Inteligenţă artificială
Laboratorul 11

Clasificatorul bayesian naiv

1. Modelul teoretic

Metoda de clasificare bayesiană naivă (engl. "Naïve Bayes") se bazează pe calcularea probabilităţilor ca o anumită instanţă să aparţină claselor problemei.

Într-o reţea bayesiană, se evita calcularea întregii distribuţii comune de probabilitate după regula de înmulţire a probabilităţilor (engl. "chain rule") presupunând că un nod depinde doar de părinţii săi din graf. În metoda naivă, presupunerea simplificatoare este şi mai puternică, considerând că toate atributele sunt independente dată fiind clasa. Acest fapt nu este neapărat adevărat, de cele mai multe ori, dimpotrivă, condiţia de independenţă poate să nu fie satisfăcută. Cu toate acestea, s-a constatat că deseori metoda are rezultate foarte bune.

Formal, se consideră fiecare atribut şi clasa ca variabile aleatorii. Se dă o instanţă definită de valorile atributelor $(A_1, \dots, A_n)$. Scopul este determinarea clasei C pentru această combinaţie de valori, ceea ce este echivalent cu găsirea valorii C_j care *maximizează* probabilitatea clasei dată fiind instanţa:

$$C^* = \underset{C_j}{argmax}\, P\left(C_j | A_1, \dots, A_n\right) \tag{1}$$

Această probabilitate trebuie estimată direct din datele mulţimii de antrenare, pe baza frecvenţelor relative de apariţie a valorilor atributelor.

Conform teoremei lui Bayes:

$$P\left(C_j | A_1, \dots, A_n\right) = \frac{P\left(A_1, \dots, A_n | C_j\right) \cdot P\left(C_j\right)}{P(A_1, \dots, A_n)}. \tag{2}$$

$P(A_1, \dots, A_n)$ este aceeaşi pentru toate valorile clasei întrucât este vorba despre aceeaşi instanţă pentru toate clasele. Ea depinde doar de valorile atributelor instanţei şi nu de clase, astfel încât o putem ignora atunci când vrem să maximizăm cantitatea din partea dreaptă a ecuaţiei (2). Problema de clasificare devine echivalentă cu alegerea valorii clasei care maximizează numărătorul:

$$C^* = \underset{C_j}{\operatorname{argmax}}\, P\left(A_1, \dots, A_n | C_j\right) \cdot P\left(C_j\right). \tag{3}$$

Rămâne de estimat probabilitatea instanţei dată fiind clasa. Considerând că toate atributele sunt independente dată fiind clasa (presupunerea fundamentală a metodei bayesiene naive), putem exprima acest produs sub forma:

$$P(A_1, \dots, A_n|C_j) = P(A_1|C_j) \cdot \dots \cdot P(A_n|C_j). \quad (4)$$

După cum vom vedea în continuare, putem estima uşor din datele mulţimii de antrenare $P(A_i|C_j)$ pentru toate valorile atributelor A_i şi clasei C_j. Problema de clasificare devine următoarea:

$$C^* = \underset{C_j}{\operatorname{argmax}} P(C_j) \cdot \prod_{i=1}^{n} P(A_i|C_j). \quad (5)$$

Metoda se reduce la găsirea clasei pentru care valoarea produsului este maximă.

2. Probleme cu atribute simbolice

Pentru a exemplifica modalitatea de calcul, vom considera o problemă privind oportunitatea de a juca golf sau nu (tabelul 1).

Tabelul 1. *Problemă de clasificare cu atribute simbolice*

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
7	Înnorat	Mică	Normală	Prezent	Da
8	Soare	Medie	Mare	Absent	Nu
9	Soare	Mică	Normală	Absent	Da
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

Metoda bayesiană naivă clasifică o instanţă dată, care poate să fie una nouă, care nu aparţine mulţimii de antrenare.

În cazul de faţă, fie această instanţă: x_q = (*Soare*, *Mare*, *Normală*, *Absent*).

Trebuie să realizăm un număr de calcule egal cu numărul de clase. Apoi vom alege clasa pentru care probabilitatea de apartenenţă a instanţei este maximă.

Mai întâi vom calcula produsul pentru clasa *Joc* = *Da* (J_D). Din tabelul sortat după valoarea clasei, se observă că această valoare apare de 9 ori.

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
7	Înnorat	Mică	Normală	Prezent	Da
9	Soare	Mică	Normală	Absent	Da

Inteligenta artificiala - Laborator, http://florinleon.byethost24.com/lab_ia.htm

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
6	Ploaie	Mică	Normală	Prezent	Nu
8	Soare	Medie	Mare	Absent	Nu
14	Ploaie	Medie	Mare	Prezent	Nu

Prin urmare:

$$P(J_D) = \frac{9}{14}.$$

Pentru calculul probabilităţilor condiţionate din produs, numărăm instanţele care au valoarea dorită a atributului, numai în clasa considerată. De exemplu, pentru atributul *Starea vremii*, ne interesează câte instanţe au valoarea *Soare* (dată de instanţa de interogare x_q).

Nr. instanţă	Starea vremii	Joc
9	Soare	Da
11	Soare	Da
3	Înnorat	Da
7	Înnorat	Da
12	Înnorat	Da
13	Înnorat	Da
4	Ploaie	Da
5	Ploaie	Da
10	Ploaie	Da

Din tabelul de mai sus, se poate vedea că numai 2 instanţe din cele 9 din clasa *Da* au valoarea *Soare*. Deci:

$$P(S_S|J_D) = \frac{2}{9}.$$

Analog se procedează şi pentru restul atributelor, obţinând:

$$P(T_H|J_D) = \frac{2}{9}$$
$$P(U_N|J_D) = \frac{6}{9}$$
$$P(V_A|J_D) = \frac{6}{9}.$$

Aceleaşi calcule se realizează pentru clasa *Nu* (*N*):

$$P(J_N) = \frac{5}{14}$$

Inteligenta artificiala - Laborator, http://florinleon.byethost24.com/lab_ia.htm

$$P(S_S|J_N) = \frac{3}{5}$$
$$P(T_H|J_N) = \frac{2}{5}$$
$$P(U_N|J_N) = \frac{1}{5}$$
$$P(V_A|J_N) = \frac{2}{5}.$$

Putem calcula acum produsele pentru fiecare clasă:

$$P(J_D) \cdot P(x_q|J_D) = \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} = 14{,}109 \cdot 10^{-3},$$
$$P(J_N) \cdot P(x_q|J_N) = \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} = 6{,}857 \cdot 10^{-3}.$$

Valoarea maximă este prima şi deci instanţa va fi clasificată în clasa *Da*.

3. Considerente practice

3.1. Corecţia Laplace

Deoarece clasificarea se bazează pe calcularea unor produse, dacă un factor este 0, întregul produs devine 0. Să considerăm următoarea mulţime de antrenare:

Starea vremii	Umiditate	Joc
Înnorat	Mare	Da
Ploaie	Mare	Da
Înnorat	Mare	Da
Soare	Mare	Nu
Soare	Mare	Nu
Ploaie	Normală	Nu

şi instanţa pe care dorim să o clasificăm: x_q = (*Înnorat*, *Normală*).

Mai întâi realizăm calculele pentru clasa *Da*:

$$P(J_D) = \frac{3}{6}$$
$$P(S_I|J_D) = \frac{2}{3}$$
$$P(U_N|J_D) = \frac{0}{3}.$$

Apoi realizăm calculele pentru clasa *Nu*:

$$P(J_N) = \frac{3}{6}$$
$$P(S_I|J_N) = \frac{0}{3}$$
$$P(U_N|J_N) = \frac{1}{3}.$$

Inteligenta artificiala - Laborator, http://florinleon.byethost24.com/lab_ia.htm

Calculând produsele de probabilităţi, se vede că ambele sunt 0 şi deci nu putem lua nicio decizie de clasificare a instanţei x_q:

$$P(J_D) \cdot P(x_q|J_D) = \frac{3}{6} \cdot \frac{2}{3} \cdot \frac{0}{3} = 0,$$

$$P(J_N) \cdot P(x_q|J_N) = \frac{3}{6} \cdot \frac{0}{3} \cdot \frac{1}{3} = 0.$$

În general, pentru mai multe clase, dacă în fiecare produs există câte un factor nul, atunci toate produsele, pentru toate clasele, se anulează. Totuşi, ceilalţi factori nenuli ai produsului ne-ar putea da informaţii relevante pentru clasificare. În acest sens, există metode care garantează că niciun produs nu va fi 0. În locul calculului tipic al frecvenţelor relative ca raport între numărul de apariţii a unei valori a atributului *i* în clasa *j* (n_{ij}) şi numărul de apariţii a valorii clasei *j* (n_j):

$$P(A_i|C_j) = \frac{n_{ij}}{n_j}, \tag{6}$$

se poate folosi *estimarea-m* (engl. "m-estimate") care „netezeşte" probabilităţile aplicând următoarea formulă de calcul:

$$P(A_i|C_j) = \frac{n_{ij} + mp}{n_j + m}. \tag{7}$$

Corecţia Laplace poate fi considerată un caz particular al estimării-m unde, dacă c este numărul de clase, putem considera $m = c$ şi $p = 1/c$, rezultând:

$$P(A_i|C_j) = \frac{n_{ij} + 1}{n_j + c}. \tag{8}$$

Practic, se adaugă la numărător 1 şi la numitor numărul de clase. În acest mod, toţi factorii vor avea valori $v \in (0, 1)$. Probabilităţile sunt estimate ca frecvenţe relative din date, dar nu cunoaştem valorile absolute. Teoretic, ar putea să mai existe instanţe pe care încă nu le-am întâlnit şi atunci considerăm a-priori că mai există câte o instanţă din fiecare clasă.

Din punct de vedere filosofic, faptul că probabilităţile nu pot fi 0 sau 1 ne conduce la ideea că nu putem fi siguri niciodată de ceva că e adevărat sau fals în mod absolut.

Pentru exemplul simplificat, vom avea acum:

$$P(J_D) \cdot P(x_q|J_D) = \frac{3}{6} \cdot \frac{2+1}{3+2} \cdot \frac{0+1}{3+2} = 0{,}06$$

$$P(J_N) \cdot P(x_q|J_N) = \frac{3}{6} \cdot \frac{0+1}{3+2} \cdot \frac{1+1}{3+2} = 0{,}04$$

şi deci se poate lua o decizie (*Da*), iar rezultatul este conform cu analiza mulţimii de antrenare, unde valoarea *Înnorat* apare în clasa *Da* de două ori iar valoarea *Normală* apare în clasa *Nu* o singură dată.

Pentru exemplul iniţial din secţiunea 2, aplicând corecţia Laplace vom avea:

$$P(J_D) \cdot P(x_q|J_D) = \frac{9}{14} \cdot \frac{2+1}{9+2} \cdot \frac{2+1}{9+2} \cdot \frac{6+1}{9+2} \cdot \frac{6+1}{9+2} = 19{,}363 \cdot 10^{-3}$$

$$P(J_N) \cdot P(x_q|J_N) = \frac{5}{14} \cdot \frac{3+1}{5+2} \cdot \frac{2+1}{5+2} \cdot \frac{1+1}{5+2} \cdot \frac{2+1}{5+2} = 10{,}71 \cdot 10^{-3}.$$

Rezultatul clasificării nu se schimbă deoarece contează doar comparaţia, nu cantităţile propriu-zise. Pentru metoda bayesiană naivă, partea calitativă este mai importantă decât partea cantitativă.

Corecţia Laplace este foarte utilă mai ales la clasificarea textelor, unde atributele sunt cuvintele însele şi, având multe documente, este probabil ca din acestea să lipsească anumiţi termeni.

3.2. Precizia calculelor

O altă problemă care poate să apară este faptul că un produs de probabilităţi subunitare cu mulţi factori poate fi afectat de precizia reprezentării numerelor. Astfel, dacă valoarea produsului devine mai mică decât cantitatea minimă care poate fi reprezentată în virgulă mobilă, rezultatul va deveni 0.

O soluţie este *logaritmarea* şi în acest caz, produsul de probabilităţi este înlocuit cu suma logaritmilor de probabilităţi.

De exemplu, pentru calculul $P(J_D) \cdot P(x_q|J_D)$ de mai sus, vom avea:

$$\ln\left(P(J_D) \cdot P(x_q|J_D)\right) = \ln\frac{9}{14} + \ln\frac{2+1}{9+2} + \ln\frac{2+1}{9+2} + \ln\frac{6+1}{9+2} + \ln\frac{6+1}{9+2}$$
$$= -0{,}442 - 1{,}299 - 1{,}299 - 0{,}452 - 0{,}452 = -3{,}924$$
$$\Rightarrow P(J_D) \cdot P(x_d|J_D) = \mathrm{e}^{-3{,}924} \cong 19{,}363 \cdot 10^{-3}.$$

Pentru simplitate, mai sus am inclus doar 3 zecimale. Desigur, pentru o precizie suficientă a rezultatului, reprezentarea termenilor sumei trebuie să fie corespunzătoare.

4. Probleme cu atribute numerice

Pentru atribute numerice, există mai multe modalităţi de abordare. Valorile acestora pot fi discretizate, rezultând atribute ordinale. De asemenea, se poate aplica partiţionarea binară: se alege o valoare de referinţă iar valorile atributului rezultat vor fi *Da* sau *Nu*, dacă valorile atributului iniţial sunt mai mari sau mai mici, respectiv, decât referinţa.

5. Concluzii

Dintre avantajele metodei de clasificare bayesiene naive menţionăm calculele simple şi robusteţea la zgomot şi atribute irelevante. De aceea, este foarte potrivită pentru mulţimi de antrenare de dimensiuni medii sau mari (de exemplu pentru clasificarea documentelor text, detecţia spam-ului, diagnoză etc.).

Chiar dacă se bazează pe independenţa atributelor dată fiind clasa, metoda funcţionează de multe ori bine chiar şi atunci când presupunerea este infirmată în realitate.

Inteligenta artificiala - Laborator, http://florinleon.byethost24.com/lab_ia.htm

6. Aplicaţie

Implementaţi clasificatorul bayesian naiv pentru problema definită mai jos. Intrările şi ieşirea (clasa) sunt precizate în formatul nume : valori-posibile. Valorile atributelor din secţiunea de antrenare sunt despărţite de spaţii albe.

INPUT
Outlook : sunny overcast rainy
Temperature : cool mild hot
Humidity : normal high
Windy : false true

OUTPUT
Play : no yes

TRAINING
sunny hot high false no
sunny hot high true no
overcast hot high false yes
rainy mild high false yes
rainy cool normal false yes
rainy cool normal true no
overcast cool normal true yes
sunny mild high false no
sunny cool normal false yes
rainy mild normal false yes
sunny mild normal true yes
overcast mild high true yes
overcast hot normal false yes
rainy mild high true no

Programul trebuie să permită clasificarea unor instanţe precizate de utilizator. Se vor afişa probabilităţile ambelor clase şi se va evidenţia clasa cu probabilitate maximă, adică decizia de clasificare. Utilizatorul va avea posibilitatea să utilizeze sau nu corecţia Laplace.

Inteligenta artificiala - Laborator, http://florinleon.byethost24.com/lab_ia.htm