Twierdzenie "No free lunch"

Jeśli uśrednimy jakość klasyfikacji po wszystkich możliwych problemach klasyfikacyjnych to wszystkie algorytmy dają ten sam błąd klasyfikacji na zbiorze testowym. To twierdzenie stanowi, że nie ma uniwersalnie najlepszego algorytmu. Zatem celem nie jest poszukiwanie takiego algorytmu, a znalezienie algorytmu optymalnego dla konkretnych danych.

Dowód.

Rozważmy problem dwuklasowy. Jeśli Algorytm A wygrywa to zamieniamy etykiety i będzie wygrywał Algorytm B. □

Wolpert, D. (1996). *The Lack of a Priori Distinctions between Learning Algorithms*. Neural Computation 8(7): 341–1390.

Wolpert, D. & Macready, G. (1997). *No Free Lunch Theorems for Optimization*. IEEE Transactions on Evolutionary Computation 1(1):67–82.

Kwadratowa analiza dyskryminacyjna

Najprostszym podejściem do zagadnienia klasyfikacji jest wykorzystanie estymacji funkcji gęstości i przyjęcie modelu parametrycznego dla gęstości, tj. przyjęcie założenia, że znana jest postać gęstości z wyjątkiem tkwiących w niej parametrów. Załóżmy, że $\mathcal{Y} = \{1, 0\}$ i $f_1(\boldsymbol{x}) = f(\boldsymbol{x}|Y = 1)$ oraz $f_0(\boldsymbol{x}) = f(\boldsymbol{x}|Y = 0)$ są gęstościami p-wymiarowego rozkładu normalnego. Wówczas $\boldsymbol{X}|Y = 1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ i $\boldsymbol{X}|Y = 0 \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Kwadratowa analiza dyskryminacyjna

Twierdzenie

Jeżeli $\boldsymbol{X}|Y=1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *i* $\boldsymbol{X}|Y=0 \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, *to klasyfikator bayesowski ma postać*

$$d_B(\boldsymbol{x}) = \begin{cases} 1, & \textit{jeżeli } r_1^2 < r_0^2 + 2\ln\left(\frac{\pi_1}{\pi_0}\right) + \ln\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right), \\ 0, & \textit{poza tym.} \end{cases}$$

gdzie

$$r_i^2 = (\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i),\ i=1,0$$

jest kwadratem odległości Mahalanobisa.

Kwadratowa analiza dyskryminacyjna

Powyższy klasyfikator najczęściej zapisujemy w następującej postaci:

$$d_B(\boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x}),$$

gdzie

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln\pi_k.$$

Powyższa funkcja jest funkcją kwadratową argumentu $\boldsymbol{x}$ i jest nazywana kwadratową funkcją klasyfikującą grupy k. Procedura klasyfikacji oparta na tej funkcji nosi nazwę kwadratowej analizy dyskryminacyjnej (QDA).

Kwadratowa analiza dyskryminacyjna

Parametry π_1, π_0, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_0$ są w praktyce nieznane, zastępujemy je ich estymatorami z próby uczącej $\mathcal{L}_n$ postaci:

$$\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\pi}_0 = 1 - \hat{\pi}_1,$$

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{Y_i=1} \boldsymbol{X}_i, \quad \hat{\boldsymbol{\mu}}_0 = \frac{1}{n_0}\sum_{Y_i=0} \boldsymbol{X}_i,$$

$$\hat{\boldsymbol{\Sigma}}_1 = \boldsymbol{S}_1 = \frac{1}{n_1 - 1}\sum_{Y_i=1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_1)',$$

$$\hat{\boldsymbol{\Sigma}}_0 = \boldsymbol{S}_0 = \frac{1}{n_0 - 1}\sum_{Y_i=0} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)',$$

gdzie $n_1 = \sum_{i=1}^{n} Y_i$, $n_0 = \sum_{i=1}^{n}(1 - Y_i)$.

Liniowa analiza dyskryminacyjna

Procedura ta znacznie się uprości, jeżeli dodatkowo założymy równość macierzy kowariancji w dwóch grupach $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$. W tym przypadku klasyfikator bayesowski ma postać

$$d_B(\boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x}),$$

gdzie

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k, \;\; k = 1, 0.$$

Funkcja ta jest funkcją liniową $\boldsymbol{x}$ i nosi nazwę liniowej funkcji klasyfikującej grupy k.

Liniowa analiza dyskryminacyjna

Powierzchnia rozdzielająca

$$\{\boldsymbol{x}:\ \delta_1(\boldsymbol{x}) = \delta_0(\boldsymbol{x})\} = \{\boldsymbol{x}:\ \delta_{10}(\boldsymbol{x}) = \delta_1(\boldsymbol{x}) - \delta_0(\boldsymbol{x}) = 0\}$$

jest w tym przypadku hiperpłaszczyzną. Funkcja $\delta_{10}(\boldsymbol{x}) = \delta_1(\boldsymbol{x}) - \delta_0(\boldsymbol{x})$ nazywa się liniową funkcją dyskryminacyjną. Ma ona postać:

$$\delta_{10}(\boldsymbol{x}) = (\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0))'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \ln\left(\frac{\pi_1}{\pi_0}\right).$$

Obserwację $\boldsymbol{x}$ klasyfikujemy do grupy 1 wówczas, gdy $\delta_{10}(\boldsymbol{x}) > 0$. W przeciwnym razie klasyfikujemy do grupy 0. Procedura klasyfikacji bazująca na liniowej funkcji klasyfikującej lub liniowej funkcji dyskryminacyjnej nosi nazwę liniowej analizy dyskryminacyjnej (LDA).

Liniowa analiza dyskryminacyjna

Nieznane parametry estymujemy z próby uczącej tak jak poprzednio, przy czym:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n_1 + n_0 - 2}[(n_1 - 1)\boldsymbol{S}_1 + (n_0 - 1)\boldsymbol{S}_0].$$

Przypadek większej liczby grup

Twierdzenie

Załóżmy, że $Y \in \{1, \ldots, K\}$. *Jeżeli* $f_k(\boldsymbol{x}) = f(\boldsymbol{x}|Y = k)$ *jest gęstością p-wymiarowego rozkładu normalnego (gaussowskiego)* $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, *to klasyfikator bayesowski ma postać*

$$d_B(\boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x}),$$

gdzie

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln \pi_k.$$

Jeżeli ponadto wszystkie macierze kowariancji są sobie równe i równe macierzy $\boldsymbol{\Sigma}$, *to*

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k.$$

Przypadek większej liczby grup

Powierzchnia rozdzielająca grupy i oraz j ma postać

$$\{\boldsymbol{x}:\ \delta_i(\boldsymbol{x}) = \delta_j(\boldsymbol{x})\} = \{\boldsymbol{x}:\ \delta_{ij}(\boldsymbol{x}) = \delta_i(\boldsymbol{x}) - \delta_j(\boldsymbol{x}) = 0\},$$

gdzie

$$\begin{aligned}\delta_{ij}(\boldsymbol{x}) = & \frac{1}{2}\boldsymbol{x}'(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1})\boldsymbol{x} \\ & + \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\ & + \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}\right) + \ln\left(\frac{\pi_i}{\pi_j}\right),\ i,j = 1,2,\ldots,K,\ j \neq i\end{aligned}$$

w przypadku funkcji kwadratowej

Przypadek większej liczby grup

oraz

$$\delta_{ij}(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) - \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \ln\left(\frac{\pi_i}{\pi_j}\right), \ i,j = 1,2,\ldots,K, \ j \neq i.$$

w przypadku funkcji liniowej.

Przypadek większej liczby grup

Występujące w powyższych wzorach parametry nie są zazwyczaj znane i w praktyce należy zastąpić je ich estymatorami z próby uczącej. Jeżeli próba ucząca zawiera n_i obserwacji z i-tej grupy, $n_1 + n_2 + \cdots + n_K = n$ oraz $\boldsymbol{X}_{ij}$ jest j-tą obserwacją z i-tej grupy, to estymatory nieznanych parametrów są równe:

$$\hat{\pi}_k = \frac{n_k}{n},$$

$$\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{X}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \boldsymbol{X}_{kj},$$

$$\hat{\boldsymbol{\Sigma}}_k = \boldsymbol{S}_k = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (\boldsymbol{X}_{kj} - \bar{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \bar{\boldsymbol{X}}_k)',$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\boldsymbol{X}_{kj} - \bar{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \bar{\boldsymbol{X}}_k)'.$$

Zbiór danych `iris`

The `iris flower` data set is a multivariate data set introduced by R. Fisher (1936). The data set consists of 50 samples from each of three species of iris (setosa, virginica and versicolor). Four features were measured: the length and the width of the sepals and petals, in centimeters. Based on it, Fisher developed a linear discriminant model to distinguish the species from each other.

Sir Ronald Aylmer Fisher (1890-1962)

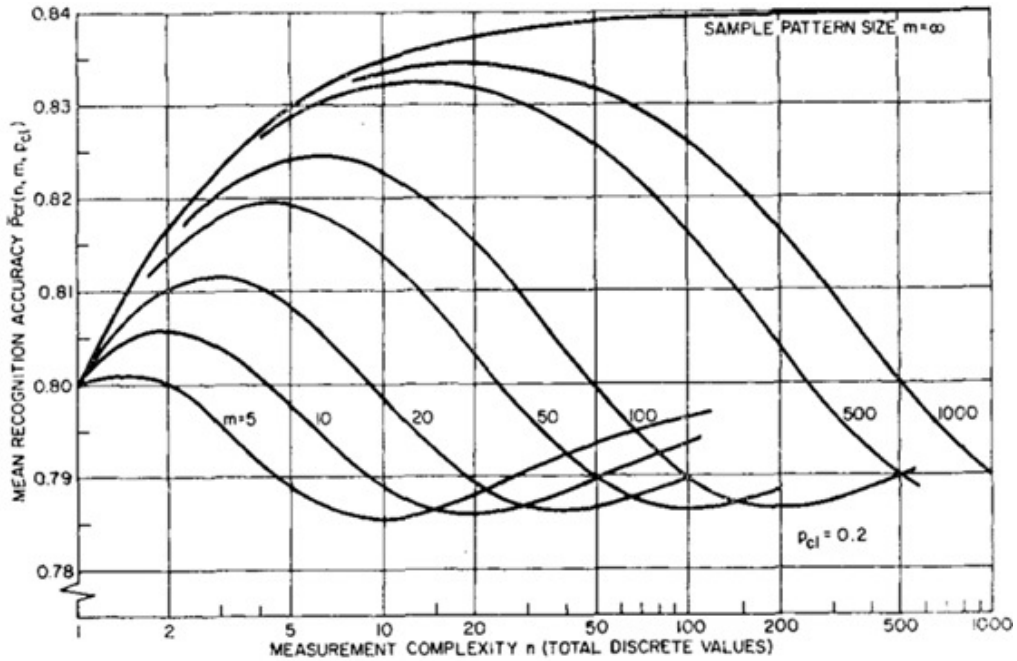
Zbiór danych `iris`

Iris Versicolor **Iris Setosa** **Iris Virginica**

Fisher R. A. (1936). *The use of multiple measurements in taxonomic problems*. Annals of Eugenics 7(2):179–188.

Przekleństwo wymiarowości

Dla ustalonej wielkości próby uczącej jakość klasyfikatora (regresji) na początku wzrasta wraz ze zwiększaniem liczby cech, a następnie spada. Zjawisko to nazywane jest paradoksem Hughes'a.



Hughes, G.F. (1968). *On the mean accuracy of statistical pattern recognizers*. IEEE Transactions on Information Theory 14(1):55–63.