1 Wstęp

Przedstawimy teraz w sposób ogólny zadanie rozpoznawania. Rozpoznawany obiekt pochodzi z jednej z kilku klas. Wykonywany jest na nim pewien zestaw pomiarów, na podstawie którego należy ustalić z której klasy on pochodzi, czyli rozpoznać. Zadanie polega na wyznaczeniu najlepszego sposobu rozpoznawania, czyli takiego który myli się najrzadziej.

Przykładem może być zadanie diagnozy medycznej, którą lekarz podejmuje wykorzystując informację, różnego rodzaju. Jest nią ogólna wiedza nabyta z literatury przedmiotu, którą będziemy nazywać aprioryczną, jak również własne praktyczne doświadczenie, czego odpwiednikiem będzie tzw. ciąg uczący. Wspomniany zestaw pomiarów, to np. wyniki badań wykonanych wobec pacjenta, czyli informacja empiryczna.

W naszych rozważaniach symbol $\int$ należy rozumieć jako $\int_{-\infty}^{\infty}$.

2 Przedstawienie problemu

W parze (θ, X) zmiennych losowych, θ przyjmuje wartości w zbiorze $M = \{1, 2, \ldots, m\}$, elementy którego nazywają się klasami, X na prostej $(-\infty, \infty)$. Gęstością prawdopodobieństwa w klasie i, czyli gęstością warunkową, jest $f_i(x) = f(x|\theta = i)$. Prawdopodobieństwa klas, to $p_i = P\{\theta = i\}$, przy czym $\sum_{i=1}^{m} p_i = 1$. Gęstości $f_i(x)$ oraz prawdopodobieństwa p_i składają się na tzw. informację aprioryczną.

Bezwarunkową gęstością zmiennej X jest zatem

$$f(x) = \sum_{i=1}^{m} p_i f_i(x).$$

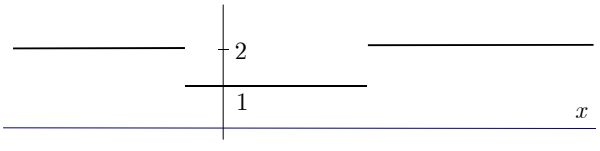
Prawdopodobieństwo *a posteriori* klasy i, czyli po wykonaniu obserwacji $X = x$, to

$$P\{\theta = i|X = x\} = \frac{p_i f_i(x)}{f(x)}, \tag{2.1}$$

$i = 1, \ldots, m$.

Zadanie polega na rozpoznaniu klasy θ na podstawie X. Rozpoznanie to, czyli decyzja o przynależności do klasy, może być poprawna lub błędna. Wyznaczymy teraz najlepszy sposób podejmowania decyzji, czyli rozpoznawania.

Zaczniemy od pojęcia reguły rozpoznawania. Definiuje się ją jako funkcję, oznaczmy ją jako $\psi(x)$, która każdemu punktowi $x \in (-\infty, \infty)$ przyporządkowuje element ze zbioru M, rys. 2.1. Jeśli zatem pomiar X wykonany na klasyfikowanym obiekcie jest równy x, to reguła ta klasyfikuje go do klasy $\psi(x)$. Ze względu na to, że estymowana zmienna θ jest natury losowej, zadanie jest tzw. problemem Bayesa.



Rys. 2.1: Przykładowa reguła rozpoznawania, $m = 2$.

Wyznaczymy teraz regułę optymalną $\psi^*(x)$, czyli taką, która zapewnia najmniejsze prawdopodobieństwo pomyłki $P\{\psi(X) \neq \theta\}$, czyli błędnej klasyfikacji. Wprowadźmy w tym celu pojęcie funkcji strat zdefiniowanej jako

$$L(i, j) = \begin{cases} 0, & \text{dla } i = j \\ 1, & \text{dla } i \neq j, \end{cases}$$

gdzie j jest klasą, do której należy X, a i decyzją. W przypadku błędnej klasyfikacji ponoszona jest zatem strata w wysokości 1, przy prawidłowej stratą jest 0. Jest więc oczywiste, że prawdopodobieństwo błędnej klasyfikacji można wyrazić jako

$$P\{\psi(X) \neq \theta\} = EL(\psi(X), \theta),$$

która to wielkość nazywa się ryzykiem.

Aby je wyliczyć, zauważmy, że

$$\begin{aligned} EL(\psi(X), \theta) &= \sum_{j=1}^{m} E\{L(\psi(X), \theta)|\theta = j\} P\{\theta = j\} \\ &= \sum_{j=1}^{m} \int L(\psi(x), j) p_j f_j(x) dx \\ &= \int \left(\sum_{j=1}^{m} L(\psi(x), j) p_j f_j(x) \right) dx. \end{aligned}$$

Zatem

$$\begin{aligned} EL(\psi^*(X), \theta) &= \min_{\psi} \int \left(\sum_{j=1}^{m} L(\psi(x), j) p_j f_j(x) \right) dx \\ &= \int \min_{\psi(x)} \left(\sum_{j=1}^{m} L(\psi(x), j) p_j f_j(x) \right) dx. \end{aligned}$$

Dla ustalonego x optymalna reguła jest więc równa temu i, które minimalizuje wyrażenie

$$\begin{aligned} &\sum_{j=1}^{m} L(i, j) p_j f_j(x) \\ &= L(i, 1) p_1 f_1(x) + \cdots + L(i, i) p_i f_i(x) + \cdots \\ &+ L(j, m) p_m f_m(x). \end{aligned}$$

Jeśli zatem $\psi(x) = i$, to powyższa wartość jest równa

$$\begin{aligned} \sum_{j=1, j\neq i}^{m} p_j f_j(x) &= \sum_{j=1}^{m} p_j f_j(x) - p_i f_i(x) \\ &= f(x) - p_i f_i(x), \end{aligned}$$

bowiem $L(i, j) = 1$ dla wszystkich $j \neq i$ i jedynie $L(i, i) = 0$. Jej wartość minimalna, to

$$f(x) - \max_{i \in M} p_i f_i(x),$$

którą przyjmuje dla

$$\psi^*(x) = \arg\max_{i \in M} p_i f_i(x). \tag{2.2}$$

Zatem

$$\begin{aligned} EL(\psi^*(X), \theta) &= \int \left(\sum_{j=1}^{m} L(\psi^*(x), j) p_j f_j(x) \right) dx \\ &= \int \left(f(x) - \max_{i \in M} p_i f_i(x) \right) dx \\ &= 1 - \int \left(\max_{i \in M} p_i f_i(x) \right) dx. \end{aligned}$$

W rezultacie

$$P\{\psi^*(X) \neq \theta\} = 1 - \int \left(\max_{i \in M} p_i f_i(x) \right) dx.$$

Regułę optymalną można przedstawić w nieco innej formie. Wprowadzając wskaźnik klasy i

$$I_{\{\theta=i\}} = \begin{cases} 1, & \text{dla } \theta = i \\ 0, & \text{dla } \theta \neq i, \end{cases}$$

zauważmy, że funkcję regresji

$$r_i(x) = E\{I_{\{\theta=i\}} | X = x\}, \tag{2.3}$$

można wyrazić wzorem

$$r_i(x) = \frac{p_i f_i(x)}{f(x)} = P\{\theta = i | X = x\}. \tag{2.4}$$

patrz (2.1). Zatem optymalna reguła (2.2) operująca gęstościami jest równoważna poniższej

$$\psi^*(x) = \arg\max_{i \in M} r_i(x). \tag{2.5}$$

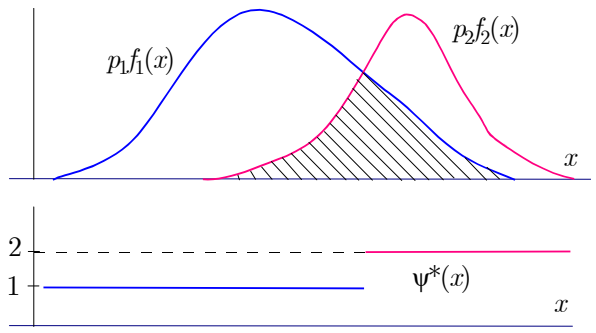
która odwołuje się do funkcji regresji. Zalicza ona obserwację X do klasy i, dla której prawdopodobieństwo *a posteriori* jest największe.

Uwaga 2.1 (dychotomia) *W problemie dychotomii, czyli dwóch klas,*

$$\begin{aligned} \psi^*(x) &= \begin{cases} 1, & \text{jeśli } p_2 f_2(x) \leq p_1 f_1(x) \\ 2, & \text{w przeciwnej sytuacji} \end{cases} \\ &= \begin{cases} 1, & \text{jeśli } r_2(x) \leq r_1(x) \\ 2, & \text{w przeciwnej sytuacji,} \end{cases} \end{aligned}$$

rys. 2.2, oraz

$$\begin{aligned} & P\{\psi^*(X) \neq \theta\} = EL(\psi^*(X), \theta) \\ = & \int [L(\psi^*(x), 1) p_1 f_1(x) + L(\psi^*(x), 2) p_2 f_2(x)] \, dx \\ = & \int \min [p_1 f_1(x), p_2 f_2(x)] \, dx. \end{aligned}$$



Rys. 2.2: Dychotomia, $m = 2$. Ilustracja reguły optymalnej $\psi^*(x)$. Pole zakreskowane jest prawdopodobieństwem jej błędu.

3 Algorytmy uczenia

Zakładamy, że ani gęstości w klasach $f_i(x)$, ani prawdopodobieństwa p_i nie są znane. Dysponujemy natomiast ciągiem uczącym

$$(\theta_1, X_1), (\theta_2, X_2), (\theta_3, X_3), \ldots, (\theta_n, X_n), \tag{3.1}$$

tzn. ciągiem niezależnych realizacji pary (θ, X). Oznacza to, że znamy kolejne obiekty X_i i klasy θ_i, do których one należą. Na podstawie tego ciągu można estymować nieznane rozkłady prawdopodobieństwa i rezultaty wykorzystać w algorytmach rozpoznawania.

Oznaczmy teraz przez $\psi_n(x)$ empiryczną regułę rozpoznawania, czyli regułę opracowaną na podstawie ciągu uczącego (3.1). Jeśli zatem $\psi_n(x) = i$, to reguła ta zalicza x do klasy i.

Wszystkie podane poniżej algorytmy są asymptotycznie optymalne, co oznacza to, że

$$\lim_{n \to \infty} P\{\psi_n(X) \neq \theta\} = P\{\psi^*(X) \neq \theta\}.$$

Własność ta wynika z tego, że wykorzystują one nieparametryczne estymatory gęstości i regresji, które są zgodne, tzn. zbieżne do gęstości w poszczególnych klasach i funkcji regresji charakterystycznych dla tych klas.

W dalszej części symbol $\#\{\}$ oznacza liczbę elementów spełniających warunek zawarty w $\{\}$, np. jeśli $a_1 = 3, a_2 = 5, a_3 = 7$, to $\#\{a_i : a_i \leq 5\} = 2$.

3.1 Estymacja gęstości

Znany nam estymator jądrowy gęstości $f_i(x)$ przyjmuje postać

$$\widehat{f}_i(x) = \frac{1}{N_i h(N_i)} \sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(N_i)} \right),$$

gdzie N_i jest liczbą obiektów z klasy i, czyli

$$N_i = \sum_{j=1}^{n} I_{\{\theta_j = i\}}.$$

Dzięki użyciu wskaźnika $I_{\{\theta_j=i\}}$ zależy on jedynie od X_j pochodzących z klasy i, czyli tych X_j, dla których $\theta_j = i$. Dla odpowiednio wybranej funkcji jądrowej i ciągu liczbowego spełniającego warunki

$$h(n) \overset{n \to \infty}{\to} 0, nh(n) \overset{n \to \infty}{\to} \infty, \tag{3.2}$$

estymator ten jest zgodny, tzn. zbiega się do $f_i(x)$, gdy liczba obserwacji narasta do nieskończoności.

Estymator prawdopodobieństwa p_i ma natomiast oczywistą postać:

$$\widehat{p}_i = \frac{N_i}{n}.$$

W rezultacie jako empiryczną regułę rozpoznawania przyjmujemy

$$\arg\max_i [\widehat{p}_i \widehat{f}_i(x)],$$

czyli

$$\widehat{\psi}_n(x) = \arg\max_i \left[\frac{1}{h(N_i)} \sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(N_i)} \right) \right].$$

Jeśli zastosuje się jądro prostokątne

$$K(x) = \begin{cases} \frac{1}{2}, & \text{dla } |x| \leq 1 \\ 0, & \text{dla } |x| > 1, \end{cases}$$

to

$$\begin{aligned} & \sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(N_i)}\right) \\ = \; & \#\{X_j : \theta_j = i, |X_j - x| \leq h(N_i)\}. \end{aligned}$$

Prawa strona tego wyrażenia jest więc liczbą obserwacji z klasy i zawartych w odcinku $[x - h(N_i), x + h(N_i)]$. Reguła przyjmuje wówczas postać

$$\arg\max_i \left[\frac{1}{h(N_i)} \#\{X_j : \theta_j = i, |X_j - x| \leq h(N_i)\}\right]$$

Wykorzystanie metody NN, czyli najbliższego sąsiada, prowadzi do estymatora

$$\overline{f}_n(x) = \frac{k(N_i)}{2N_i d_i(x; k(N_i))},$$

gdzie $d_i(x; k(N_i))$ jest odległością pomiędzy punktem x a k-tym najbliższym mu pomiarem spośród pochodzących z klasy i. Ciąg liczbowy $k(n)$ spełnia warunki

$$k(n) \overset{n\to\infty}{\to} \infty, \frac{k(n)}{n} \overset{n\to\infty}{\to} 0, \tag{3.3}$$

ponieważ, jak wiemy, zapewniają one zgodność estymatora. W rezultacie reguła rozpoznawania przyjmuje postać

$$\overline{\psi}_n(x) = \arg\max_i \frac{k(N_i)}{d_i(x; k(N_i))}.$$

3.2 Estymacja regresji

Nieco inne i prostrze algorytmy uzyskuje się wychodząc z (2.5) i estymując funkcje regresji (2.3).

Metoda jądrowa prowadzi do estymatora regresji o postaci

$$\widehat{r}_i(x) = \frac{1}{nh(n)} \sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(n)}\right),$$

przy czym funkcja K i ciąg liczbowy $h(n)$ spełniają warunki jak w jądrowym estymatorze gęstości. Wynika stąd poniższy algorytm rozpoznawania:

$$\widehat{\psi}_n(x) = \arg\max_i \left[\sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(n)}\right)\right].$$

Dla jądra prostokątnego przyjmuje on postać

$$\widehat{\psi}_n(x) = \arg\max_i \#\{X_j : \theta_j = i, |X_j - x| \leq h(n)\}, \tag{3.4}$$

w której wyrażenie w nawiasie kwadratowym jest liczbą obserwacji z klasy i zawartych w odcinku $[x - h(n), x + h(n)]$.

Estymator NN wymaga uporządkowania par w (3.1), co doprowadza do ciągu

$$(\theta_{[1]}, X_{(1)}), (\theta_{[2]}, X_{(2)}), (\theta_{[3]}, X_{(3)}), \ldots, (\theta_{[n]}, X_{(n)}),$$

w którym

$$|X_{(1)} - x| < |X_{(2)} - x| < \cdots < |X_{(n)} - x|.$$

Kryterium porządkowania jest odległość X_i od punktu x, bowiem pozycja, na której para (Y_i, X_i) z ciągu (3.1) znajdzie się w ciągu uporządkowanym zależy od tej odległości, czyli od $|X_i - x|$. Jeśli odległość ta jest j-ta co do wielkości, to pozycją tą jest j, co oznacza, że $(\theta_i, X_i) = (\theta_{[j]}, X_{(j)})$. Z uwagi na (2.3), estymatorem funkcji regresji $r_i(x)$ jest

$$\overline{r}_i(x) = \frac{1}{k(n)} \#\{X_j : \theta_j = i, |X_j - x| \leq k(n)\}$$

Wartość $\#\{X_j : \theta_j = i, |X_j - x| \leq k(n)\}$ jest liczbą tych X_j, które pochodzą z klasy i oraz znajdują się wśród $k(n)$ wszystkich obserwacji najbliższych punktowi x. Estymator ten prowadzi do algorytmu rozpoznawania jak poniżej:

$$\overline{\psi}_n(x) = \arg\max_i \#\{X_j : \theta_j = i, |X_j - x| \leq k(n)\}, \tag{3.5}$$

czyli zalicza x do klasy, która jest najliczniej reprezentowanej wśród $k(n)$ jego najbliższych sąsiadów.

Niekiedy stosuje się algorytmy, w których $k(n) = const$ nie zmienia się wraz z długością ciągu uczącego. W szczególności może być tak, że $k(n) = 1$, co oznacza, że termin "najbliższy sąsiad" rozumiany jest dosłownie. W celu odróżnienia tych algorytmów i wskazania, że chodzi np. o k najbliższych sąsiadów stosuje się niekiedy oznaczenie k-NN.

3.3 Wiele wymiarów

Uogólnienie na przypadek wielowymiarowy może nastąpić przez użycie pojęcia normy wektora x oznaczanej jako $||x||$. Dla wektora o dwóch wymiarach, czyli $x = [x_1, x_2]^T$, przykłady norm, to

$$||x|| = \sqrt{x_1^2 + x_2^2}, \tag{3.6}$$

$$||x|| = |x_1| + |x_2|. \tag{3.7}$$

$$||x|| = \max(x_1, x_2), \tag{3.8}$$

3.3.1 Algorytm jądrowy

Dla wielu wymiarów reguła (3.4) przyjmuje postać

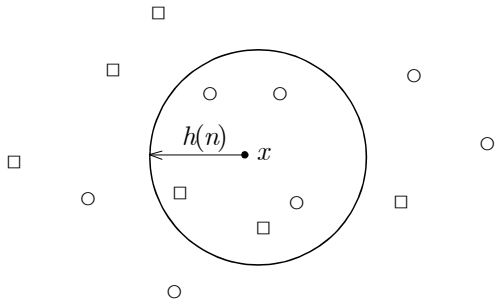
$$\widehat{\psi}_n(x) = \arg\max_i \#\{X_j : \theta_j = i, \|X_j - x\| \leq h(n)\},$$

przy czym wyrażenie w nawiasie kwadratowym jest liczbą obserwacji z klasy i zawartych w sferze o promieniu $h(n)$ i środku w punkcie x. Modyfikacji musi jednak ulec drugi z warunków w (3.2), gdyż należy go zastąpić przez

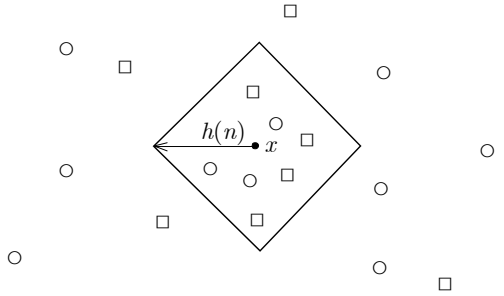
$$\lim_{n\to\infty} nh^d(n) \to \infty,$$

gdzie d jest wymiarem wektora X.

Na rys. 3.1 zilustrowano zasadę działania reguły przy zastosowaniu normy (3.6). Dla z góry ustalonego $h(n)$, wektor x zalicza ona do klasy ◯, ponieważ jest ona liczniej reprezentowana w kole o promieniu $h(n)$ niż klasa □. Na rys. 3.2 normą jest (3.7), a z tych samych powodów wektor x zaliczony jest do klasy □.



Rys. 3.1: Algorytm jądrowy. Ilustracja reguły (3.4); norma (3.6).



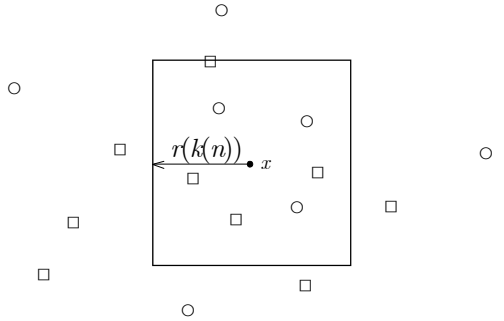
Rys. 3.2: Algorytm jądrowy. Ilustracja reguły (3.4); norma (3.7).

3.3.2 Algorytm NN

Jeśli X jest wektorem, to w algorytmie NN obserwacje porządkuje się według odległości mierzonej normą. Reguła (3.5) przyjmuje wtedy postać

$$\overline{\psi}_n(x) = \arg\max_i \#\{X_j : \theta_j = i, ||X_j - x|| \leq k(n)\}.$$

Ilustracją tego algorytmu jest rys. 3.3. Dla ustalonego z góry $k(n) = 7$, rozpoznaje on x jako pochodzący z klasy $\square$, ponieważ jest on liczniej reprezentowany wśród jego najbliższych 7 sąsiadów niż $\bigcirc$. Zastosowano przy tym normę (3.8).



Rys. 3.3: Algorytm NN. Ilustracja reguły (3.5); $m = 2$, $k(n) = 7$, norma (3.8).

3.4 Podsumowanie

W algorytmie jądrowym bezpośredni wpływ na decyzję mają obserwacje zawarte w odcinku $[x - h(n), x + h(n)]$, gdzie $h(n)$ jest deterministyczne, niezależne od ciągu uczącego. Ich liczba jest losowa a wartość oczekiwana, to

$$nP\{||X - x|| \leq h(n)\} = n \int_{x-h(n)}^{x+h(n)} f(\xi)d\xi \approx nh(n)f(x),$$

która to wielkość narasta do nieskończoności gdy $n \to \infty$, patrz (3.2). W algorytmie NN liczba ta jest zdeterminowana, wynosi $k(n)$ i też narasta do nieskończoności, patrz (3.3).

Zauważmy na koniec, że algorytmy (3.4) i (3.5) można zapisać jedną wspólną formułą

$$\psi_n(x) = \arg\max_i \#\{X_j : \theta_j = i, ||X_j - x|| \in S(x, r(n))\},$$

w której $S(x, r(n))$ jest sferą o promieniu $r(n)$ umieszczoną w punkcie x, czyli zbiorem punktów v takich, że $\|x - v\| \leq r(n)$. Różnica polega na sposobie ustalania tego promienia. W pierwszym $r(n) = h(n)$ jest ustalane z góry i nie zależy od ciągu uczącego. W drugim $r(n) = ||x - X_{(k(n))}||$ zależy, bo jest odległością pomiędzy x a $k(n)$-tą najbliższą mu obserwacją w ciągu uczącym.

4 Problemy parametryczne

Załóżmy teraz, że informacja o rozkładach w poszczególnych klasach jest znacznie większa, że znane są ich postaci funkcyjne. Dla przykładu przyjmijmy, że rozkłady w klasach są normalne, tzn. że

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}.$$

Optymalna reguła zalicza zatem x do klasy

$$\arg\max_i \left[\frac{p_i}{\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}\right].$$

Ponieważ

$$\ln\left(\frac{p_i}{\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}\right) = \ln\frac{p_i}{\sigma_i} - \frac{(x-\mu_i)^2}{2\sigma_i^2}.$$

jest ona równoważna regule zaliczającej ten punkt do klasy

$$\arg\min_i \left[\frac{(x-\mu_i)^2}{2\sigma_i^2} - \ln\frac{p_i}{\sigma_i}\right].$$

Naturalną propozycją empirycznej reguły jest zastąpienie nieznanych wartości ich estymatorami.