

Învățare automată

— Licență, anul III, 2020-2021, examenul parțial 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 prezența atributelor care au un număr mare de valori. De exemplu, dacă fiecare instanță de antrenament are un ID unic, atunci câștigul de informație va fi maxim atunci când folosim acest ID ca atribut de intrare, ceea ce nu este deloc bine pentru faza de generalizare / predicție, în care putem întâlni adeseori instanțe având ID-uri care nu se regăsesc în datele de antrenament. Algoritmul C4.5 remediază acest aspect din funcționarea lui ID3 folosind în locul câștigului de informație, pentru a evalua atributele, *raportul câștigului de informație* (engl., information gain ratio).

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

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

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

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

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

Raportul câștigului de informație se definește în felul următor:²

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

Algoritmul C4.5 (succesorul lui ID3) folosește *raportul câștigului de informație* pentru a determina „cel mai bun“ atribut de pus în nodul curent. Intuiția spune că $SplitInfo(X)$ acționează ca un „normalizator“ (engl., normalizer), care penalizează atributele care au un număr mare de valori. De exemplu, dacă pentru $\forall i, j$ avem

¹Se poate observa din formula de definiție că $SplitInfo(X)$ este *entropia* atributului X calculată pentru cele D instanțe asociate la nodul curent. (Ea *nu* este însă nicidecum același lucru cu *entropia condițională medie* a atributului X în raport cu variabila de ieșire Y , pe care o folosim la calcularea câștigului de informație.)

²Cazul când $SplitInfo(X) = 0$ corespunde situației când X are o singură valoare (pentru instanțele din nodul în care se calculează acest $SplitInfo$), deci nu are putere discriminativă în raport cu variabila de ieșire. În consecință, astfel de atribute nu vor fi luate în considerare atunci când, pentru nodul curent, se va pune problema să alegem atributul cu $GainRatio$ maxim.

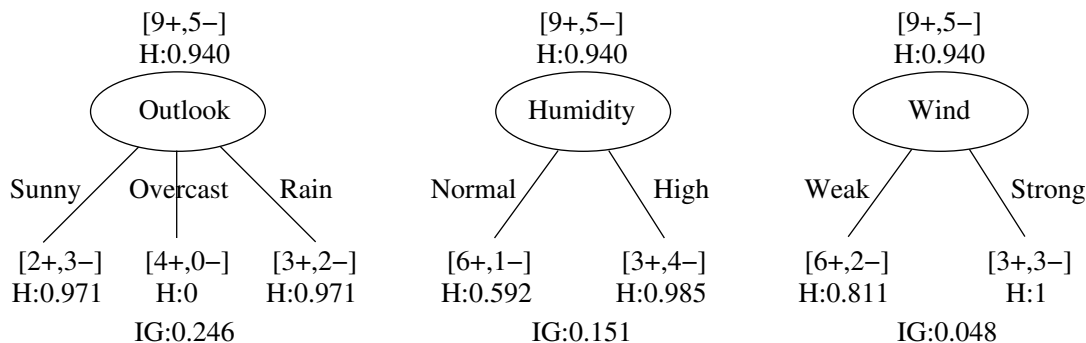
$|D_i| = |D_j|$, atunci $SplitInfo(X) = \log_2 V$ (adică, maximum posibil), așadar atributele care au mulțimea de valori (V) mai restrânsă vor fi preferate.

În acest exercițiu veți elabora arborele de decizie corespunzător setului de date din tabelul de mai jos, folosind pe de o parte algoritmul ID3 și pe de altă parte algoritmul C4.5. Atributele de intrare sunt *Outlook*, *Temperature*, *Humidity* și *Wind*, iar eticheta este asociată cu variabila de ieșire *EnjoyTennis*. Veți 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 dispoziție 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 desenați [în mod similar] un arbore de decizie de adâncime 1, corespunzător atributului *Temperature*. Asociați mai întâi la nodul lui rădăcină *partiția* de instanțe corespunzătoare, iar după aceea procedați similar pentru fiecare dintre descendenții lui direcți (adică nodurile-fi). Calculați apoi entropia condițională medie a acestui atribut și, în fine, câștigul de informație în raport cu atributul de ieșire *EnjoyTennis*.

Indicație: Următoarele valori pentru entropia distribuției 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 „învață“?

c. (1p, defalcat cf. cu ceea ce urmează)

Calculați valorile $SplitInfo(\cdot)$ pentru toate atributele de intrare, relativ la întregul set de date de antrenament.

(0.2p)

Atenție! $SplitInfo(Humidity)$ și $SplitInfo(Wind)$ se calculează foarte ușor folosind informațiile din tabelul de mai sus.

(0.5p)

Pentru $SplitInfo(Outlook)$ și $SplitInfo(Temperature)$, vă cerem elaborați calculul cât mai complet posibil.

(0.3p)

Presupunând că $SplitInfo(Outlook) = 1.577$, și $SplitInfo(Temperature) = 2.646$, ce atribut va selecta algoritmul C4.5 pentru nodul rădăcină al arborelui pe care-l „învață“?

d. (0.3p)

Elaborați complet arborele de decizie produs de algoritmul ID3. (Atenție! Dacă ați procedat corect la punctele a și b, atunci aici nu veți avea de făcut calcule aproape deloc!)

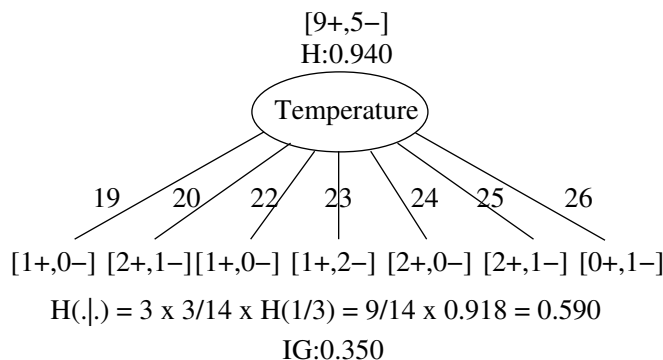
e. (0.7p)

Elaborați complet arborele de decizie produs de algoritmul C4.5. (La calcule, puteți folosi aproximațiile $\log_2 3 = 1.584$, $\log_2 5 = 2.322$ și $\log_2 7 = 2.807$.)

Ce observați comparând arborele obținut aici cu arborele care a fost obținut la punctul d?

Răspuns:

a-b.



\Rightarrow *Temperature* is selected by the ID3 algorithm at the root node.

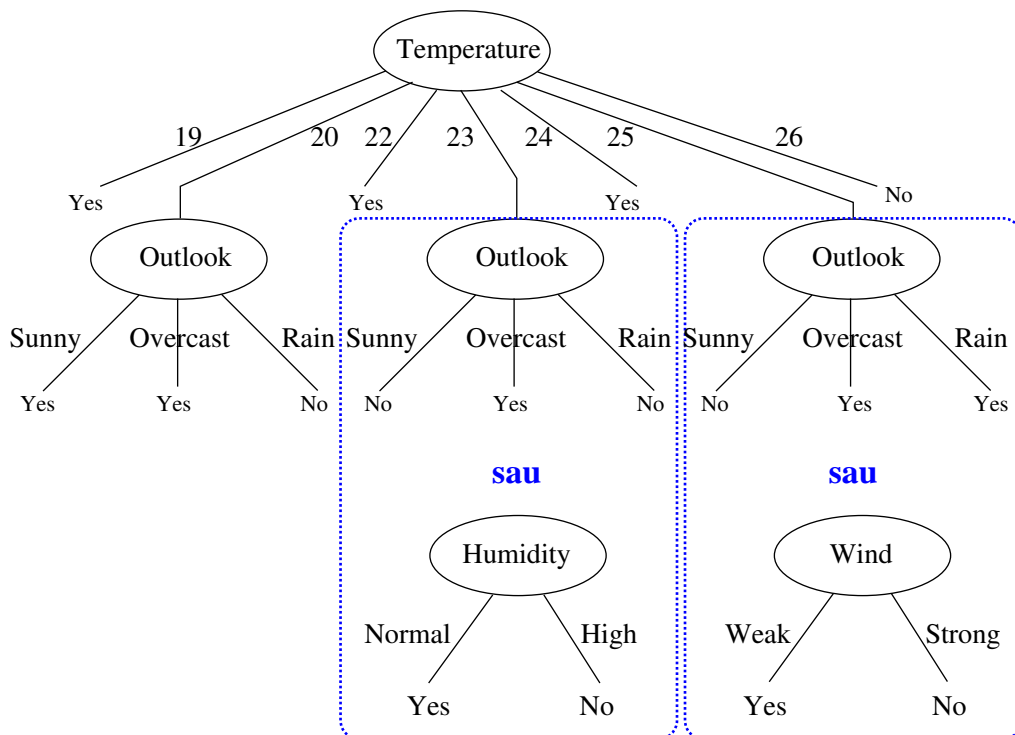
C.

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

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

⇒ *Outlook* is selected by the C4.5 algorithm at the root node.

d.



e. Pentru nodul corespunzător [descendentului din nodul rădăcină care este determinat de testul] lui $Outlook = Sunny$:

$$SplitInfo(Temperature) = H[1_{20}; 1_{23}; 1_{24}; 1_{25}; 1_{26}] = \log_2 5 = 2.322$$

$$SplitInfo(Humidity) = H[2N; 3H] = H(2/5) = 0.970$$

$$SplitInfo(Wind) = H[3W; 2S] = H(2/5) = 0.970$$

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

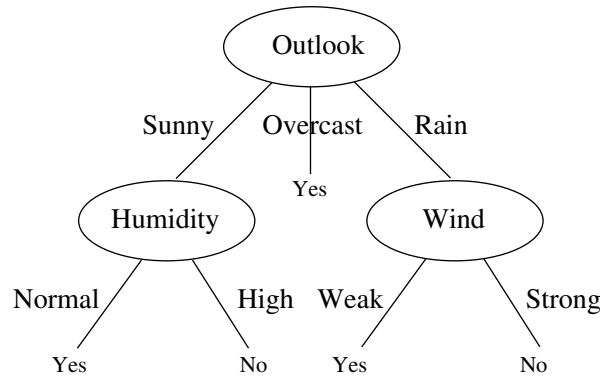
Pentru nodul corespunzător [descendentului din nodul rădăcină care este determinat de testul] lui $Outlook = Rain$:

$$SplitInfo(Temperature) = H[1_{19}; 1_{20}; 1_{23}; 1_{24}; 1_{25}] = \log_2 5 = 2.322$$

$$SplitInfo(Humidity) = H[3N; 2H] = H(2/5) = 0.970$$

$$SplitInfo(Wind) = H[3W; 2S] = H(2/5) = 0.970$$

Ca și mai sus, din date se observă imediat că aici $IG(X, PlayTennis)$ este maxim pentru $X = Wind$, fiindcă entropia lui condițională medie este 0. Aceasta conduce ușor — ținând 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,³ while in the ID3 tree it is present at the root node! It can be 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!

³LC: Este o pură coincidență faptul că arborele C4.5 care a fost învățat pe datele de aici coincide cu arborele ID3 învățat pe datele de antrenament din cartea *Machine Learning* de Tom Mitchell (vedeți pag. 53 și 59).

2. (0.9p)

(Algoritmii Bayes Naiv și 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 și z variabile de intrare și U variabila de ieșire.

Presupunem că trebuie să prezicem ieșirea U folosind clasificatorul Bayes Naiv (fără a aplica regula de „netezire” a probabilităților (“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. (0.4p)

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 instanța de test $(x = 0, y = 1, z = 0)$?

b. (0.2p)

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

Pentru următoarele două puncte, vom presupune că „învățați” 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 instanța 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 instanța 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 instanța de test $(x = 0, y = 1, z = 0)$ este $U = 1$.

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

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

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

c.

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

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

d. $P(U = 0|x = 0) = \frac{1}{2}$. (Nu este nicio diferență față de punctul b, fiindcă avem un singur atribut de intrare, deci nu se pune problema aplicării presupuziției de independență condițională.) Eticheta prezisă de clasificator pentru instanța de test $x = 0$ este fie $U = 1$ fie $U = 0$ (cu probabilitate de $1/2$).

3. (0.9p+0.75p)

(O corespondență (interesantă, dar imperfectă!) între regulile de decizie ale clasificatorilor K -NN și Bayes Naiv)

prelucrare de Liviu Ciortuz, după

• ◦ CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 3.b

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

Presupunem că avem un set de date format din N_k puncte (adică, instanțe de antrenament) din clasa C_k , pentru $k = 1, \dots, m$, unde m este numărul de clase cu care se lucrează. Pentru un punct nou x (adică, o instanță de test x), vom determina o sferă în jurul lui x în așa fel încât ea să conțină exact cei mai apropiați K vecini ai lui x . Vom presupune că această sferă are volumul V și că ea conține n_k puncte din clasa C_k .

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

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

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

$$P(x) = \frac{K}{NV}.$$

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

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

Observație: Aceste mărimi / probabilități sunt calculate pentru o mică regiune situată în jurul punctului de test x , iar mărimea acestei zone depinde de distribuția datelor de test.

a. (0.9p: 0.6p pt. calculul lui $P(C_k|x)$ și 0.3p pt. aplicarea operatorului $\arg \max$)

Demonstrați că în acest model probabilist, *regula de decizie* de tip Bayes [LC: Comun / Optimal] va produce exact aceeași clasificare ca și 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)

Arătați că funcția densitate de probabilitate $P(x)$ din acest model nu este [LC: întotdeauna] bine definită, integrala sa pe întreg spațiul [de instanțe] nefiind în mod neapărat egală cu 1.

Sugestie: Este suficient să identificați un singur caz particular [de set de date de antrenament] pentru care să demonstrați acest fapt. Vă sugerăm să lucrați în \mathbb{R} (deci cu un singur atribut de intrare) și să observați când anume volumul V poate fi 0.

Răspuns:

a.

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

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 $p(x) = \frac{1}{|x|}$ (LC: which is not properly defined when the “new” data point x from the problem’s statement is also 0),⁴ and

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

⁴Integrala $\int_{-\infty}^{+\infty} p(x) dx$ este o *integrală improprie*, fiindcă funcția $p(x)$ nu este definită în punctul $x = 0$. (LC: Mulțumesc 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)

(Comparații între algoritmi 1-NN și ID3:
[o clasă de] seturi de date de antrenament pe care
cei doi clasificatori obțin 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 instanțele de antrenament $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ având respectiv etichetele y_1, y_2, \dots, y_n .

Întrebare: Este oare posibil ca aplicând algoritmul ID3 cu attribute numerice continue să obținem aceleași rezultate la testare / generalizare (deci și aceleași zone de decizie) ca și cele produse de clasificatorul 1-NN folosind distanța euclidiană?

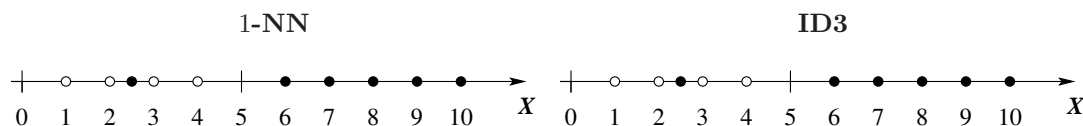
Veți aborda separat cazurile $d = 1$ și $d > 1$.

a. (0.5p, defalcăt cf. cu ceea ce urmează)

În cazul $d = 1$, pentru a înțelege mai bine cerința problemei, pentru setul de date de mai jos trasați diagrama Voronoi pe figura din stânga. Veți identifica în mod clar separatorii decizionali.

(0.3p = $2 \times 0.15p$)

Pe figura din dreapta, desenați granițele de decizie și zonele de decizie determinate de algoritmul ID3. (Vă readucem aminte convenția noastră: semnul ◦ desemnează un exemplu negativ, iar semnul • desemnează un exemplu pozitiv.)

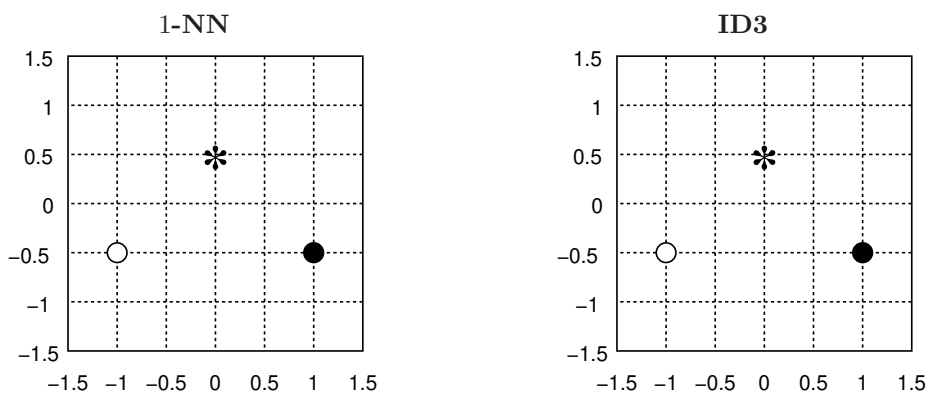


(0.2p)

Ce concluzie puteți trage referitor la calculul erorilor de tip CVLOO pentru cei doi clasificatori (1-NN și ID3 cu attribute numerice continue) pe seturi de date din \mathbb{R} ?

b. (0.75p = $0.15p \times 5$ diagrame)

În cazul $d > 1$, procedați similar pentru setul de date reprezentat mai jos, unde simbolii ◦, • și * denotă trei clase diferite. Adică, trasați diagrama Voronoi pe figura din stânga, iar pe figura din dreapta desenați granițele de decizie și zonele de decizie determinate de algoritmul ID3 (justificați riguros!).



În cazul algoritmului ID3, rezultatul este unic determinat? Dacă da, explicați de ce. Dacă nu, arătați câte variante se pot obține în total.

c. (0.25p)

Ce răspuns puteți formula acum referitor la *Întrebarea* de mai sus, din enunț? Încercați să *generalizați*, pornind de la exemplele de la punctele *a* și *b*.

Răspuns:

• CMU:

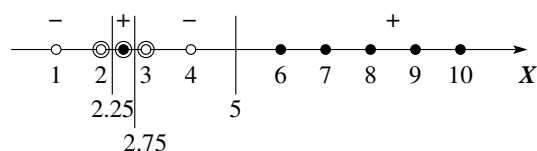
If $d = 1$, we can compute the distance between two points by using only one feature, so we can build a decision tree with the same decision boundary as 1-NN.

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 și pentru ID3:



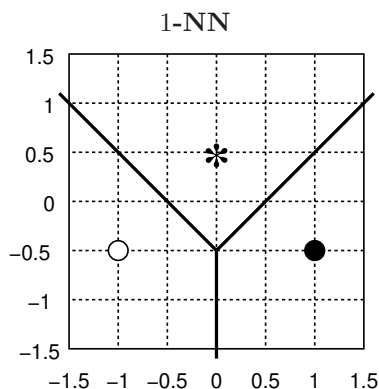
În această figură am încercuit exemplele care conduc la eroare la cross-validarea de tip “Leave-One-Out”, atât pentru algoritmul 1-NN cât și pentru algoritmul ID3.

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

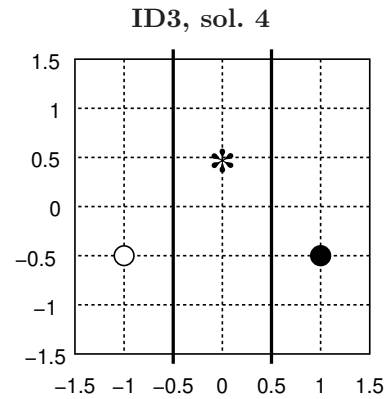
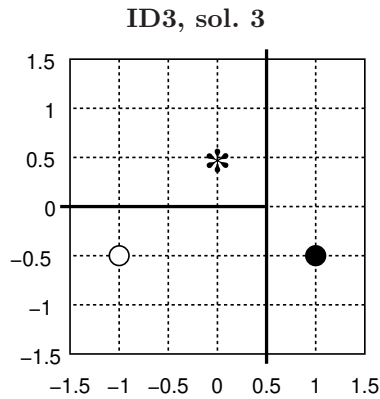
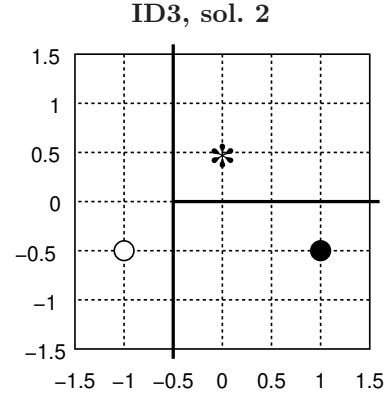
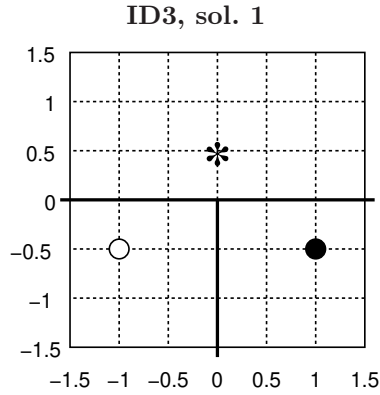
Explicația ține de folosirea distanței euclidiene la algoritmul 1-NN și respectiv modul cum se stabilesc pragurile de splitare de către algoritmul ID3 (și anume, la jumătatea distanței dintre două instanțe consecutive pe axa reală, care au etichete diferite).

În consecință, pe date de antrenament din \mathbb{R} , cei doi algoritmi produc exact aceleași zone de decizie și aceiași separatori decizionali, precum și aceleași erori la CVLOO.

b.



În cazul algoritmului ID3, split-urile sunt $x = -0.5$, $x = +0.5$, și $y = 0$, iar compașii de decizie care se formează au toți același tip de structură. (De exemplu: unul dintre ei are partiția $[1-, 1+, 1^*]$ în rădăcină, iar descendenții lui au partițiile $[1-, 0+, 0^*]$ și respectiv $[0-, 1+, 1^*]$. Ceilalți compași de decizie sunt ușor de „vizualizat“.) Asta face să se obțină mai multe soluții (a se vedea mai jos). Toate aceste soluții comportă zone de decizie diferite de cele determinate de algoritmul 1-NN (vedeți diagrama Voronoi de mai sus).



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

5. (Bonus: 1.25p)

(Probabilități condiționate, formula lui Bayes; proprietatea de aditivitate numărabilă.

O dilemă cu trei deținuți și un gardian; ipoteze MAP.)

prelucrare de Liviu Ciortuz, după

• ◦ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 3

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

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

Indicație importantă: La rezolvarea exercițiului va trebui să presupunem că, răspunzând la întrebarea deținutului a , gardianul a ales în mod aleatoriu — și uniform, dacă a fost cazul să aleagă între mai multe posibilități —, și că el a respectat însă atât adevărul cât și cerința formulată de guvernator (înțelesă într-un sens mai lax, și anume că *gardianul nu are voie să comunice niciunui deținut soarta sa, în mod direct*).

a. (0.95p, defalcat cf. cu ceea ce găsiți în rezolvare: i : 0.15p, ii : 0.5p, iii : 0.3p,)

Fie $X = a$, respectiv b sau c evenimentul care reprezintă faptul că deținutul a , respectiv b sau c a fost grațiat.

Precizați care sunt valorile numerice pentru probabilitățile a priori $P(X = a)$, $P(X = b)$ și $P(X = c)$.

Notăm cu $Y = b$ evenimentul (comunicat de gardian) că deținutul b urmează să moară (adică, să fie executat).

Calculați $P(X = a|Y = b)$. În urma obținerii de către deținutul a informației adiționale (de la gardian) că deținutul b va muri, a crescut oare probabilitatea ca el (deținutul a) să supraviețuiască?

Indicație (1): Comparați probabilitatea a posteriori $P(X = a|Y = b)$ cu probabilitatea a priori $P(X = a)$. În prealabil veți completa un tabel în care veți specifica valorile tuturor probabilităților condiționate $P(Y = y|X = x)$, cu x și $y \in \{a, b, c\}$.

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

b. (0.3p)

Presupunem că deținutul a a comunicat toate cele de mai sus deținutului c . Arătați că probabilitatea deținutului c de a supraviețui a devenit acum $2/3$.

Indicație (2): Demonstrați că $P(X = c|Y = b) = 2/3$.

Răspuns:

a.

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

(ii.) Probabilitățile a posteriori — pentru cine anume va fi grațiat, dacă știm răspunsul gardianului — sunt cele date în tabelul următor. Ele au fost calculate în funcție de datele / specificațiile din enunțul 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 $P(X = a | Y = b)$ poate fi calculată astfel:

$$\begin{aligned} P(X = a | Y = b) &\stackrel{F. Bayes}{=} \frac{P(Y = b | X = a) \cdot P(X = a)}{P(Y = b)} \\ &\stackrel{F.P.T.}{=} \frac{P(Y = b | X = a) \cdot P(X = a)}{P(Y = b | X = a) \cdot P(X = a) + P(Y = b | X = b) \cdot P(X = b) + P(Y = b | X = c) \cdot P(X = c)} \\ &\stackrel{prob. a priori}{=} \frac{P(Y = b | X = a)}{P(Y = b | X = a) + P(Y = b | X = b) + P(Y = b | X = c)} \\ &= \frac{\frac{1}{2}}{\frac{1}{2} + 0 + 1} = \frac{1}{3}. \end{aligned}$$

Așadar, răspunsul furnizat de gardian nu a condus la creșterea probabilității ca deținutul a să supraviețuiască.

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

Prin urmare, *ipoteza* ca deținutul c să fie grațiat este *mai probabilă a posteriori* decât ipoteza ca deținutul a să fie grațiat. Așadar, în acest exercițiu am identificat care este ipoteza de probabilitate maximă a posteriori (MAP) ...în același mod / spirit în care se lucrează în clasificarea bayesiană.

Observație: (indicație pentru corectarea lucrărilor)

Ca și în raționamentul precedent referitor la valoarea lui $P(X = c | Y = b)$, din **Indicația (2)** din enunț se poate deduce că

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

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