Ex: 26, 27, 31, 32, 35, 51

31. **(Aplicarea algoritmului Bayes Naiv: chestiunea valorilor lipsă (engl., missing values) în datele de antrenament)**

□ • *CMU, 2013 fall, A. Smola, G. Gordon, midterm practice, pr. 9*

Fie setul de date de antrenament (x, y) şi datele de test z:

A 1 2 3 4 5 6 7

$$x_1 = (0,0,0,1,0,0,1) \quad y_1 = 1$$
$$x_2 = (0,0,1,1,0,0,0) \quad y_2 = 1$$
$$x_3 = (1,1,0,0,0,1,0) \quad y_3 = -1$$
$$x_4 = (1,0,0,0,1,1,0) \quad y_4 = -1$$

$$z_1 = (1,0,0,0,0,1,0)$$
$$z_2 = (0,1,1,0,0,1,1)$$

Ce *problemă* va întâmpina clasificatorul Bayes Naiv pe aceste date? (*Indicaţie*: Pentru ca răspunsul dumneavoastră să fie cât mai bine justificat, veţi estima toţi parametrii necesari şi veţi aplica algoritmul pe cele două instanţe de test. Veţi nota atributele cu $A1, A2, \ldots$.)

La curs am prezentat un „remediu" standard pentru o astfel de *problemă*. Precizaţi cum se numeşte „tehnica" respectivă şi aplicaţi-o pe aceste date. După aceea, veţi aplica algoritmul Bayes Naiv pentru a clasifica instanţele de test z_1 şi z_2.

$$y_{MAP} = \underset{y \in \{0,1\}}{\arg\max}\, P(v_j | a_i) \overset{F.B}{=} \underset{y \in \{0,1\}}{\arg\max} \frac{P(a_i | v_j) \cdot P(v_j)}{P(a_i)} =$$

$$= \underset{y \in \{0,1\}}{\arg\max} \prod P(a_i | v_j) \cdot P(v_j)$$

P.P. ind. cond.

$$z_1 = (1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0) \qquad P_{y=1} = P(A_1 = 1 | y = 1) \cdot P(A_2 = 0 | Y = 1) \ldots P(y=1)$$

$$z_1: \quad P_{y=1} = \frac{0}{2} \cdot \frac{2}{2} \cdot \frac{1}{2} \cdot \frac{0}{2} \cdot \frac{2}{2} \cdot \frac{0}{2} \cdot \frac{1}{2} \cdot \frac{2}{4} = 0$$

$$z_1: \quad P_{y=-1} = \frac{2}{2} \cdot \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{2}{2} \cdot \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{2}{2} \cdot \frac{2}{4} = \frac{1}{8}$$

$$z_2 = (0, 1, 1, 0, 0, 1, 1)$$

$$P_{z_2 = 1} = \frac{2}{2} \cdot \frac{0}{2} \cdot \frac{1}{2} \cdot \frac{0}{2} \cdot \frac{2}{2} \cdot \frac{0}{2} \cdot \frac{1}{2} = 0$$

$$P_{z_2 = -1} = \frac{0}{2} \cdot \ldots = 0$$

Observăm că avem probabilităţi care se înmulţesc cu 0

Folosim regula „add-one" a lui Laplace

pentru a „netezi" probabilităţile.

$$z_1 = (1\ 0\ 0\ 0\ 0\ 1\ 0)$$

$$P_{1y=1} = \frac{0+1}{2+2} \cdot \frac{0+1}{4} \cdot \frac{1+1}{4} \cdot \frac{0+1}{4} \cdot \frac{2+1}{4} \cdot \frac{0+1}{4} \cdot \frac{1+1}{4} =$$

$$= \frac{1}{4} \cdot \frac{1}{4} + \cancel{\frac{2}{4}}_{2} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{3}{6} = \frac{1}{16} \cdot \frac{1}{8} \cdot \frac{3}{16} \cdot \frac{1}{8} = 0.001$$

$$P_{1y=-1} = \frac{2+1}{2+2} \cdot \frac{2}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{3}{5} \cdot \frac{3}{4} \cdot \frac{3}{6} = 0.029$$

Prin urmare $P_{y=-1} > P_{y=1}$ $\Rightarrow$ BN va clasifica z_1 în clasa $y=-1$

$$z_2 = (0, 1, 1, 0, 0, 1, 1)$$

$$P_{y=1} = \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{3}{6} = 0.001$$

$$P_{y=-1} = \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{3}{6} = 0.001$$

$P_{y=1} = P_{y=-1}$, Deci BN va prezice z_2 ca fiind în $y=1$ sau $y=-1$ cu probabilitatea $\frac{1}{2}$.

26. **(Adevărat sau Fals?)**

CMU, 2002 fall, Andrew Moore, final exam, pr. 11.a

Fie D un set de exemple (date de antrenament), iar H o mulţime de ipoteze pentru (un algoritm de) învăţare automată pe datele D. Precizaţi care este *valoarea de adevăr* a următoarelor afirmaţii:

$\text{argmax}_{h \in H} P(D|h)$ este ipoteză de probabilitate maximă a posteriori, [iar] $\text{argmax}_{h \in H} P(h|D)$ este ipoteză de verosimilitate maximă.

Fals

$v_{MAP} = \underset{h \in H}{\text{argmax}}\, P(D|h)$ → fals în. corectă este:

$$v_{MAP} = \underset{h \in H}{\text{argmax}}\, P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

La MAP țin cont de date din trecut (inițiale) înainte de a „observa" noi date.

La ipoteza ML nu țin cont de nimic, observ datele direct și după trag concluziile în funcție de datele observate. ⇒ $P(h|D)$

Deci pentru $v_{MLE} = \underset{h \in H}{\text{argmax}}\, P(D|h)$

Deci în MLE căutăm acea ipoteză $h \in H$ care-mi maximizează probabilitățile datelor D observate fără a ține cont de alte date inițiale.

Și la MAP căutăm

la fel să maximizăm $P(D|h)$ ținând cont și de prob. datelor inițiale $\underbrace{P(h)}_{\text{„priori"}}$ ⇒ $\underset{h \in H}{\text{argmax}}\, P(D|h) \cdot P(h)$

32. **(Algoritmul Bayes Naiv: calculul ratei medii a erorii – exemplificare)**

• ○ *CMU, 2011 spring, Tom Mitchell, midterm, pr. 5.1-2*

Considerăm o problemă de clasificare binară în care se foloseşte o variabilă $X_1 \in \{0,1\}$ şi eticheta $Y \in \{0,1\}$. Distribuţia generativă „adevărată“ $P(X_1, Y) = P(Y) \cdot P(X_1|Y)$ este determinată conform tabelelor următoare:

Y	0	1
$P(Y)$	0.8	0.2

$P(X_1\|Y)$	$Y=0$	$Y=1$
$X_1=0$	0.7	0.3
$X_1=1$	0.3	0.7

a. Presupunem că am antrenat un clasificator Bayes Naiv folosind o infinitate de *date de antrenament* generate conform celor două tabele de mai sus. Scrieţi în tabelul următor *predicţiile* făcute de algoritmul Bayes Naiv pentru diferitele valori ale lui X_1. Remarcaţi faptul că $\hat{Y}(X_1)$ din acest tabel este decizia [lui Bayes Naiv] cu privire la valoarea lui Y dat fiind X_1. În coloanele care corespund probabilităţilor, veţi scrie atât valorile concrete ale acestor probabilităţi, cât şi modul cum au fost ele calculate (de exemplu, $0.8 \cdot 0.7 = 0.56$), iar în coloana care corespunde deciziei, veţi scrie [pe fiecare linie] fie $\hat{Y}=0$ fie $\hat{Y}=1$.

	$P(X_1, Y=0)$	$P(X_1, Y=1)$	$\hat{Y}(X_1)$
$X_1=0$	0.56	0.06	0
$X_1=1$	0.24	0.14	0

$$P(X_1, Y=0) \overset{R.C}{=} P(X_1=0 \mid Y=0)\cdot P(Y=0) = 0.7 \cdot 0.8 = 0.56$$
$$= P(X_1=1 \mid Y=0)\cdot P(Y=0) = 0.3 \cdot 0.8 = 0.24$$

$$P(X_1, Y=1) \overset{R.C}{=} P(X_1=0 \mid Y=1)\cdot P(Y=1) = 0.3 \cdot 0.2 = 0.06$$
$$= P(X_1=1 \mid Y=1)\cdot P(Y=1) = 0.7 \cdot 0.2 = 0.14$$

$$\hat{y}(X_1) \overset{MAP}{=} \underset{y\in\{0,1\}}{\operatorname{argmax}}\ P(X_1 \mid Y=y)\cdot P(Y=y)$$

- $\hat{y}(X_1=0)_{,y=0} = P(X_1=0 \mid Y=0)\cdot P(Y=0) = P(X_1=0, Y=0) = 0.56$ $\vee$ $\Rightarrow$
- $\hat{y}(X_1=0)_{,y=1} = P(X_1=0 \mid y=1)\cdot P(Y=1) = P(X_1=0, y=1) = 0.06$

$$\hat{y}(X_1) = 0$$

altfel, comparăm direct pe tabel

	$P(X_1, Y=0)$	$P(X_1, Y=1)$	$\hat{Y}(X_1)$
$X_1 = 0$	0.56	0.06	0
$X_1 = 1$	0.24	0.14	0

instanțe test

b. Cât este *rata medie a erorilor* (engl., expected error rate) produsă de acest clasificator Bayes Naiv pe *instanțele de test* care sunt generate conform primelor două tabele de mai sus? Cu alte cuvinte, calculați $P(\hat{Y}(X_1) \neq Y)$ unde perechile (X_1, Y) sunt generate conform celor două tabele.

Indicație: $P(\hat{Y}(X_1) \neq Y) = P(\hat{Y}(X_1) \neq Y, X_1 = 0) + P(\hat{Y}(X_1) \neq Y, X_1 = 1)$.

? $y = 1 - \hat{y}(X_1)$

Pentru următoarele trei puncte ale acestui exercițiu vom considera două variabile, și anume $X_1 \in \{0,1\}$ și $X_2 \in \{0,1\}$, precum și eticheta $Y \in \{0,1\}$. Y și X_1 sunt și de data aceasta generate conform primelor două tabele de mai sus, iar apoi X_2 este creat ca o *copie exactă* (adică, duplicat) după X_1.

$$P(\hat{y}(X_1) \neq Y) = \sum_{X_1=0}^{1} P(X_1, Y \neq \hat{y}(X_1)) = P(X_1=0, y = 1 - \hat{y}(X_1=0))$$

$$+ P(X_1=1, Y = 1 - \hat{y}(X_1=1)) = P(X_1=0, Y=1-1) + P(X_1=1, y=1-1) =$$

$$= 0.06 + 0.14 = 0.2$$

c) **c.** Acum vom presupune că am antrenat un clasificator Bayes Naiv folosind o infinitate de *exemple de antrenament* care au fost generate în conformitate cu primele două tabele de mai sus *și* cu regula de duplicare. Scrieți în tabelul următor *predicțiile* făcute de către acest clasificator Bayes Naiv pentru diferitele valori ale perechii (X_1, X_2). În privința probabilităților din tabel, puteți să scrieți doar cum anume sunt ele calculate (de exemplu, în loc de $0.8 \cdot 0.3 \cdot 0.3 = 0.072$ veți putea scrie doar $0.8 \cdot 0.3 \cdot 0.3$ pentru a economisi un pic de timp).

	$\hat{P}(X_1, X_2, Y=0)$	$\hat{P}(X_1, X_2, Y=1)$	$\hat{Y}(X_1, X_2)$
$X_1 = 0, X_2 = 0$	0.392 >	0.018	0
$X_1 = 0, X_2 = 1$	0.168	0.042	0
$X_1 = 1, X_2 = 0$	0.168	0.042	0
$X_1 = 1, X_2 = 1$	0.072	0.098	1

1

$$P(X_1, X_2, Y=0) = P(X_1=0, X_2=0 \mid Y=0) \cdot P(Y=0) \overset{\text{P.P. ind. cond.}}{=}$$

$$= P(X_1=0 \mid Y=0) \cdot P(X_2=0 \mid Y=0) \cdot P(Y=0) =$$

$$= 0.7 \cdot 0.7 \cdot 0.8 = 0.392$$

$$P(X_1=0, X_2=1 \mid Y=0) \cdot P(Y=0) =$$

$= P(X_1=0 \mid Y=0) \cdot P(X_2=1 \mid Y=0) \cdot P(Y=0) = 0.7 \cdot 0.3 \cdot 0.8 =$

$= 0.168$

- $P(X_1=1, X_2=0 \mid Y=0) \cdot P(Y=0) = \quad = 0.168$

- $P(X_1=1, X_2=1 \mid Y=0) \cdot P(Y=0) = P(X_1=1 \mid Y=0) \cdot P(X_2=1 \mid Y=0) \cdot P(Y=0)$

$= 0.072$

$\hat{P}(X_1, X_2, Y=1) = P(X_1=0, X_2=0 \mid Y=1) \cdot P(Y=1) =$ P.P. ind. cond.

$= P(X_1=0 \mid Y=1) \cdot P(X_2=0 \mid Y=1) \cdot P(Y=1)$

$= 0.018$

$P(X_1=0, X_2=1 \mid Y=1) \cdot P(Y=1) =$

$= P(X_1=0 \mid Y=1) \cdot P(X_2=1 \mid Y=1) \cdot P(Y=1) = 0.042$

- $P(X_1=1, X_2=0 \mid Y=1) \cdot P(Y=1) = \quad = 0.042$

- $P(X_1=1, X_2=1 \mid Y=1) \cdot P(Y=1) = P(X_1=1 \mid Y=1) \cdot P(X_2=1 \mid Y=1) P(Y=1)$

$= 0.098$

---

d) **d. Cât este *rata medie a erorilor* pentru acest clasificator Bayes Naiv pe instanţe de test care sunt generate în conformitate cu primele două tabele de mai sus *şi* cu regula de duplicare?**

$\sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(y \neq \hat{y}(x_1, x_2)) = P(X_1=0, X_2=0, y=1-0) +$

$+ P(X_1=0, X_2=1, Y=1-0) + P(X_1=1, X_2=0, y=1-0) +$

$+P(X_1=1, X_2=1, Y=1-1) = 0.018 + 0.042 + 0.042 +$

$+0.072 = 0.09 + 0.084 = 0.174$

e. **Comparativ cu cazul precedent (adică, fără X_2), cum s-a schimbat rata medie a erorilor (adică, a crescut ori a scăzut)? În tabelul de la punctul** c, ***ce*** **linie este responsabilă pentru această schimbare? Cum explicaţi ce s-a întâmplat?**

Observăm că rata medie a erorii a scăzut pentru datele duplicate.

Ultima linie face schimbarea, deoarece dacă ne uităm în tabelul de la a) obs. că pentru $X_1 = 1$ NB prezice doar clasa 0. Însă odată ce am adăugat și X_2, $X_1 = 1$ este împărțit în 2 clase, atât 0 cât și 1, astfel având și X_2, alg. NB va clasifica „puțin" mai bine datele care conțin atrib. $X_1 = 1$, având o eroare medie mică. $0.174 < 0.2$. În acest caz ne ajută duplicarea unui atribut, deoarece nu am și alte atribute diferite, dacă aș fi avut, duplicarea ar fi putut să facă să conteze mai puțin datele neduplicate și mai mult datele duplicate, distorsionând distribuția reală a datelor, lucru care duce în general la o rată medie a erorii mai mare.

41. **(Comparaţie între clasificatorul Bayes Naiv şi algoritmul ID3)**

□ • ∘ *CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 5.b*

Care dintre afirmaţiile de mai jos sunt adevărate atât pentru clasificatorul Bayes Naiv cât şi pentru algoritmul ID3 pentru învăţarea de arbori de decizie? (Veţi putea alege nu neapărat una singură dintre aceste afirmaţii.)

1. În cazul ambilor clasificatori se presupune că orice pereche de atribute X_i şi X_j cu $i \neq j$ — văzute ca variabile aleatoare — sunt independente.
2. În cazul ambilor clasificatori se presupune că orice pereche de atribute X_i şi X_j cu $i \neq j$ sunt dependente.
3. În cazul ambilor clasificatori se presupune că orice pereche de atribute sunt independente în raport cu eticheta (adică variabila care reprezintă clasa).
4. În cazul ambilor clasificatori se presupune că orice pereche de atribute sunt dependente în raport cu eticheta.

1. Fals, doar la NB sunt considerate independente.
ID3 caută atr. cu câștig de inf. maxim pentru a crea arborele de decizie.

2. Fals. (opusul afirmației 1)

3. Fals la BN, deși el presupune indep. pt toate atributele, acestea sunt condiționate de etichetă.
Fals la ID3, calculul câștigului de inf. al tuturor atr. este în funcție de etichetă, sugerând o relație între atribut și eticheta sa.

4. Adevărat, atributele alg NB sunt independente condiționat de etichetă, rezultând dependența între atribute și etichetă.

34. **(Algoritmul Bayes Naiv: calculul ratei medii a erorii – exemplificare; comparaţie cu regresia logistică)**

prelucrare de Liviu Ciortuz, după
• * *CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.4*

Considerăm o problemă de clasificare binară în care fiecare exemplu de antrenament are două atribute binare $X_1, X_2 \in \{T, F\}$ şi eticheta / clasa $Y \in \{T, F\}$.

Presupunem că $P(Y = T) = 0.5$, iar $P(X_1 = T|Y = T) = 0.8$, $P(X_1 = F|Y = F) = 0.7$, $P(X_2 = T|Y = T) = 0.5$ şi $P(X_2 = F|Y = F) = 0.9$. **(Se poate observa că atributul X_1 furnizează / constituie un indiciu întrucâtva mai puternic decât atributul X_2 în ce priveşte determinarea clasei unei instanţe oarecare.)**

În cele ce urmează vom presupune că X_1 şi X_2 sunt independente în raport cu Y.

a. Calculaţi probabilităţile $P(X_1 = F|Y = T)$, $P(X_1 = T|Y = F)$, $P(X_2 = F|Y = T)$ şi $P(X_2 = T|Y = F)$. Asociaţi răspunsului dumneavoastră o justificare generală, sub forma unei formule din teoria probabilităţilor:

$P(\neg A|B) = \ldots$, **unde A şi B sunt evenimente aleatoare oarecare.**

$P(Y=F) = 1 - P(Y=T)$
$= 1 - 0.5 = 0.5$

Y	T	F
$P(Y)$	0.5	0.5

	$P(X_1 \mid Y=T)$	$P(X_1 \mid Y=F)$
$X_1 = T$	0.8	0.3
$X_1 = F$	0.2	0.7

	$P(X_2 \mid Y=T)$	$P(X_2 \mid Y=F)$
$X_2 = T$	0.5	0.1
$X_2 = F$	0.5	0.9

Restul abr. le-am obț. prin complementaritate.

a) $X_1, X_2 \to$ indep. cond

$P(Y=F) = 1 - P(Y=T)$

Folosim prob. Totală $P(A) = P(A,B) + P(A,\bar{B}) = 1$

$1 = P(X_1=T \mid Y=T) + P(X_1=F \mid Y=T) \Rightarrow P(X_1=F \mid Y=T) =$

$= 1 - 0.8 = 0.2$

⋮

la fel și la celelalte.

b)

b. Scrieţi regula de decizie a algoritmului Bayes Naiv pentru $X_1 = x_1$ şi $X_2 = x_2$, justificând în mod succint obţinerea ei.

$$\hat{y}(X_1, X_2) = \underset{y \in \{T,F\}}{\operatorname{argmax}} P(Y=y \mid X_1 = x_1, X_2 = x_2) =$$

$$= \underset{y \in \{T,F\}}{\operatorname{argmax}} P(X_1 = x_1, X_2 = x_2 \mid Y=y)\, P(Y=y) =$$

D.P. indep. cond.

$$= \underset{y \in \{T,F\}}{\operatorname{argmax}} P(X_1 = x_1 \mid Y=y) \cdot P(X_2 = x_2 \mid Y=y) \cdot P(Y=y)$$

P_1, $Y=T$: $\operatorname{argmax} P(X_1 = x_1 \mid Y=T) \cdot P(X_2 = x_2 \mid Y=T) \cdot P(Y=T)$

P_2, $Y=F$: $\operatorname{argmax} P(X_1 = x_1 \mid Y=F) \cdot P(X_2 = x_2 \mid Y=F) \cdot P(Y=F)$

În fct. de param. x_1 și x_2, se vor calcula P_1 și P_2 și alg va alege maximul dintre acestea

c. Calculaţi rata medie a erorii produse de algoritmul Bayes Naiv, atunci când se folosesc ambele atribute, X_1 şi X_2. (Veţi da în prealabil definiţia ratei medii a erorii.) Este oare această rată mai bună decât în cazul în care se foloseşte un singur atribut (X_1 sau X_2)? De ce?

Def: $E_P\left[\mathbb{1}\{y_{NB}(x_1,x_2)\neq y\}\right]$

Rata medie a erorii ne spune cat de prost clasifică alg. datele.

$$= \sum_{X_1,X_2,Y} 1_{\{\hat{y}(X_1,X_2)\neq Y\}} \cdot P(X_1,X_2,Y)$$

X_1	X_2	Y	$P(X_1,X_2,Y) = P(X_1\mid Y)\cdot P(X_2\mid Y)\cdot P(Y)$	$y_{NB}(X_1,X_2)$
T	T	T	$0.8\cdot0.5\cdot0.5=0.2$	T
T	T	F	$0.3\cdot0.1\cdot0.5=0.015$	T
T	F	T	$0.8\cdot0.5\cdot0.5=0.2$	T
T	F	F	$0.3\cdot0.9\cdot0.5=0.135$	T
F	T	T	$0.2\cdot0.5\cdot0.5=0.05$	T
F	T	F	$0.7\cdot0.1\cdot0.5=0.035$	T
F	F	T	$0.2\cdot0.5\cdot0.5=0.05$	F
F	F	F	$0.7\cdot0.9\cdot0.5=0.315$	F

Calc: $\sum_{X_1=T}^{1}\sum_{X_2=T}^{1} P(\hat{y}(X_1,X_2)\neq y) = P(X_1=T,X_2=T,y=1-\hat{y}(x_1,x_2))$

$+ P(X_1=F,X_2=T,y=1-\hat{y}(x_1,x_2)) + P(X_1=T,X_2=F,y=1-\hat{y}(x_1,x_2)) +$

$+ P(X_1=F,X_2=F,y=1-\hat{y}(x_1,x_2)) = 0.015+0.135+0.035+0.05=$

$= 0.185 + 0.09 = 0.235$

Da, această rată este mai bună, deoarece avem mai multe date, iar acest aspect face ca alg NB să clasifice mai bine.

d)

d. Să presupunem acum că se crează un nou atribut, X_3, care este o copie exactă a lui X_2. Aşadar, pentru fiecare exemplu de antrenament, atributele X_2 şi X_3 au aceeaşi valoare, $X_2 = X_3$. Răspundeţi la următoarele întrebări:

– Sunt X_2 şi X_3 independente condiţional în raport cu Y?

– Cât este rata medie a erorii pentru Bayes Naiv acum? (Atenţie! Distribuţia „adevărată“ a datelor nu s-a modificat.)

– Explicaţi ce se întâmplă cu algoritmul Bayes Naiv. Oare *regresia logistică* are aceeaşi problemă? Explicaţi de ce.

Dacă X_3 este copia lui X_2, atunci atributele în mod clar sunt dependente între ele, X_3 fiind la fel ca X_2 depinde de valorile sale, știind valori din X_2 vom ști și cele din X_3 ceea ce nu reprezintă o independență.

Dacă mai adăugăm date noi, rata medie a erorii probabil va crește, deoarece X_3 fiind copie a lui X_2 alg va calc. prob. ca și cum ele ar fi independente. Astfel alg. va ține în mod indirect cont mai mult de prob. lui X_2 decât X_1, ceea ce va rezulta cel mai probabil într-o clasificare greșită a atr., ceea ce duce la creșterea ratei medii a erorii.

În Regresia logistică nu există aceeași problemă, deoarece ponderile atr. duplicate se ajustează a.î. modelul va evita supraestimarea importanței atributelor duplicate.

	$P(Y)$
$Y=T$	0.5
$Y=F$	0.5

	$P(X_1 \mid Y=T)$
$X_1=T$	0.8
$X_1=F$	0.2

	$P(X_2 \mid Y=T)$
$X_2=T$	0.5
$X_2=F$	0.5

	$P(X_1 \mid Y=F)$
$X_1=T$	0.3
$X_1=F$	0.7

	$P(X_2 \mid Y=F)$
$X_2=T$	0.1
$X_2=F$	0.9

X_3	X_1	X_2	Y	$P(X_1,X_2,X_3,Y)$	$P_{NB}(X_1,X_2,X_3,Y) = P(X_1\|Y)\cdot P(X_2\|Y)\cdot P(X_3\|Y)\cdot P(Y)$	$y_{NB}(X_1,X_2,\dots$
T	T	T	T	0	$0.8 \cdot 0.5 \cdot 0.5 = 0.2 \cdot 0.5 = 0.1$	T
T	T	T	F	0	$0.3 \cdot 0.1 \cdot 0.5 = 0.015 \cdot 0.1 = 0.0015$	T
T	T	F	T	0.2	$0.8 \cdot 0.5 \quad 0.5 = 0.2 \cdot 0.5 = 0.1$	T
T	T	F	F	0.135	$0.3 \cdot 0.9 \quad 0.5 = 0.135 \cdot 0.1 = 0.0135$	T
T	F	T	T	0	$0.2 \cdot 0.5 \quad 0.5 = 0.05 \cdot 0.5 = 0.025$	T
T	F	T	F	0	$0.7 \cdot 0.1 \quad 0.5 = 0.035 \cdot 0.1 = 0.0035$	T
T	F	F	T	0.05	$0.2 \cdot 0.5 \quad 0.5 = 0.05 \cdot 0.5 = 0.025$	F
T	F	F	F	0.315	$0.7 \cdot 0.9 \quad 0.5 = 0.315 \cdot 0.1 = 0.0315$	F
F	T	T	T	0.2	$0.8 \cdot 0.5 \cdot 0.5 = 0.2 \cdot 0.5 = 0.1$	T
F	T	T	F	0.015	$0.3 \cdot 0.1 \cdot 0.5 = 0.015 \cdot 0.9 = 0.0135$	T
F	T	F	T	0	$0.8 \cdot 0.5 \quad 0.5 = 0.2 \cdot 0.5 = 0.1$	F
F	T	F	F	0	$0.3 \cdot 0.9 \quad 0.5 = 0.135 \cdot 0.9 = 0.315$	F
F	F	T	T	0.05	$0.2 \cdot 0.5 \quad 0.5 = 0.05 \cdot 0.5 = 0.025$	F
F	F	T	F	0.035	$0.7 \cdot 0.1 \quad 0.5 = 0.035 \cdot 0.9 = 0.0315$	F
F	F	F	T	0	$0.2 \cdot 0.5 \quad 0.5 = 0.05 \cdot 0.5 = 0.025$	F
F	F	F	F	0	$0.7 \cdot 0.9 \quad 0.5 = 0.315 \cdot 0.9 = 0.2835$	F

sursa erorii

[sunt dependente, sunt la fel, cale este invalid] pun 0 ?

distribuția calc. de NB care își asumă atr. ca fiind indep. (în cazul nostru X_2 și X_3 sunt dependente, deci distribuția este greșită)

$$= E_P\left[1_{\{y_{NB}(X_1,X_2,X_3) \neq y\}}\right]$$

$$= \sum_{X_1,X_2,X_3,Y} 1_{\{y(X_1,X_2,X_3) \neq y\}} \cdot P(X_1,X_2,X_3,Y)$$

Așa ar fi calculul dacă X_3 ar fi fost indep. de X_2.

$$= P(X_1=T, X_2=T, X_3=T, Y=1-y(x_1,x_2,x_3)) + P(X_1=T, X_2=F, X_3=T, Y=1-y(x_1,x_2,x_3)) +$$
$$P(X_1=F, X_2=T, X_3=T, Y=1-y(x_1,x_2,x_3)) + P(X_1=F, X_2=F, X_3=T, Y=1-y(x_1,x_2,x_3)) +$$
$$P(X_1=T, X_2=T, X_3=F, Y=1-y(x_1,x_2,x_3)) + P(X_1=T, X_2=F, X_3=F, Y=1-y(x_1,x_2,x_3)) +$$
$$P(X_1=F, X_2=T, X_3=F, Y=1-y(x_1,x_2,x_3)) + P(X_1=F, X_2=F, X_3=F, Y=1-y(x_1,x_2,x_3)) =$$

$$= 0.0015 + 0.0135 + 0.0035 + 0.025 + 0.0135 + 0.1 + 0.025 + 0.025 =$$

$= 0.005 + 0.027 + 0.05 + 0.125 = 0.13 + 0.077 =$

$= 0.207$

Rata erorii se face pe dist. reală, P, nu pe P_{NB}. $\Rightarrow$

$$E_P[1_{\{\hat{y}(x_1,x_2,x_3)\neq Y\}}] = 0.05 + 0.015 + 0.05 + 0.135 = 0.115 + 0.135 =$$
$$= 0.250$$

Se observă cum rata este mai mare

0.250 > 0.235, de unde reiese că duplicarea atributelor pentru clasificarea NB duce la o rată a erorii mai mare.

27. **(Algoritmii Bayes Naiv şi Bayes Optimal; aplicare; numărul minimal de parametri de estimat)**

• ○ ***Liviu Ciortuz, 2017, pornind de la setul de date din Machine Learning, Tom Mitchell, 1997, ch. Decision Trees, page 59***

Considerăm următorul set de date de antrenament, în care variabila de ieşire este *EnjoyTennis*:

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

a. Determinaţi decizia luată de către algoritmul Bayes Naiv pentru instanţa de test

$$X = \langle Outlook = sunny,\ Temp = cool,\ Humidity = high,\ Wind = strong\rangle,$$

precum şi probabilitatea cu care este luată această decizie.

b. Care este numărul *minim* de parametri pe care trebuie să-l estimeze algoritmul Bayes Naiv pe aceste date [pentru a face apoi predicţii pe un set oarecare de instanţe de test]? Dar în cazul clasificatorului Bayes Optimal?

c. Implementaţi algoritmul Bayes Naiv, iar apoi cu ajutorul acestei implementări calculaţi eroarea la antrenare şi eroarea la CVLOO pe acest set de date.

a)

$$\hat{y}(x) = \arg\max_{e\in\{No,Yes\}} P(E=e \mid O=s, T=c, H=h, W=s) =$$

$$= \arg\max \frac{P(O=s, T=c, H=h, W=s \mid E=e)\cdot P(E=e)}{P(O=s, T=c, H=h, W=s)} =$$

$$= \arg\max P(O=s, T=c, H=h, W=s \mid E=e)\cdot P(E=e) = NB$$

P_1: E = No: $P(O=s, T=c, H=h, W=s \mid E=n) \cdot P(E=n) =$

n.p. ind. cond

$P(O=s \mid E=n) \cdot P(T=c \mid E=n) \cdot P(H=h \mid E=n) \cdot$

MLE

$\cdot P(W=s \mid E=n) \cdot P(E=n) =$

$= \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{\cancel{5}_1} \cdot \frac{\cancel{5}}{14} = 0.0411$ ⌊Observăm în tabel, MLE,⌉

P_2: E = yes: $P(O=s \mid E=y) \cdot P(T=c \mid E=y) \cdot P(H=h \mid E=y) \cdot P(W=s \mid E=y) \cdot P(E=y) \overset{MLE}{=} \frac{2}{3\cancel{9}} \cdot \frac{\cancel{3}}{\cancel{9}} \cdot \frac{\cancel{3}}{9} \cdot \frac{3}{\cancel{9}_1} \cdot \frac{\cancel{9}^1}{14} = 0.0052$

Deci, $P_1 > P_2$, alg. NB va prezice pt. $X = (..)$ Enjoy Tennis = no.

b) Nr. minim de param. de estimat pentru NB? Dar pentru JB?

Pentru NB vom avea un nr. liniar de param de estimat. $2d+1$.

Pentru JB avem un nr exponențial

$2^{d+1} - 1$.

$P(E=yes) \Rightarrow P(E=n) = 1 - P(E=yes)$

$P(W=s \mid E=y) \Rightarrow P(W=w \mid E=y) = 1 - P(W=s \mid E=y)$

$P(W=s \mid E=n) \Rightarrow P(W=w \mid E=n) = 1 - P(W=s \mid E=n)$

$P(H=n \mid E=y) \Rightarrow P(H=h \mid E=y) = 1 - P(H=n \mid E=y)$

$P(H=n \mid E=n) \Rightarrow P(H=h \mid E=n) = 1 - P(H=n \mid E=n)$

$\left.\begin{array}{l} P(T=c \mid E=y) \\ P(T=m \mid E=y) \end{array}\right\} \Rightarrow P(T=h \mid E=y) = 1 - P(T=c \mid E=y) - P(T=m \mid E=y)$

$$\left.\begin{array}{l} P(T=c|E=n) \\ P(T=m|E=n) \end{array}\right\} \Rightarrow P(T=h|E=n) = 1 - P(T=c|E=n) - P(T=m|E=n)$$

$$\left.\begin{array}{l} P(O=s|E=y) \\ P(O=r|E=y) \end{array}\right\} \Rightarrow P(O=o|E=y) = 1 - P(O=r|E=y) - P(O=r|E=y)$$

$$\left.\begin{array}{l} P(O=s|E=n) \\ P(O=r|E=n) \end{array}\right\} \Rightarrow P(O=o|E=n) = 1 - P(O=r|E=n) - P(O=r|E=n)$$

$\rightarrow 2d(m-1)+1$ → nr. d valori ale atr. d.

(dacă outputul este binar)

Pentru NB vom avea nevoie minim de 13 param. pentru a putea estima.

Iar pentru JB avem param:

pentru datele de ieșire care nu sunt binare $\Rightarrow$

$$2\prod_{i=1}^{d} K_i - 1,$$

$\uparrow$ y, binar

d – nr. de atribute

$K_i \rightarrow$ nr. de valori pe care le poate avea fiecare atribut d.

$$2\cdot(3\cdot3)\cdot2\cdot2 - 1 = 2\cdot 9\cdot 4 - 1 = 72 - 1 = 71.$$

Pentru JB vom avea nevoie minim de 71 param. pentru a putea estima.

sau JB:

$$P(O=\ldots,\ T=\ldots,\ H=\ldots,\ W=\ldots,\ y=y)$$

$$3\cdot 3\cdot 2\cdot 2\cdot 2 - 1$$

$$= 71.$$

c) Am făcut la temă opționala doar că pe alt set de date.
(MNIST)