# Naive Bayes

Naive Bayes este metodă de clasificare ce are la bază teorema elaborată de Thomas Bayes (Eq. 1). În caz general, presupunem că datele de intrare se reprezintă printr-un vector de caracteristici (*feature vector*) notat $x = (x_1, ..., x_n)$. Acest vector conține o serie de valori pentru $n$ atribute specifice fiecărei instanțe din setul de date (altfel spus, conține valorile caracteristicilor fiecărui element din mulțimea de obiecte inițială). Pentru obiectele dintr-o mulțime de date de un anumit tip se folosește denumirea mult mai generală de *instanță*. Presupunem că instanțele se încadrează în $K$ clase, notate $c_1$ ... $c_K$. Pentru o nouă instanță, neclasificată, dorim să determinăm probabilitățile de încadrare a acesteia în cele $K$ clase, pe baza analizei caracteristicilor instanțelor cărora li se cunosc deja clasele.

$$P(c_k \mid X) = \frac{P(X \mid c_k)\, P(c_k)}{P(X)}, k = 1..K \qquad (1)$$

Eq. 1 se interpretează astfel:

$P(c_k \mid X)$ este probabilitatea ca o nouă instanță cu atributele date de vectorul $X$ să aparțină clasei $c_k$ (Este ceea de dorim sa determinăm.)

$P(X \mid c_k)$ este probabilitatea ca în clasa $c_k$ să existe o instanță cu vectorul de atribute $X$

$P(c_k)$ este probabilitatea ca o instanță să se încadreze în clasa $c_k$

$P(X)$ este probabilitatea de apariție a unei instanțe cu vectorul de atribute $X$

În cazul (cel mai frecvent întâlnit) în care instanțele au mai multe atribute, probabilitatea de apariție a unei instanțe cu un vector de valori ale acestor atribute în clasa oarecare $c_k$ se calculează conform Eq. 2.

$$P(X \mid c_k) = \prod_{i=1}^{n} P(x_i \mid c_k) \qquad (2)$$

Interpretarea Eq. 2 este următoarea: presupunând că atributele sunt independente unele de altele, probabilitatea ca în clasa $c_k$ să apară o instanță cu vectorul $X$ cu $n$ atribute se calculează ca fiind produsul probabilităților ca în clasa $c_k$ să apară o instanță cu valoarea $x_i$ a atributului $i$.

Odată determinate probabilitățile de încadrare a noii instanțe în cele $c_k$ *clase,* se consideră că aceasta aparține clasei pentru care s-a obținut probabilitatea maximă.

Având în vedere faptul că termenul *P(X)* din Eq. 1 este același pentru toate clasele $c_k$, el poate fi ignorat în contextul în care trebuie maximizat termenul din partea dreaptă a Eq. 1. Luând în considerare acest fapt, precum și Eq. 2, calculul probabilităților se poate simplifica ( Eq. 3).

$$P(c_k \mid X) = P(c_k) \prod_{i=1}^{n} P(x_i \mid c_k) \qquad (3)$$

**Exemplu:**

Considerăm următorul set de date:

| 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 |

| Nr. instanţă | Starea vremii | Temperatură | Umiditate | Vânt | Joc |
|---|---|---|---|---|---|
| 15 | Soare | Mare | Normală | Absent | ??? |

Problema care se se pune e legată de oportunitatea de a practica un sport oarecare (“Joc”), având în vedere condițiile date de starea vremii, temperatură, umiditate și vânt. Setul de date cu care se pornește este dat de instanțele 1..14, pentru care se cunoaște clasa (sunt două clase, Da și Nu). Fiecare instanță constă în vectorul de valori ale celor 4 atribute.

Fiind dată o nouă instanță, a 15-a, în ce clasă se încadrează aceasta? Altfel spus, se dorește sau nu practicarea sportului în condițiile meteorologice definite de valorile atributelor instanței 15, având în vedere deciziile care s-au luat anterior, în cazul celorlalte 14 instanțe?

Trebuie să calculăm probabilitățile ca instanța 15 să se încadreze în cele două clase Da, Nu, și să alegem clasa corespunzătoare probabilității maxime.

- Calculăm probabilitatea de apariție a clasei Da, notată $J_D$. Observăm că ea apare de 9 ori, din cele 14 cazuri posibile. Prin urmare,

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

- Calculăm probabilitățile condiționate de apartenența la clase (termenii produsului din Eq. 3). Pentru atributul *Starea vremii*, observăm că 2 instanțe din cele 9 din clasa *Da* au valoarea *Soare* (cea care provine din instanța 15, nou apărută), notată $S_s$. Așadar,

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

În mod similar, determinăm probabilitățile valorilor atributelor instanței 15, condiționate de apartenența la clasa *Da*.

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

Calculăm probabilitatea ca noua instanță să aparțină clasei *Da*, conform cu Eq. 3:

$$P(J_D) \cdot P(x_i|J_D) = \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} = 14.109 \cdot 10^{-3}$$

Efectuând aceleași calcule pentru clasa *Nu*, obținem:

$$P(J_N) \cdot P(x_i|J_N) = \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} = 6.857 \cdot 10^{-3}$$

Valoarea maximă este cea corespunzătoare clasei *Da*, în care se va încadra noua instanță.

| Nr. instanță | Starea vremii | Temperatură | Umiditate | Vânt | Joc |
|---|---|---|---|---|---|
| 15 | Soare | Mare | Normală | Absent | **Da** |

## Corecția Laplace

Există posibilitatea ca, pentru anumite seturi de date, valorile unora dintre termenii produselor din Eq. 3 să fie nule, caz în care produsul corespunzător va fi, la rândul său, nul.

**Exemplu:**

Fie următorul set de date:

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

Dorim să clasificăm următoarea instanță:

| Starea vremii | Umiditate | Joc |
|---|---|---|
| Înnorat | Normală | Nu |

Realizând aceleași calcule ca și în cazul anterior, obținem:

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

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

Pentru a preîntâmpina această situație, considerăm modul în care se calculează probabilitatea de apariție a atributului $x_i$ printre valorile instanțelor din clasa $c_j$ (Eq. 4), unde $n_{ij}$ este numărul de apariții a unei valori a atributului *i* în clasa *j*, iar $n_j$ este numărul de apariții al clasei *j*.

$$P(x_i|c_j) = \frac{n_{ij}}{n_j} \tag{4}$$

Corecția Laplace se aplică conform cu Eq. 5, unde *c* este numărul de clase:

$$P(x_i|c_j) = \frac{n_{ij} + 1}{n_j + c} \tag{5}$$

Această abordare garantează faptul că termenii produsului vor fi nenuli. În cazul exemplului anterior, calculele vor fi:

$$P(J_D) \cdot P(x_i|J_D) = \frac{3}{6} \cdot \frac{2+1}{3+2} \cdot \frac{0+1}{3+2} = 0.06$$

$$P(J_N) \cdot P(x_i|J_N) = \frac{3}{6} \cdot \frac{0+1}{3+2} \cdot \frac{1+1}{3+2} = 0.04$$

Prin urmare, se poate decide faptul că noua instanță se va încadra în clasa *Da*.

În cazul exemplului inițial, vom avea:

$$P(J_D) \cdot P(x_i|J_D) = \frac{9}{14} \cdot \frac{2+1}{9+2} \cdot \frac{2+1}{9+2} \cdot \frac{6+1}{9+2} \cdot \frac{6+1}{9+2} = 19.363 \cdot 10^{-3}$$

$$P(J_N) \cdot P(x_i|J_N) = \frac{5}{14} \cdot \frac{3+1}{5+2} \cdot \frac{2+1}{5+2} \cdot \frac{1+1}{5+2} \cdot \frac{2+1}{5+2} = 10.71 \cdot 10^{-3}.$$

Rezultatul clasificării nu se schimbă, deoarece contează doar comparația, nu și valorile propriu–zise.

**Cerințe:**

1) Implementați metoda de clasificare Naive Bayes folosind setul de date din fișierul *data_vreme1.csv:*

| Starea vremii | Temperatura | Umiditate | Vant | Joc |
|---|---|---|---|---|
| Soare | Mare | Mare | Absent | Nu |
| Soare | Mare | Mare | Prezent | Nu |
| Innorat | Mare | Mare | Absent | Da |
| Ploaie | Medie | Mare | Absent | Da |
| Ploaie | Mica | Normala | Absent | Da |
| Ploaie | Mica | Normala | Prezent | Nu |
| Innorat | Mica | Normala | Prezent | Da |
| Soare | Medie | Mare | Absent | Nu |
| Soare | Mica | Normala | Absent | Da |
| Ploaie | Medie | Normala | Absent | Da |
| Soare | Medie | Normala | Prezent | Da |
| Innorat | Medie | Mare | Prezent | Da |
| Innorat | Mare | Normala | Absent | Da |
| Ploaie | Medie | Mare | Prezent | Nu |

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.
Pentru verificare, clasificați instanța 15 menționată anterior.

2) Implementați metoda de clasificare Naive Bayes folosind setul de date din fișierul *data_vreme2.csv:*

| Starea vremii | Temperatura | Umiditate | Vant | Joc |
|---|---|---|---|---|
| Soare | 17 | Mare | Absent | Nu |
| Soare | 15 | Mare | Prezent | Nu |
| Innorat | 24 | Mare | Absent | Da |
| Ploaie | 19 | Mare | Absent | Da |
| Ploaie | 19 | Normala | Absent | Da |
| Ploaie | 12 | Normala | Prezent | Nu |
| Innorat | 20 | Normala | Prezent | Da |
| Soare | 18 | Mare | Absent | Nu |
| Soare | 18 | Normala | Absent | Da |
| Ploaie | 17 | Normala | Absent | Da |
| Soare | 19 | Normala | Prezent | Da |
| Innorat | 15 | Mare | Prezent | Da |
| Innorat | 20 | Normala | Absent | Da |
| Ploaie | 18 | Mare | Prezent | Nu |

De data aceasta, atributul **Temperatura** are valori numerice, a căror prelucrare necesită o altă abordare: pentru valorile acestui atribut, vom determina probabilitățile folosind funcția densitate de probabilitate a distribuției acestora.

Pentru fiecare clasă (Da, Nu) valorile temperaturii sunt generate conform cu distribuția normală. Așadar pentru aceste valori se poate obține probabilitatea în raport cu clasa corespunzătoare folosind funcția densitate de probabilitate a distribuției normale, cu următoarea expresie:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6)$$

unde $\mu$ (“miu”) și $\sigma$ (“sigma”) sunt *media*, respectiv *abaterea standard* a distribuției. Aceste valori trebuie determinate separat pentru fiecare clasă, din valorile cunoscute ale temperaturii.

Media $\mu$ se determină ca fiind media aritmetică a valorilor. Presupunând n valori ale temperaturii, $t_1$ .. $t_n$ , formula este următoarea:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} t_i \quad (7)$$

Abaterea standard $\sigma$ se determină astfel:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \mu)^2} \quad (8)$$

**Pentru verificare:**

Pentru datele din fișierul *data_vreme2.csv*, rezultă următoarele valori:

$\mu_{DA}$ = 19
$\sigma_{DA}$ = 2.3094
$\mu_{NU}$ = 16
$\sigma_{NU}$ = 2.28035

Pentru următoarea instanță:

| Nr. instanţă | Starea vremii | Temperatură | Umiditate | Vânt | Joc |
|---|---|---|---|---|---|
| 15 | Soare | 14 | Normală | Absent | ??? |

Probabilitatea valorii temperaturii rezultă, pentru cele două clase, din funcția densitate de probabilitate (Eq. 6) cu parametrii $\mu$ și $\sigma$ determinați pentru fiecare clasă:

*P(Temp = 14 | Da)* = 0.0166
*P(Temp = 14 | Nu)* = 0.1190

Calculăm probabilitatea ca noua instanță să aparțină clasei *Da*, conform cu Eq. 3:

$$P(J_D) \cdot P(x_i|J_D) = 1.05 \cdot 10^{-3}$$

Efectuând aceleași calcule pentru clasa *Nu*, obținem:

$$P(J_N) \cdot P(x_i|J_N) = 2.04 \cdot 10^{-3}$$

Așadar, instanța anterioară se clasifică astfel:

| Nr. instanţă | Starea vremii | Temperatură | Umiditate | Vânt | Joc |
|---|---|---|---|---|---|
| 15 | Soare | 14 | Normală | Absent | Nu |