NIEPARAMETRYCZNE METODY UCZENIA ROZPOZNAWANIA

Włodzimierz Greblicki

Instytut Cybernetyki Technicznej, Wydział Elektroniki
Politechnika Wrocławska
Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław

STRESZCZENIE

Omówiono zasady tworzenia algorytmów uczenia rozpoznawania na podstawie nieparametrycznych estymatorów funkcji gęstości prawdopodobieństwa i regresji. Podano takie algorytmy i ich własności.

WSTĘP

Ze względu na informację wstępną problemy uczenia rozpoznawania można podzielić na parametryczne i nieparametryczne. W pierwszych, rozkłady prawdopodobieństwa w klasach znane są z dokładnością do skończonej i znanej liczby parametrów. W zadaniach nieparametrycznych, którym poświęcona jest ta praca, brak wspomnianej informacji jest bardziej posunięty, w skrajnym przypadku rozkłady te mogą być nawet całkowicie nieznane. Wydaje się, że w zastosowaniach często spotykamy się z takimi właśnie sytuacjami. Nieparametryczne metody uczenia rozpoznawania rozwijają się już od lat sześćdziesiątych i mają bogatą literaturę, patrz np. monografia [6], a także [2] i [8]. Warto zaznaczyć jednak, że najstarsza, jak się zdaje, praca[1] na ten temat, w której zaproponowano algorytm *k-NN*, powstała już w 1951 roku.

UCZENIE ROZPOZNAWANIA

Przedstawimy teraz zadanie rozpoznawania, a następnie uczenia. Niech (θ, X) będzie parą zmiennych losowych. Pierwsza z nich przyjmuje wartości w zbiorze $M = \{i = 1,2,\ldots,m\}$, elementy którego nazywają się klasami, a druga na prostej R. Rozkład pary (θ, X) opisuje się przez prawdopodobieństwa $p_i = P\{\theta = i\}$ poszczególnych klas oraz warunkowe dystrybuanty $F_1, F_2, \cdots, F_m$, zmiennej θ w tych klasach. Rozpoznawanie (klasyfikacja) polega na estymacji θ na podstawie X.

Regułę rozpoznawania definiuje się jako funkcję, która każdemu punktowi $x \in R$ przyporządkowuje element ze zbioru M. Jakością reguły ψ jest $P\{\psi(X) \neq \theta\}$, czyli prawdopodobieństwo błędnej klasyfikacji. Najlepszą regułę oznaczymy przez ψ^*, a jej

[1] Fix, E., Hodges, J.L., *Discriminatory Analysis. Nonparametric Discrimination: Consistency properties.* Report 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

jakość przez R^*. Jak wiadomo, $\psi^*(x) = \arg\max_{i \in M} P\{\theta = i | X = x\}$. Zauważając, że $P\{\theta = i \mid X = x\} = E\{I_{\{\theta=i\}} \mid X = x\}$, gdzie $I_{\{\theta=i\}} = 1$ dla $\theta = i$ oraz $I_{\{\theta=i\}} = 0$ dla $\theta \neq i$, otrzymujemy bardziej użyteczną postać reguły optymalnej:

$$\psi^*(x) = \arg\max_{i \in M} r_i(x), \tag{1}$$

gdzie $r_i(x) = E\{I_{\{\theta=i\}} \mid X = x\}$ jest funkcją regresji. Jeśli natomiast wszystkie dystrybuanty $F_1, \cdots, F_m$ posiadają gęstości prawdopodobieństwa, co ma miejsce, gdy są one np. różniczkowalne, to oznaczając je jako $f_1, \cdots f_m$, można łatwo zauważyć, że $P\{\theta = i | X = x\} = p_i f_i(x) \Sigma_{i=1}^m p_i f_i(x)$. Wtedy

$$\psi^*(x) = \arg\max_{i \in M} p_i f_i(x). \tag{2}$$

Zadanie uczenia pojawia się, gdy prawdopodobieństwa klas lub rozkłady w klasach nie są znane, a brak tej wstępnej informacji jest rekompensowany przez ciąg uczący $V_n = \{(\theta_1, X_1), \cdots, (\theta_n, X_n)\}$, tzn. ciąg niezależnych obserwacji pary (θ, X), czyli ciąg n prawidłowo rozpoznanych obserwacji $X_1, \cdots, X_n$. Algorytmem uczenia rozpoznawania nazywa się ciąg funkcji $\{\psi_1(x, V_1), \cdots, \psi_n(x, V_n)\}$ o wartościach w M. Warunkowa jakość reguły ψ_n, przy ciągu uczącym V_n, jest równa $P\{\psi_n(X, V_n) \neq \theta \mid V_n\}$ i jest oczywiście losowa. Od poprawnych algorytmów uczenia będziemy oczekiwać, że warunkowe prawdopodobieństwo ich błędu zmierza do R^*, co można wyrazić w formie definicji.

Definicja: Algorytm uczenia $\{\psi_1, \psi_2, \cdots\}$ nazywa się słabo (mocno) asymptotycznie optymalny, jeśli $P\{\psi_n(X, V_n) \neq \theta | V_n\} \xrightarrow{n} R^*$ według prawdopodobieństwa (z p. 1).

Powyższe rozważania prowadzą do dwóch naturalnych sposobów konstrukcji algorytmów uczenia rozpoznawania. Dla pierwszego, który zakłada istnienie gęstości w klasach, punktem wyjściowym jest (2). Na podstawie ciągu uczącego estymuje się zarówno p_i jak i f_i, $i = 1, \cdots, m$. Oznaczając te estymatory odpowiednio jako $\hat{p}_i$ oraz $\hat{f}_i$, algorytm uczenia określa się następująco: dla $n = 1,2, \cdots$,

$$\varphi_n(x, V_n) = \arg\max_{i \in M} \hat{p}_i \hat{f}_i(x). \tag{3}$$

Drugi sposób bierze pod uwagę (1) i jako algorytm uczenia przyjmuje, dla $n = 1,2, \cdots$,

$$\phi_n(x, V_n) = \arg\max_{i \in M} \hat{r}_i(x), \tag{4}$$

gdzie $\hat{r}_i(x)$ jest estymatorem funkcji regresji $r_i(x)$.

ASYMPTOTYCZNA OPTYMALNOŚĆ

Twierdzenia podane poniżej stwierdzają, że zgodne estymatory gęstości i regresji prowadzą do asymptotycznie optymalnych algorytmów uczenia rozpoznawania.

Twierdzenie 1 ([9])***:*** Jeśli, dla $i = 1,2, \cdots, m$ oraz prawie wszystkich (według miary Lebesgue'a) $x \in R$, $\hat{f}_i(x) \xrightarrow{n} f_i(x)$ według prawdopodobieństwa (z p. 1), to algorytm (3) jest słabo (mocno) asymptotycznie optymalny.

Twierdzenie 2 ([28])***:*** Dla dowolnych gęstości w klasach,

$$0 \leq P\{\varphi_n(X, V_n) \neq | V_n\} - R^* \leq \sum_{i=1}^m |\hat{p}_i - p_i| + \sum_{i=1}^m \int_{-\infty}^{\infty} |\hat{f}_i(x) - f_i(x)| dx.$$

W następnym twierdzeniu rozkłady w klasach mogą być zupełnie dowolne.

Twierdzenie 3 ([3], [18])***:*** Dla dowolnych rozkładów w klasach,

$$0 \le P\{\phi_n(X,V_n) \ne |V_n\} - R^* \le \sum_{i=1}^{m} \int_{-\infty}^{\infty} |\hat{r}_i(x) - \hat{r}_i(x)| \mu(dx),$$

gdzie μ jest (dowolną) miarą prawdopodobieństwa rozkładu zmiennej losowej X.

REGUŁY UCZENIA ROZPOZNAWANIA

Stosując różne estymatory funkcji gęstości i regresji, patrz Dodatek, otrzymuje się różne algorytmy uczenia. W metodzie (3), oznaczając przez N_i liczbę obserwacji z klasy i, przyjmujemy przy tym $\hat{p}_i = N_i / n$ jako naturalny estymator nieznanego p_i.

A) Algorytmy jądrowe

Jądrowy estymator gęstości daje algorytm, który każdemu x przyporządkowuje klasę

$$\arg\max_{i \in M} \frac{1}{h(N_i)} \sum_{j=1}^{n} I_{\{\theta_j = i\}} K\left(\frac{x - X_j}{h(N_i)}\right).$$

Stosując estymator regresji każde x zalicza się do klasy

$$\arg\max_{i \in M} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h(n)}\right).$$

Korzystając z Twierdzenia 1 oraz wyników podanych w Dodatku, zauważamy, że pierwszy z powyższych algorytmów jest asymptotycznie optymalny jeśli wszystkie gęstości w klasach są funkcjami ciągłymi. Można wykazać, że ma on tę własność także przy zupełnie dowolnych gęstościach. Twierdzenie 3 doprowadza natomiast do godnego uwagi wniosku, a mianowicie, że drugi z podanych algorytmów jest asymptotycznie optymalny przy całkowicie dowolnych rozkładach w poszczególnych klasach, czyli, że jest on uniwersalnie asymptotycznie optymalny, [3], [11], [21].

B) Algorytmy typu najbliższy sąsiad

Estymator gęstości doprowadza do algorytmu, która zalicza każde x do klasy

$$\arg\max_{i \in M} \frac{k(N_i)}{D_i(x; k(N_i))},$$

gdzie $D_i(x;k)$ jest odległością pomiędzy punktem x a k-tą najbliższą mu obserwacją spośród tych, które pochodzą z klasy i. Estymator regresji daje natomiast algorytm, który klasyfikuje x jako pochodzące z klasy

$$\arg\max_{i \in M} \sum_{j=1}^{n} I_{\{(\theta_j = i) \wedge (j \in J(x;k(n)))\}},$$

gdzie $J(x;k)$ zdefiniowano w Dodatku. Jest to znany algorytm $k(n)$-NN.

Twierdzenie 1 i wyniki podane w Dodatku dotyczące zbieżności estymatora gęstości prowadzą do wniosku, że pierwszy algorytm jest asymptotycznie optymalny, jeśli wszystkie gęstości w klasach są np. funkcjami ciągłymi. Drugi jest natomiast uniwersalnie asymptotycznie optymalny, [4].

C) Algorytmy ortogonalne

Ortogonalny estymator gęstości wykorzystujący szereg Hermite'a, patrz Dodatek, daje algorytm, który zalicza każde x jako pochodzące z klasy

$$\arg\max_{i \in M} \sum_{j=1}^{n} \sum_{k=0}^{N(N_i)} I_{\{\theta_j = i\}} h_k(X_j) h_k(x).$$

Drugi sposób prowadzi do algorytmu, który przypisuje x do klasy

$$\arg\max_{i\in M} \sum_{j=1}^{n} \sum_{k=0}^{N(n)} I_{\{\theta_j=i\}} h_k(X_j) h_k(x).$$

Asymptotyczna optymalność tych algorytmów wynika z podanych wcześniej twierdzeń i własności estymatorów podanych w Dodatku, [10].

ZAKOŃCZENIE

Oprócz asymptotycznej optymalności, ważny jest także problem szybkości zbieżności algorytmów uczenia. Zauważmy, że Twierdzenia 2 i 3 wiążą dokładność estymacji gęstości i regresji z ich jakością. Jeśli gęstości w klasach są np. dwukrotnie różniczkowalne, to, $P\{\psi_n(X,V_n) \neq \theta\} - R^* = O(n^{-2/5})$ gdzie $\{\psi_n\}$ jest dowolnym z omówionych algorytmów. W porównaniu z $n^{-1/2}$, tzn. z szybkością typową dla uczenia parametrycznego, jest to zachęcający rezultat. Znane są także inne sposoby nieparametrycznej estymacji gęstości i regresji, [5], [20], [25], które prowadzą do kolejnych algorytmów. Badane są także estymatory rekurencyjne, [14].

DODATEK

Odnośnie nieparametrycznej estymacji gęstości i regresji odsyłamy do monografii [20], [25] i [25]. Tutaj omówimy krótko trzy metody.

Estymacja gęstości prawdopodobieństwa

Skalarna zmienna losowa X posiada gęstość prawdopodobieństwa f, którą estymuje się na podstawie niezależnych obserwacji $X_1, X_2, \cdots, X_n$.

A) Estymator jądrowy

Parzen zaproponował następujący estymator:

$$\hat{f}(x) = \frac{1}{nh(n)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(n)}\right),$$

gdzie K jest tzw. jądrem, a $\{h(n)\}$ ciągiem liczbowym, [24]. Jeśli K jest ograniczone, $\int_{-\infty}^{\infty} K(x)dx = 1$, $\lim_{|x|\to\infty} xK(x) = 0$, oraz $h(n) \xrightarrow{n} 0$, i $nh(n) \xrightarrow{n} \infty$, to $\hat{f}(x) \xrightarrow{n} f(x)$ według prawdopodobieństwa w każdym punkcie x, w którym gęstość f jest ciągła. Jeśli ponadto $nh(n)/\log n \xrightarrow{n} \infty$, to zachodzi zbieżność z p. 1, [7]. Jako jądro można wybrać np. $1/\pi(1+x^2)$, $(1/2)\exp(-|x|)$, $(1/\sqrt{\pi})\exp(-x^2)$, lub jądro prostokątne równe 1/2 i 0 odpowiednio dla $|x|<1$ i $|x|\geq 1$. Jako ciąg liczbowy można natomiast zastosować $h(n) = cn^{-\alpha}$, $c > 0$, $0 < \alpha < 1$.

B) Estymator typu najbliższy sąsiad

Loftsgaarden i Quesenberry podali następujący estymator:

$$\hat{f}(x) = \frac{k(n)}{2nD(x;k(n))},$$

gdzie $\{k(n)\}$ jest ciągiem liczb naturalnych, a $D(x;k(n))$ odległością pomiędzy punktem x a $k(n)$-tą najbliższą mu obserwacją, [22]. Jeśli $k(n) \xrightarrow{n} \infty$, $k(n)/n \xrightarrow{n} 0$, to $\hat{f}(x) \xrightarrow{n} f(x)$ według prawdopodobieństwa w każdym punkcie $x \in R$, w którym gęstość f jest ciągła. Jeśli ponadto $k(n)\log n/n \xrightarrow{n} 0$, to ma miejsce zbieżność z p. 1, [6]. Jako ciąg liczbowy można przyjąć $k(n) = [cn^{\beta}]$, przy czym $0 < c, 0 < \beta < 1$, gdzie $[\beta]$ oznacza część całkowitą liczby β.

C) Estymator ortogonalny

Estymację ortogonalną zaproponował Čencov, [1]. Jak wiadomo, funkcje Hermite'a $\{h_k; k = 0,1,2,\cdots\}$, gdzie $h_k(x) = (2^k k!\pi^{1/2})^{-1/2} H_k(x)\exp(-x^2/2)$, przy czym $H_k(x) = \exp(x^2/2)(d^k/dx^k)\exp(-x^2/2)$, tworzą zupełny system ortonormalny na prostej R. Estymator wykorzystujący ten szereg ma następującą postać:

$$\bar{f}(x) = \sum_{k=0}^{N(n)} \bar{a}_k(x) h_k(x),$$

przy czym $\bar{a}_k = n^{-1}\Sigma_{i=1}^{n} h_k(X_i)$. Jeśli $f \in L^p(R)$, $p = 2$, $N(n) \xrightarrow{n} \infty$ oraz $N^{1/2}(n)/n \xrightarrow{n} 0$, to $\int_{-\infty}^{\infty}(\bar{f}(x) - f(x))^2 dx \xrightarrow{n} 0$ według prawdopodobieństwa. Jeśli ponadto $N^{1/2}(n)\log n/n \xrightarrow{n} 0$, to zachodzi zbieżność z p. 1, [10], [12]. Ciągiem liczbowym może być np. $N(n) = [cn^{\gamma}]$, $0 < c$, $0 < \gamma < 2$. Ma miejsce także zbieżność punktowa, i to nawet dla $p > 1$. Inne układy funkcji ortogonalnych, np. trygonometryczny, [16], Legendre'a, Laguerre'a, czy też Haara, prowadzą do kolejnych estymatorów. Interesującą modyfikacją jest tzw. uśrednianie Cesàro, [17].

ESTYMACJA FUNKCJI REGRESJI

Niech (Y,X) będzie parą skalarnych zmiennych losowych takich, że $E|Y| < \infty$. Regresję $r(x) = E\{Y \mid X = x\}$ estymuje się na podstawie niezależnych obserwacji $(Y_1,X_1),\cdots,(Y_n,X_n)$. Przez μ oznaczymy miarę prawdopodobieństwa zmiennej losowej X. Może być ona zupełnie dowolna, a zatem w szczególności może nie posiadać gęstości.

A) Estymator jądrowy

Jądrowy estymator regresji ma następującą postać, [23], [29]:

$$\hat{r}(x) = \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h(n)}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(n)}\right).$$

Jeśli jądro K i ciąg $\{h(n)\}$ spełniają warunki podane przy estymatorze gęstości, to $\hat{r}(x) \xrightarrow{n} r(x)$ według prawdopodobieństwa w prawie każdym punkcie (według miary μ) $x \in R$, [3], [11], [27]. Podobnie można zapewnić zbieżność z p. 1, [3], [11], [27]. Wersje rekurencyjne badano w pracach [13] i [19].

B) Estymator typu najbliższy sąsiad

Ustalmy punkt $x \in R$ oraz liczbę naturalną k. Wśród obserwacji $X_1,\cdots,X_n$ jest k najbliższych temu punktowi. Oznaczmy przez $J(x;k)$ zbiór ich indeksów. Niech

$$\hat{r}(x) = \frac{1}{k(n)} \sum_{i=1}^{n} Y_i I_{\{i \in J(x;k(n))\}}.$$

Jeśli $k(n) \xrightarrow{n} \infty$ i $k(n)/n \xrightarrow{n} 0$, to $\hat{r}(x) \xrightarrow{n} r(x)$ według prawdopodobieństwa w prawie każdym punkcie (według miary μ) $x \in R$. Jeśli ponadto $k(n)\log n/n \xrightarrow{n} 0$, to ma miejsce zbieżność z p. 1, [4].

C) Estymator ortogonalny

Zakładamy teraz, że μ posiada gęstość f, a estymator ma następującą postać, [13]:

$$\hat{r}(x) = \sum_{k=0}^{N(n)} \bar{a}_k h_k(x) \Big/ \sum_{k=0}^{N(n)} \bar{b}_k h_k(x),$$

gdzie $\bar{a}_k = n^{-1}\Sigma_{i=1}^{n} Y_i h_k(X_i)$, $\bar{b}_k = n^{-1}\Sigma_{i=1}^{n} h_k(X_i)$. Jeśli $f \in L^2(R)$, $f(.)r(.) \in L^2(R)$ oraz $\{N(n)\}$ spełnia warunki jak przy estymatorze gęstości, to $\hat{r}(x) \xrightarrow{n} r(x)$ według

prawdopodobieństwa dla prawie wszystkich (według miary Lebesgue'a) $x \in R$, takich, że $0 < f(x)$. Wybierając odpowiednio ciąg liczbowy, można zapewnić zbieżność z p. 1.

BIBLIOGRAFIA

[1] Čencov, N.N., Evaluation of an unknown distribution density from observations, *Soviet Mathematics*, vol. 3, 1559-1562, 1962.

[2] Devijver, P.A., J. Kittler, J., *Pattern Recognition: A Statistical Approach*, Prentice Hall, Englewood Cliffs, 1982.

[3] Devroye, L., Distribution-free consistency results in nonparametric discrimination and regression function estimates, *Annals of Statistics*, vol. 8, 231-239, 1980.

[4] Devroye, L., Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates, *Zeitschrift für Wahrscheinlichkıstheorie und verwandte Gebiete* , vol. 61, 467-481, 1982.

[5] Devroye, L., Györfi, L., *Nonparametric Density Estimation: The L_1 View*, Wiley, New York, 1985.

[6] Devroye, L, Györfi, L., Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[7] Devroye, L.P., Wagner, T.J., *Nonparametric Discrimination and Density Estimation*, Technical Report 183, Electronics Research Centre, University of Texas, Austin, Texas, 1976.

[8] Duda R.O., Hart, P.E., *Pattern Recognition and Scene Analysis*, Wiley, New York, 1973.

[9] Greblicki, W., Asymptotically optimal pattern recognition procedures with density estimates, *IEEE Transactions on Information Theory*, vol. IT-24, 250-251, 1978.

[10] Greblicki, W., Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities, *IEEE Transactions on Information Theory*, vol. IT- 27, 364-366, 1981.

[11] Greblicki, W., Krzyżak, A., Pawlak, M., Distribution-free consistency of kernel regression estimate, *Annals of Statistics*, vol. 12, 1570-1575, 1984.

[12] Greblicki, W., Pawlak, M., Hermite series estimates of a probability density and its derivatives, *Journal of Multivariate Analysis*, vol. 15, 174-182, 1984.

[13] Greblicki, W., Pawlak, M., Fourier and Hermite series estimates of regression functions, *Annals of Institute of Statistical Mathematics*, vol. 37, Part A, 443-454, 1985.

[14] Greblicki, W., Pawlak, M., Necessary and sufficient conditions for Bayes risk consistency of a recursive kernel classification rule, *IEEE Transactions on Information Theory*, vol. IT-33, 408-412, 1987.

[15] Greblicki, W., Pawlak, M., Necessary and sufficient consistency conditions for a recursive kernel regression estimate, *Journal of Multivariate Analysis*, vol. 23, 67-76, 1987.

[16] Greblicki, W., Pawlak, M., A classification procedure using the multiple Fourier series, *Information Sciences*, vol. 26, 115-126, 1982.

[17] Greblicki, W., Rutkowski, L., Density-free Bayes risk consistency of nonparametric pattern recognition procedures, *Proceedings of the IEEE*, vol. 69, 482-483, 1981.

[18] Györfi, L., Recent results on nonparametric regression estimate and multiple classification, *Problems of Control and Information Theory*, vol. 10, 13-52, 1981.

[19] Györfi, L., Walk, H., On the strong universal consistency of a recursive regression estimate by Pàl Rèvèsz, *Statistics and Probability Letters*, vol. 31, 177-183, 1997.

[20] Härdle, W., *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.

[21] Krzyżak, A., Pawlak, M., Distribution-free consistency of a nonparametric kernel regression estimate and classification, *IEEE Transactions on Information Theory*, vol. IT-30, 78-81, 1984.

[22] Loftsgaarden, D.O., Quesenberry, C.P., A nonparametric estimation of a multivariate density function, *Annals of Mathematical Statistics*, vol. 36, 1049-1051, 1965.

[23] Nadaraya, E.A., On estimating regression, *Theory of Probability and Its Applications*, vol. 9, 141-142, 1964.

[24] Parzen, E., On the estimation of a probability density function and the mode, *Annals of Mathematical Statistics*, vol. 33, 1065-1076, 1962.

[25] Prakasa Rao, B.L.S., *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.

[26] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London 1986.

[27] Stone, C., Consistent nonparametric regression, *Annals of Statistics*, vol. 8, 1348-1360, 1977.
[28] Van Ryzin, J., Bayes risk consistency of classification procedures using density estimation, *Sankhyā*, Parts 2&3, vol. 28, 261-270, 1966.
[29] Watson, G.S., Smooth regression analysis, *Sankhyā*, Ser. A, vol. 26, 359-372, 1964.