnjoyTennis, Wind)$$

Aşadar, atributul $Temperature$ **este selectat de către algoritmul ID3 în nodul rădăcină.**

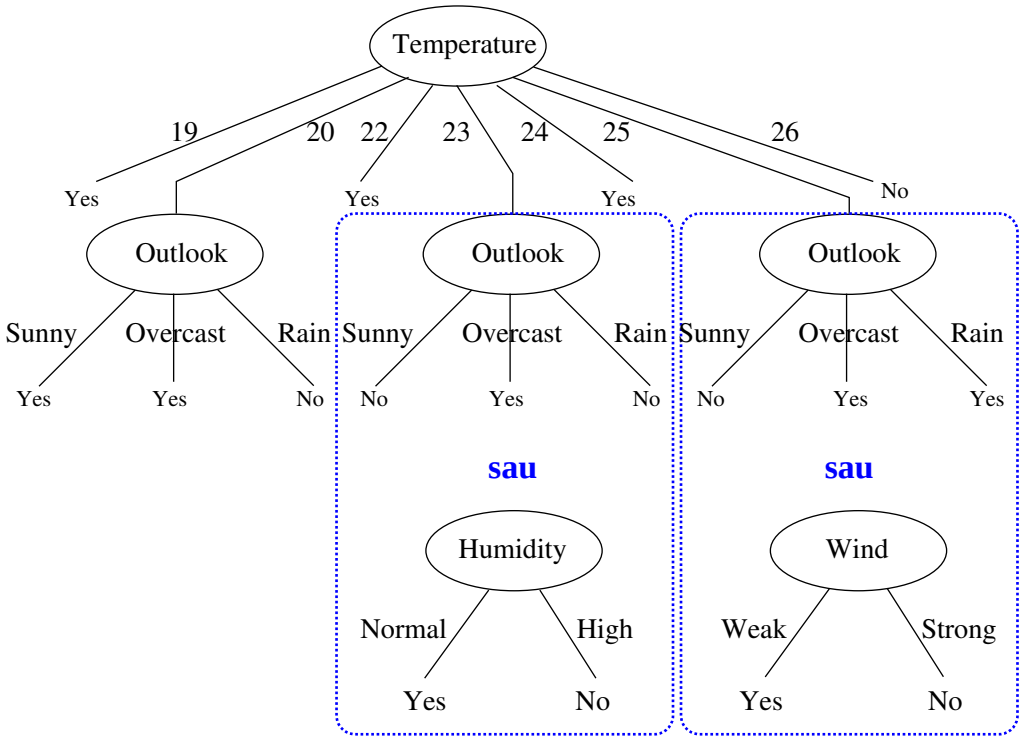
c. Aplicând definiţiile date în enunţ pentru *SplitInfo* şi *Gain Ratio*, vom obţine:

$$\begin{aligned}
SplitInfo(Outlook) &= H[5S;4O;5R] = H\Big(\frac{5}{14},\frac{4}{14},\frac{5}{14}\Big) = 2\cdot\frac{5}{14}\log_2\frac{14}{5}+\frac{4}{14}\log_2\frac{14}{4}\\
&= \log_2 14 - 2\cdot\frac{5}{14}\log_2 5 - \frac{4}{14}\log_2 4 = \frac{3}{7}+\log_2 7-\frac{2}{7}\log_2 5\\
&= 1.577406283\\
SplitInfo(Humidity) &= H[7N;7H] = H(1/2) = 1\\
SplitInfo(Wind) &= H[8L;6H] = H(3/7) = 0.985\\
SplitInfo(Temperature) &= H[1_{19};3_{20};1_{22};3_{23};2_{24};3_{25};1_{26}] = H\Big(\frac{1}{14},\frac{3}{14},\frac{1}{14},\frac{3}{14},\frac{2}{14},\frac{3}{14},\frac{1}{14}\Big)\\
&= 3\cdot\frac{1}{14}\log_2 14+3\cdot\frac{3}{14}\log_2\frac{14}{3}+\frac{2}{14}\log_2\frac{14}{2}\\
&= \log_2 14-\frac{9}{14}\log_2 3-\frac{2}{14} = \frac{6}{7}+\log_2 7-\frac{9}{14}\log_2 3\\
&= 2.645593314
\end{aligned}$$

$$\begin{aligned}
Gain\ Ratio(Outlook) &= \frac{0.246}{1.577} = 0.155\\
Gain\ Ratio(Humidity) &= \frac{0.151}{1} = 0.151\\
Gain\ Ratio(Wind) &= \frac{0.048}{0.985} = 0.0487\\
Gain\ Ratio(Temperature) &= \frac{0.350}{2.646} = 0.132
\end{aligned}$$

$\Rightarrow$ *Outlook* **is selected by the C4.5 algorithm at the root note.**

d. Arborele ID3 este următorul:



e. Folosim rezultatul de la punctul *c* şi primul arbore din *Comentariul* din enunţ. Pentru nodul corespunzător [descendentului din nodul rădăcină care este determinat de testul] lui *Outlook* = *Sunny*:

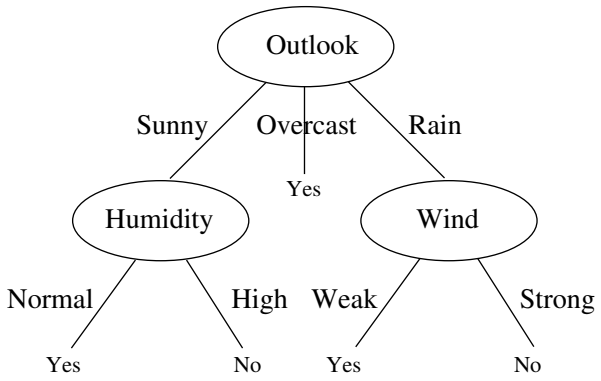
$$\begin{aligned} \mathit{SplitInfo}(\mathit{Temperature}) &= H[1_{20}; 1_{23}; 1_{24}; 1_{25}; 1_{26}] = \log_2 5 = 2.322 \\ \mathit{SplitInfo}(\mathit{Humidity}) &= H[2N; 3H] = H(2/5) = 0.970 \\ \mathit{SplitInfo}(\mathit{Wind}) &= H[3W; 2S] = H(2/5) = 0.970 \end{aligned}$$

Din date, se observă imediat că aici $IG(X, PlayTennis)$ este maxim pentru $X = Humidity$, fiindcă entropia lui condiţională medie este 0.[3] Acest fapt conduce uşor — ţinând cont de valorile pentru *SplitInfo* calculate mai sus — la concluzia că *GainRatio* este maxim pentru atributul *Humidity*.

Pentru nodul corespunzător [descendentului din nodul rădăcină care este determinat de testul] lui *Outlook* = *Rain*:

$$\begin{aligned} \mathit{SplitInfo}(\mathit{Temperature}) &= H[1_{19}; 1_{20}; 1_{23}; 1_{24}; 1_{25}] = \log_2 5 = 2.322 \\ \mathit{SplitInfo}(\mathit{Humidity}) &= H[3N; 2H] = H(2/5) = 0.970 \\ \mathit{SplitInfo}(\mathit{Wind}) &= H[3W; 2S] = H(2/5) = 0.970 \end{aligned}$$

Ca şi mai sus, din date se observă imediat că aici $IG(X, PlayTennis)$ este maxim pentru $X = Wind$, fiindcă entropia lui condiţională medie este 0.[4] Aceasta conduce uşor — ţinând cont de valorile pentru *SplitInfo* calculate mai sus — la concluzia că *GainRatio* este maxim pentru atributul *Wind*.



LC: The two decision trees are quite different. The *Temperature* attribute is not present in the C4.5 tree,[5] while in the ID3 tree it is present at the root node! It can be easily seen that this [attribute] makes some predictions impossible (see for instance the case of *Temperature* = 21). In other cases (*Temperature* = 19, 22, 24 and 26), the prediction is based only on *Temperature*; all the remaining attributes (which normally could influence someone's decision to go to tennis) are ignored!

[3] *Observaţie*: De fapt, ţinând cont de acest fapt, nici nu mai era nevoie să calculăm valorile pentru *SplitInfo*. Era suficient să observăm că valorile pentru *SplitInfo* pentru *Wind* şi *Humidity* sunt egale şi mai mici decât 1, în vreme ce *SplitInfo*(*Temperature*) este maxim şi deci mai mare decât 1, întrucât acest atribut ia aici mai mult de 2 valori.

[4] Aceeşi *Observaţie* ca mai sus este valabilă şi aici. Chiar mai mult, se observă că la nivel de *structură*, în comparaţie cu situaţia de mai sus, *Wind* şi *Humidity* au roluri schimbate!

[5] LC: Este o pură coincidenţă faptul că arborele C4.5 care a fost învăţat pe datele de aici coincide cu arborele ID3 învăţat pe datele de antrenament din cartea *Machine Learning* de Tom Mitchell (vedeţi pag. 53 şi 59).

2. **(Algoritmii Bayes Naiv şi Bayes Optimal: aplicare)**

prelucrare de Liviu Ciortuz, după

• ○ *CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 3*

Se dă setul de date de antrenament din tabelul alăturat, cu x, y şi z variabile de intrare şi U variabila de ieşire.
Presupunem că trebuie să preziceţi ieşirea U folosind clasificatorul Bayes Naiv (fără a aplica regula de „netezire“ a probabilităţilor (“add-one”) a lui Laplace).

x	y	z	U
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

a. După terminarea antrenării, care va fi valoarea prezisă pentru probabilitatea $P(U = 0|x = 0, y = 1, z = 0)$? Care este deci eticheta prezisă de clasificator pentru instanţa de test $(x = 0, y = 1, z = 0)$?

b. Care va fi valoarea prezisă pentru probabilitatea $P(U = 0|x = 0)$? Care este deci eticheta prezisă de clasificator pentru instanţa de test $x = 0$?

Pentru următoarele două puncte, vom presupune că „învăţaţi“ clasificatorul Bayes Optimal. În acest caz,

c. (0.2p)
care va fi valoarea prezisă pentru probabilitatea $P(U = 0|x = 0, y = 1, z = 0)$? Care este deci eticheta prezisă de clasificator pentru instanţa de test $(x = 0, y = 1, z = 0)$?

d. (0.1p)
care va fi valoarea prezisă pentru probabilitatea $P(U = 0|x = 0)$? Care este deci eticheta prezisă de clasificator pentru instanţa de test $x = 0$?

Răspuns:

a.

$$
\begin{aligned}
&P(U=0|x=0,y=1,z=0)\\
&= \frac{P(x=0,y=1,z=0|U=0)\cdot P(U=0)}{P(x=0,y=1,z=0)}\\
&= \frac{P(x=0|U=0)\cdot P(y=1|U=0)\cdot P(z=0|U=0)\cdot P(U=0)}{P(x=0,y=1,z=0|U=0)\cdot P(U=0)+P(x=0,y=1,z=0|U=1)\cdot P(U=1)}\\
&= \frac{\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{3}{7}}{\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{3}{7}+\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4}\cdot\frac{4}{7}} = \frac{\frac{2}{3^3}}{\frac{2}{3^3}+\frac{1}{4}} = \frac{8}{8+27} = \frac{8}{35} = 0.229 < 1/2.
\end{aligned}
$$

Eticheta prezisă de clasificator pentru instanţa de test $(x = 0, y = 1, z = 0)$ este $U = 1$.

b. $P(U=0|x=0) = \dfrac{P(x=0|U=0)\cdot P(U=0)}{P(x=0|U=0)\cdot P(U=0)+P(x=0|U=1)\cdot P(U=1)} = \dfrac{\frac{2}{7}}{\frac{2}{7}+\frac{2}{7}} = \dfrac{1}{2}.$

Sau, direct: $P(U=0|x=0) = \dfrac{P(x=0,U=0)}{P(x=0)} = \dfrac{\frac{2}{7}}{\frac{4}{7}} = \dfrac{1}{2}.$

Eticheta prezisă de clasificator pentru instanţa de test $x = 0$ este fie $U = 1$ fie $U = 0$ (cu probabilitate de $1/2$).

c.

$$\begin{aligned} & P(U=0|x=0,y=1,z=0) \\ = \; & \frac{P(x=0,y=1,z=0|U=0)\cdot P(U=0)}{P(x=0,y=1,z=0|U=0)\cdot P(U=0)+P(x=0,y=1,z=0|U=1)\cdot P(U=1)} \\ = \; & \frac{0\cdot\frac{3}{7}}{0\cdot\frac{3}{7}+\frac{1}{4}\cdot\frac{4}{7}}=0<1/2. \end{aligned}$$

Sau, direct: $P(U=0|x=0,y=1,z=0)=\dfrac{P(x=0,y=1,z=0,U=0)}{P(x=0,y=1,z=0)}=\dfrac{0}{\frac{1}{7}}=0$**.**

Eticheta prezisă de clasificator pentru instanţa de test $(x=0,y=1,z=0)$ **este** $U=1$ **(cu probabilitatea 1).**

d. $P(U=0|x=0)=\dfrac{1}{2}$**. (Nu este nicio diferenţă faţă de punctul** b**, fiindcă avem un singur atribut de intrare, deci nu se pune problema aplicării presupoziţiei de independenţă condiţională.) Eticheta prezisă de clasificator pentru instanţa de test** $x=0$ **este fie** $U=1$ **fie** $U=0$ **(cu probabilitate de** $1/2$**).**

3. (0.9p+0.75p) **(O corespondenţă (interesantă, dar imperfectă!) între regulile de decizie ale clasificatorilor K-NN şi Bayes Optimal)**

prelucrare de Liviu Ciortuz, după
• ◦ CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 3.b

În acest exerciţiu vom arăta că *regula de decizie* a algoritmului K-NN poate avea şi o *interpretare probabilistă*.

Presupunem că avem un set de date format din N_k puncte (adică, instanţe de antrenament) din clasa C_k, pentru $k = 1, \ldots, m$, unde m este numărul de clase cu care se lucrează, iar N este numărul total de instanţe de antrenament. Pentru un punct nou x (adică, o instanţă de test x), vom determina o sferă în jurul lui x în aşa fel încât ea să conţină exact cei mai apropiaţi K vecini ai lui x. Vom presupune că această sferă are volumul V şi că ea conţine n_k puncte din clasa C_k.

În aceste condiţii, putem estima probabilitatea a priori a fiecărei clase C_k astfel:

$$P(C_k) = \frac{N_k}{N}.$$

Putem de asemenea să estimăm funcţia de densitate condiţionată [a lui x] în raport cu clasa C_k astfel:

$$P(x|C_k) = \frac{n_k}{N_k V}.$$

În mod similar, putem să estimăm o funcţie de densitate de probabilitate (p.d.f.), notată cu $P(x)$, în felul următor:[6]

$$P(x|C_k) = \frac{n_k}{N_k V}.$$

***Observaţie*: Aceste mărimi / probabilităţi sunt calculate pentru o mică regiune situată în jurul punctului de test x, iar mărimea acestei zone depinde de distribuţia datelor de test.**

a. (0.9p: 0.6p pt. calculul lui $P(C_k|x)$ şi 0.3p pt. aplicarea operatorului $\arg\max$)
Demonstraţi că în acest model probabilist, *regula de decizie* de tip Bayes [LC: Comun / Optimal] va produce exact aceeaşi clasificare ca şi algoritmul K-NN.

(În *concluzie*, algoritmul K-NN încearcă să aproximeze regula de decizie de tip Bayes [Comun / Optimal] pe un subset din datele de antrenament.)

b. [Bonus] (0.75p)
Arătaţi că funcţia densitate de probabilitate $P(x)$ din acest model nu este [LC: întotdeauna] bine definită, integrala sa pe întreg spaţiul [de instanţe] nefiind în mod neapărat egală cu 1.

***Sugestie*: Este suficient să identificaţi un singur caz particular [de set de date de antrenament] pentru care să demonstraţi acest fapt. Vă sugerăm să lucraţi în $\mathbb{R}$ (deci cu un singur atribut de intrare) şi să observaţi când anume volumul V poate fi 0.**

[6] Formula de definiţie pentru $P(x)$ se justifică în mod natural folosind formula probabilităţii totale:

$$P(x) = \sum_{k=1}^{m} P(x|C_k) \cdot P(C_k) = \sum_{k=1}^{m} \frac{n_k}{N_k V} \cdot \frac{N_k}{N} = \frac{1}{NV} \sum_{k=1}^{m} n_k = \frac{K}{NV}.$$

Răspuns:

a.

$$P(C_k|x) \overset{F.\ Bayes}{=} \frac{P(x|C_k)\cdot P(C_k)}{P(x)} = \frac{\frac{n_k}{\cancel{N_k}\cdot\cancel{V}}\cdot\frac{\cancel{N_k}}{\cancel{N}}}{\frac{K}{\cancel{N}\cdot\cancel{V}}} = \frac{n_k}{K}.$$

Therefore,

$$\arg\max_k P(C_k|x) = \arg\max_k \frac{n_k}{K},$$

which is exactly the decision rule of K-NN.

b. [As suggested, we will] Consider one-dimension data $\mathbb{R}$ classification by 1-NN. Assume [also] that we have only one training data point located at $x_1 = 0$. Then $p(x) = \dfrac{1}{|x|}$ (LC: which is not properly defined when the "new" data point x from the problem's statement is also 0),[7] and

$$\int_{-\infty}^{+\infty} p(x)\,dx = 2\int_0^{+\infty} p(x)\,dx = 2\int_0^{+\infty} (\ln x)'\,dx = 2\ln x\,|_0^{+\infty} = 2(+\infty - (-\infty)) = \infty \neq 1.$$

[7] Integrala $\int_{-\infty}^{+\infty} p(x)\,dx$ este o *integrală improprie*, fiindcă funcţia $p(x)$ nu este definită în punctul $x = 0$. (LC: Mulţumesc domnului Adrian Zălinescu pentru această precizare.) Ea se poate scrie în mod explicit ca $\int_{-\infty}^{0} p(x)\,dx + \int_0^{+\infty} p(x)\,dx$.

4. (1.5p)

(Comparaţii între algoritmii 1-NN şi ID3:
[o clasă de] seturi de date de antrenament pe care
cei doi clasificatori obţin rezultate identice)

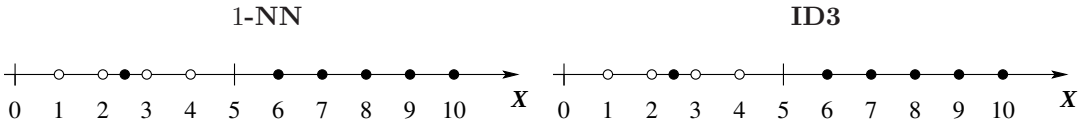
prelucrare de Liviu Ciortuz, după
• ◦ CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.1
CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 3.a

Considerăm instanţele de antrenament $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ având respectiv etichetele $y_1, y_2, \ldots, y_n$.

***Întrebare*: Este oare posibil ca aplicând algoritmul ID3 cu atribute numerice continue să obţinem aceleaşi rezultate la testare / generalizare (deci şi aceleaşi zone de decizie) ca şi cele produse de clasificatorul 1-NN folosind distanţa euclidiană?**

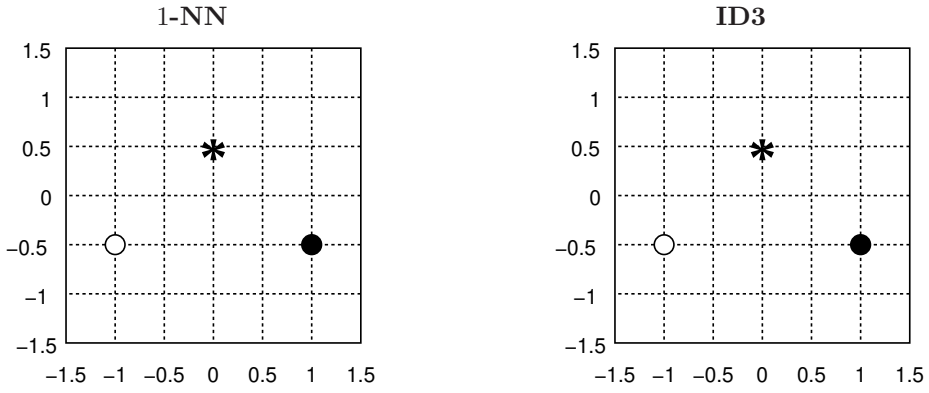
Veţi aborda separat cazurile $d = 1$ şi $d > 1$.

a. (0.5p, defalcat cf. cu ceea ce urmează)
În cazul $d = 1$, pentru a înţelege mai bine cerinţa problemei, pentru setul de date de mai jos trasaţi diagrama Voronoi pe figura din stânga. Veţi identifica în mod clar separatorii decizionali.
(0.3p = 2 × 0.15p)
Pe figura din dreapta, desenaţi graniţele de decizie şi zonele de decizie determinate de algoritmul ID3. (Vă readucem aminte convenţia noastră: semnul ◦ desemnează un exemplu negativ, iar semnul • desemnează un exemplu pozitiv.)



(0.2p)
Ce concluzie puteţi trage referitor la calculul erorilor de tip CVLOO pentru cei doi clasificatori (1-NN şi ID3 cu atribute numerice continue) pe seturi de date din $\mathbb{R}$?

b. (0.75p = 0.15p × 5 diagrame)
În cazul $d > 1$, procedaţi similar pentru setul de date reprezentat mai jos, unde simbolii ◦, • şi $*$ denotă trei clase diferite. Adică, trasaţi diagrama Voronoi pe figura din stânga, iar pe figura din dreapta desenaţi graniţele de decizie şi zonele de decizie determinate de algoritmul ID3 (justificaţi riguros!).



În cazul algoritmului ID3, rezultatul este unic determinat? Dacă da, explicaţi de ce. Dacă nu, arătaţi câte variante se pot obţine în total.

c. (0.25p)
Ce răspuns puteţi formula acum referitor la *Întrebarea* de mai sus, din enunţ? Încercaţi să *generalizaţi*, pornind de la exemplele de la punctele a şi b.

Răspuns:

• CMU:

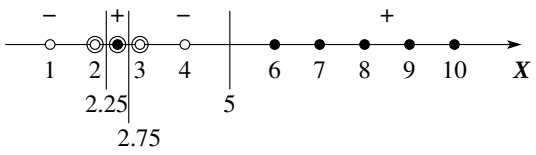
If $d = 1$, we can compute the distance between two points by using only one feature, so we can build a decision tree with the same decision boundary as 1-NN.

If $d > 1$, the decision boundary of 1-NN lies on the Voronoi diagram edges of the training points. A 2-D illustration can be seen from Figure.... On the other hand, the decision boundary of a decision tree consists of piecewise hyper-planes (lines for $d = 2$) parallel to feature axes. A 2-D illustration is shown in Figure... Hence decision tree cannot behave exactly the same as 1-NN in general.

• LC:

a.

atât pentru 1-NN cât şi pentru ID3:



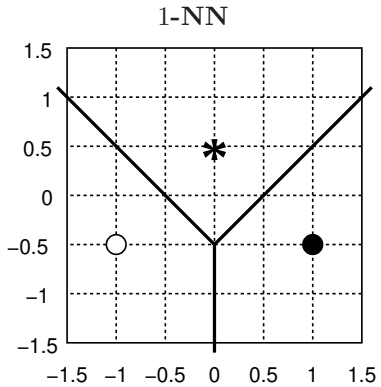
În această figură am încercuit exemplele care conduc la eroare la cross-validarea de tip "Leave-One-Out", atât pentru algoritmul 1-NN cât şi pentru algoritmul ID3.

Se constată că întotdeauna, pe orice set de date de antrenament din $\mathbb{R}$ (deci cu un singur atribut de intrare, care este numeric şi continuu), atât algoritmul 1-NN cât şi algoritmul ID3 produc aceleaşi rezultate la testare / generalizare.

Explicaţia ţine de folosirea distanţei euclidiene la algoritmul 1-NN şi respectiv modul cum se stabilesc pragurile de splitare de către algoritmul ID3 (şi anume, la jumătatea distanţei dintre două instanţe consecutive pe axa reală, care au etichete diferite).

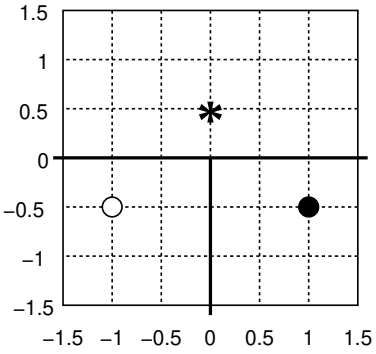
În consecinţă, pe date de antrenament din $\mathbb{R}$, cei doi algoritmi produc exact aceleaşi zone de decizie şi aceiaşi separatori decizionali, precum şi aceleaşi erori la CVLOO.

b.

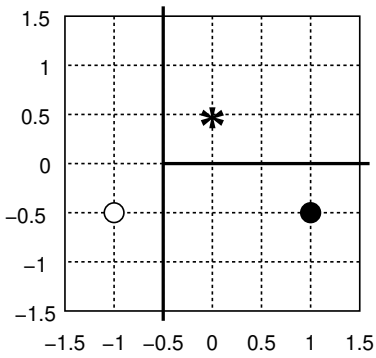


În cazul algoritmului ID3, split-urile sunt $x = -0.5$, $x = +0.5$, şi $y = 0$, iar compaşii de decizie care se formează au toţi acelaşi tip de structură. (De exemplu: unul dintre ei are partiţia [1-,1+,1*] în rădăcină, iar descendenţii lui au partiţiile [1-,0+,0*] şi respectiv [0-,1+,1*]. Ceilalţi compaşi de decizie sunt uşor de „vizualizat".) Asta face să se obţină mai multe soluţii (a se vedea mai jos). Toate aceste soluţii comportă zone de decizie diferite de cele determinate de algoritmul 1-NN (vedeţi diagrama Voronoi de mai sus).

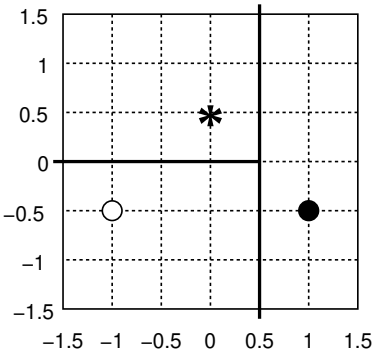
ID3, sol. 1



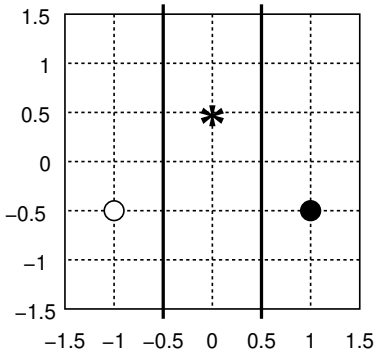
ID3, sol. 2



ID3, sol. 3



ID3, sol. 4



c. Spre deosebire de cazul $d = 1$, atunci când $d > 1$ se constată că în general algoritmul 1-NN şi algoritmul ID3 pot produce la testare / generalizare rezultate diferite (deci şi zone de decizie diferite).

5. (Bonus: 1.25p) **(Probabilităţi condiţionate, formula lui Bayes; proprietatea de aditivitate numărabilă. O dilemă cu trei deţinuţi şi un gardian; ipoteze MAP.)**

prelucrare de Liviu Ciortuz, după
• ◦ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 3

Trei deţinuţi din SUA — desemnaţi aici în mod simplu prin literele a, b, şi c — au fost condamnaţi la moarte şi aşteptau să fie executaţi. Guvernatorul statului în care se afla închisoarea respectivă a decis să-l graţieze pe unul dintre acesti trei deţinuţi (adică, să-i comute pedeapsa cu moartea) şi a ales în mod uniform aleatoriu pe care anume dintre ei să îl graţieze. Guvernatorul l-a informat pe gardianul închisorii despre decizia sa, precum şi despre rezultatul alegerii sale aleatorii, însă i-a cerut să păstreze secretul.

După ce în închisoare au început să circule zvonuri despre graţiere, deţinutul a a încercat să-l convingă pe gardian să-i spună care este soarta lui. Gardianul a refuzat. Atunci deţinutul a i-a cerut gardianului să-i spună cine — mai precis, *unul* — dintre deţinuţii b sau c care va fi executat. Gardianul s-a gândit un pic şi apoi i-a spus că deţinutul b va fi executat.

***Indicaţie importantă*: La rezolvarea exerciţiului va trebui să presupunem că, răspunzând la întrebarea deţinului a, gardianul a ales în mod aleatoriu — şi uniform, dacă a fost cazul să aleagă între mai multe posibilităţi —, şi că el a respectat însă atât adevărul cât şi cerinţa formulată de guvernator (înţelesă într-un sens mai lax, şi anume că *gardianul nu are voie să comunice niciunui deţinut soarta sa, în mod direct*).**

a. (0.95p, defalcat cf. cu ceea ce găsiţi în rezolvare: i: 0.15p, ii: 0.5p, iii: 0.3p,)
Fie $X = a$, respectiv b sau c evenimentul care reprezintă faptul că deţinutul a, respectiv b sau c a fost graţiat.

Precizaţi care sunt valorile numerice pentru probabilităţile a priori $P(X = a)$, $P(X = b)$ şi $P(X = c)$.

Notăm cu $Y = b$ evenimentul (comunicat de gardian) că deţinutul b urmează să moară (adică, să fie executat).
Calculaţi $P(X = a|Y = b)$. În urma obţinerii de către deţinutul a informaţiei adiţionale (de la gardian) că deţinutul b va muri, a crescut oare probabilitatea ca el (deţinutul a) să supravieţuiască?

***Indicaţie* (1): Comparaţi probabilitatea a posteriori $P(X = a|Y = b)$ cu probabilitatea a priori $P(X = a)$. În prealabil veţi completa un tabel în care veţi specifica valorile tuturor probabilităţilor condiţionate $P(Y = y|X = x)$, cu x şi $y \in \{a, b, c\}$.**

	$P(Y = a\|X)$	$P(Y = b\|X)$	$P(Y = c\|X)$
$X = a$			
$X = b$			
$X = c$			

b. (0.3p)
Presupunem că deţinutul a a comunicat toate cele de mai sus deţinutului c. Arătaţi că probabilitatea deţinutului c de a supravieţui a devenit acum $2/3$.

***Indicaţie* (2): Demonstraţi că $P(X = c|Y = b) = 2/3$.**

Răspuns:

a.
(*i.*) **Probabilităţile a priori — pentru cine anume va fi graţiat — sunt** $P(X = a) = P(X = b) = P(X = c) = 1/3$**.**

(*ii.*) **Probabilităţile a posteriori — pentru cine anume va fi graţiat, dacă ştim răspunsul gardianului — sunt cele date în tabelul următor. Ele au fost calculate în funcţie de datele / specificaţiile din enunţul problemei.**

	$P(Y = a\|X)$	$P(Y = b\|X)$	$P(Y = c\|X)$
$X = a$	0	1/2	1/2
$X = b$	0	0	1
$X = c$	0	1	0

(*iii.*) **Probabilitatea a posteriori** $P(X = a \mid Y = b)$ **poate fi calculată astfel:**

$$P(X = a \mid Y = b) \overset{F.\ Bayes}{=} \frac{P(Y = b \mid X = a) \cdot P(X = a)}{P(Y = b)}$$

$$\overset{F.P.T.}{=} \frac{P(Y = b \mid X = a) \cdot P(X = a)}{P(Y = b \mid X = a) \cdot P(X = a) + P(Y = b \mid X = b) \cdot P(X = b) + P(Y = b \mid X = c) \cdot P(X = c)}$$

$$\overset{prob.\ a\ priori}{=} \frac{P(Y = b \mid X = a)}{P(Y = b \mid X = a) + P(Y = b \mid X = b) + P(Y = b \mid X = c)}$$

$$= \frac{\frac{1}{2}}{\frac{1}{2} + 0 + 1} = \frac{1}{3}.$$

Aşadar, răspunsul furnizat de gardian nu a condus la creşterea probabilităţii ca deţinutul a **să supravieţuiască.**

b. Probabilitatea a posteriori $P(X = c \mid Y = b)$ **poate fi calculată la fel ca mai sus sau, chiar mai simplu, astfel: Observăm că** $P(X = b \mid Y = b) = 0$ **(datorită cerinţelor guvernatorului), deci** $P(X = c \mid Y = b) = 1 - P(X = a \mid Y = b) = \frac{2}{3}$**.**

Prin urmare, ***ipoteza*** **ca deţinutul** c **să fie graţiat este** ***mai probabilă a posteriori*** **decât ipoteza ca deţinutul** a **să fie graţiat. Aşadar, în acest exerciţiu am identificat care este ipoteza de probabilitate maximă a posteriori (MAP) ...în acelaşi mod / spirit în care se lucrează în clasificarea bayesiană.**

Observaţie**: (indicaţie pentru corectarea lucrărilor)**
Ca şi în raţionamentul precedent referitor la valoarea lui $P(X = c \mid Y = b)$**, din** ***Indicaţia*** **(2) din enunţ se poate deduce că**

$$P(X = a \mid Y = b) = 1 - P(X = c \mid Y = b) = 1 - \frac{2}{3} = \frac{1}{3}.$$

Acesta trebuie socotit ca fiind un ***răspuns valid*** **pentru ultima parte (*iii.*) a punctului** a**.**