Discriminarea Bayesiană

Probabilitate totală şi probabilitate Bayesiană

Fie Ω un sistem complet de evenimente. Probabilitatea de apariţie a unui eveniment A aparţinând lui Ω este dată de relaţia probabilităţii totale:

$P(A) = P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + ... + P(A_n)P(A/A_n)$,

unde $P(A_i)$, i=1,n, sunt probabilităţile de apariţie a celorlalte evenimente, iar $P(A/A_i)$, i=1,n, sunt probabilităţile de apariţie în simultaneitate a evenimentului A cu celelalte evenimente - probabilitățile condiționate.

$P(A)$ este deci o medie ponderată a probabilităţii de apariţie a evenimentului A împreună cu fiecare din celelalte evenimente din Ω.

O importanţă deosebită o are cunoaşterea probabilităţii ca un eveniment să se producă pornind de la o anumită cauză din mulţimea de evenimente care-l determină, $P(A_k/A)$. Relaţia prin care se determină această probabilitate se numeşte relaţia lui Bayes: $P(A_k / A) = \dfrac{P(A_k)P(A/A_k)}{\sum_{i=1}^{n} P(A_i)P(A/A_i)}$.

Probabilităţile $P(A_i)$, i=1,n se numesc probabilităţi apriorice, iar probabilităţile $P(A_i/A)$ se numesc probabilităţi aposteriorice.

Exemplu.

Conducătorii auto se împart în două categorii:

1. cei care produc accidente
2. cei care nu produc accidente.

S-a constatat că 5% din femeile conducător auto şi 20% din bărbaţi au avut cel puţin un accident. Deci probabilitatea ca o femeie să producă un accident este de 0.05, iar probabilitatea ca un bărbat să producă accident este de 0.2 ($P(A/A_1$=0.05, $P(A/A_2$=0.2). Cunoscând ponderile pe sexe în cadrul populaţiei cu carnet auto, 35% şi 65% (probabilitățile ca un conducător auto să fie femeie, respectiv, bărbat, $P(A_1)$=0.35, $P(A_2)$=0.65), se poate determina *probabilitatea ca un conducător auto să producă accident*:

$P(A) = 0.35\times0.05+0.65\times0.2 = 0.1475$

Probabilitatea ca un accident să fie produs de o femeie este:

$P(A_1/A) = \dfrac{0.35\times0.05}{0.1475} \approx 0.1186$

Probabilitatea ca un accident să fie produs de un bărbat este:

$P(A_2/A) = \dfrac{0.65\times0.2}{0.1475} \approx 0.8814$

Clasificatorul bayesian

Clasificatorul bayesian este de natură stohastică (probabilistică) și permite o clasificare cu bune rezultate la un cost scăzut. Notăm cu $g_k = \begin{bmatrix} g_{k1} \\ ... \\ g_{km} \end{bmatrix}$ centrul grupei k, $k = \overline{1,q}$. Vom eticheta grupele cu $c_1, c_2, ..., c_q$.

Regula lui Bayes pentru asignarea unei instanțe $x = \begin{bmatrix} x_1 \\ ... \\ x_m \end{bmatrix}$ la o clasă c_k mai degrabă decât la orice altă clasă c_l este următoarea:

$P(c_k/x) > P(c_l /x)$, $l=\overline{1,q}$, $l\neq k$ (1)

unde $P(c_k/x)$ reprezintă probabilitatea ca instanța x să aparţină clasei c_k, iar $P(c_l /x)$ reprezintă probabilitatea ca aceeași instanță să aparţină clasei c_l. Acestea sunt numite ***probabilităţi aposteriorice***.

Vom numi ***probabilităţi apriorice***, probabilităţile de apartenenţă la o anumită clasă în cazul în care nici o informaţie despre indivizii claselor nu este disponibilă. De obicei probabilităţile apriorice sau sunt determinate pe baza ponderii claselor, sau sunt egale cu $1/q$ sau au valori alese în mod subiectiv.

Vom considera probabilitatea apriorică pentru o clasă oarecare c_l, $P(c_l) = \dfrac{n_l}{n}$, $l=\overline{1,q}$.

Relaţia lui Bayes leagă probabilităţile apriorice de cele aposteriorice în felul următor:

$$P(c_k / x) = \frac{P(x / c_k) \cdot P(c_k)}{\sum_{l=1}^{q} P(x / c_l) P(c_l)} . \qquad (2)$$

În relaţia (2) $P(x / c_l)$ reprezintă probabilitatea ca în distribuția grupei c_l să avem o observație x. Aceste probabilităţi, probabilitățile condiționate din relația lui Bayes, reprezintă estimări ale densităților de repartiție pentru fiecare grupă. Estimarea lor corectă este decisivă pentru o clasificare cu bune rezultate (acuratețea modelului) Există două tipuri de metode de estimare a probabilităților condiționate: parametrice și neparametrice.

Estimarea neparamerică a probabilităților condiționate. Metoda histogramelor

Pentru cazul unidimensional, estimarea pentru o valoarea oarecare x este: $\hat{p}(x) = \frac{n_j}{n \cdot dx}$,

unde n este numărul total de valori, dx este mărimea intervalului (lățimea histogramei), j este numărul intervalului (histograma) în care se află x, iar n_k este numărul de valori din distribuție aflate în intervalul k.

Pentru cazul bidimensional: $\hat{p}(x) = \frac{n_j}{n \cdot dx_1 \cdot dx_2}$,

unde dx_1 este mărimea intervalului pentru prima dimensiune iar dx_2 pentru a doua dimensiune (Figura 1). Produsul $dx_1 \cdot dx_2$ este aria histogramei în care se află x.

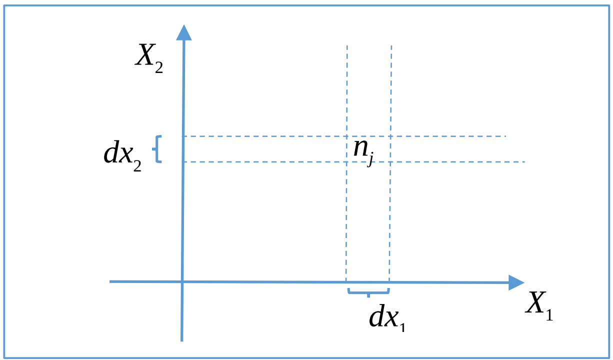

Figura 1. Estimare densitate de probabilitate

Pentru cazul m-dimensional: $\hat{p}(x) = \frac{n_j}{n \cdot H}$,

unde $H = \prod_{j=1}^{m} dx_j$ este volumul histogramei (partea din spaţiul R^m în care se află x).

Alte metode neparametrice de estimare a probabilităților condiționate: Ferestre Parzen, Metode Kernel.

Metode parametrice

Presupun asumarea unui model parametric pentru funcţia de densitate şi aplicarea ca atare a relaţiei lui Bayes.

Dacă folosim distribuţia normală Gausiană pentru estimarea probabilităţilor condiţionate, vom avea:

$$P(x / c_k) = (2\pi)^{-\frac{m}{2}} |T|^{-\frac{1}{2}} \cdot e^{\left(-\frac{1}{2}(x-g_k)^t T^{-1}(x-g_k)\right)},$$

unde T este matricea de covarianţă, iar g_k este centrul clasei c_k.

Ţinând cont de relaţia lui Bayes regula (1) devine:

vom asigna individul x clasei c_l dacă

$$P(x/c_l)P(c_l) > P(x/c_k)P(c_k), \quad k=\overline{1,q}, \ l \neq k. \qquad (3)$$

Înlocuind probabilitățile condiționate în regula lui Bayes (3), logaritmând şi simplificând termenii comuni, obţinem:

este asignat x clasei c_l dacă:

$$2 \cdot \ln P(c_l) - (x-g_l)^t T^{-1}(x-g_l) > 2 \cdot \ln P(c_k) - (x-g_k)^t T^{-1}(x-g_k) \text{ pentru orice } k=\overline{1,q}, \ l \neq k. \qquad (4)$$

Rezultă:

este asignată instanța x clasei c_l dacă

$2 \cdot \ln P(c_l)$ - $d^2_{T^{-1}}(x, g_l)$ > $2 \cdot \ln P(c_k)$ - $d^2_{T^{-1}}(x, g_k)$, pentru orice $k=\overline{1,q}$, $l \neq k$. (5)

unde $d^2_{T^{-1}}(x, g_k) = (x - g_k)^t T^{-1} (x - g_k)$ $k=\overline{1,q}$, sunt distanțele Mahalanobis dintre instanța x și centrii celor q grupe.

Se poate observa că diferența față de clasificarea liniară este dată de ponderile grupelor (termenii $2 \cdot \ln P(c_k)$).

Pentru a pune în evidență mai bine această diferență, vom calcula funcțiile de clasificare și scorurile pentru modelul bayesian parametric. Regula lui Bayes (5) de mai sus se poate scrie:

este asignată instanța x clasei c_l dacă

$d^2_{T^{-1}}(x, g_l)$ - $2 \cdot \ln P(c_l)$ < $d^2_{T^{-1}}(x, g_k)$ - $2 \cdot \ln P(c_k)$, pentru orice $k=\overline{1,q}$, $l \neq k$. (6)

Dacă înlocuim distanțele Mahalanobis așa cum am făcut la analiza discriminantă liniară pentru calculul funcțiilor de clasificare, obținem:

$(x - g_l)^t T^{-1} (x - g_l)$ - $2 \cdot \ln P(c_l)$ < $(x - g_k)^t T^{-1} (x - g_k)$ - $2 \cdot \ln P(c_k)$, pentru orice $k=\overline{1,q}$, $l \neq k$.

Dezvoltând relațiile și simplificând exact în același mod ca la analiza discriminantă liniară (funcții de clasificare), obținem regula de clasificare Bayes exprimată cu ajutorul funcțiilor de clasificare:

este asignată instanța x clasei c_l dacă

$(F_l)^t \cdot x + F_{l0} > (F_k)^t \cdot x + F_{k0}$, pentru orice $k=\overline{1,q}$, $l \neq k$. (7)

unde F_l și F_k sunt funcțiile de clasificare calculate exact ca la analiza liniară discriminantă, iar F_{l0} = $-\frac{1}{2}(F_l)^t \cdot g_l + \ln P(c_l)$, $F_{k0} = -\frac{1}{2}(F_k)^t \cdot g_k + \ln P(c_k)$.

Relațiile de calcul pentru scoruri, prin intermediul funcțiilor de clasificare sunt:

$S_k(x) = F_{k0} + F_{k1} \cdot x_1 + F_{k2} \cdot x_2 + \ldots + F_{km} \cdot x_m$, $k = 1,q$,

unde:

$F_k = \begin{bmatrix} F_{k1} \\ \ldots \\ F_{km} \end{bmatrix} = g_k^t T^{-1}$, $x = \begin{bmatrix} x_1 \\ \ldots \\ x_m \end{bmatrix}$ este instanța clasificată, iar $F_{k0} = -\frac{1}{2}(F_k)^t \cdot g_k + \ln P(c_k)$.

Instanța x va fi clasificată în acea grupă, l, pentru care $S_l(x) > S_k(x)$, $k = 1,q$, $l \neq k$.