1 Algorytm iteracyjny regresji liniowej

Niech X i Y będą równolicznymi krotkami obserwacji (po n obserwacji). Celem algorytmu jest dobranie „najlepszych" współczynników funkcji afinicznej, która pozwala obliczać Y na podstawie X. Zatem jeśli oznaczymy:

$$f(X) = aX + b, \tag{1}$$

to chcemy aby wartości $f(X)$ dla poszukiwanych a i b były „możliwie zbliżone" do Y.

Wzajemne podobieństwo dwóch sekwencji liczb można definiować na wiele sposobów, jednak w proponowanej metodzie zostanie zastosowana metoda najmniejszych kwadratów, lub minimalizacja średniego błędu kwadratowego. Najprościej to zapisać jako:

$$\overline{(f(X) - Y)^2} \to \min, \tag{2}$$

Przejdźmy teraz do bardziej „życiowego"X przykładu. Zapiszmy na początek funkcję (1) w nieco zmienionej postaci:

$$f(x, p) = \sum_{i=0}^{N-1} p_i x_i = p_0 x_0 + p_1 x_1, \tag{3}$$

gdzie N = 2, p jest listą/tablicą parametrów, przyjmującą w tym konkretnym wypadku wartość $[b, a]$, a wartości x pochodzą z poszczególnych elementów X uzupełnionych o 1. Przykładowo, jeżeli pierwszy element krotki X miał wartość 3, to lista x będzie miała postać $[1, 3]$.

Zdefiniujmy również funkcje średniego błędu kwadratowego w następujący sposób:

$$Q(p) = \frac{1}{2N} \sum_{i=0}^{N-1} (f(x^i, p) - y^i)^2, \tag{4}$$

gdzie N - ilość elementów w krotce X.

Na początku działania algorytmu losujemy wartości p, najlepiej z przedziału $[0, 1]$. Następnie, iteracyjnie zmieniamy każdą z wartości p zgodnie ze wzorem:

$$p_i(t + 1) = p_i(t) - \alpha \cdot \frac{\partial Q}{\partial p_i}, \tag{5}$$

W celu lepszego zrozumienia mechanizmu zmiany p_i rozpiszmy na początek wzór (5) dla p_0:

$$\frac{\partial Q}{\partial p_0} = \frac{\partial}{\partial p_0} \frac{1}{2N} \sum_{i=0}^{N-1} (f(x^j, p) - y^j)^2 \tag{6}$$

$$= \frac{1}{2N} \sum_{j=0}^{N-1} \frac{\partial}{\partial p_0} (f(x^j, p) - y^j)^2$$

$$= \frac{1}{2N} \sum_{j=0}^{N-1} 2 \cdot (f(x^j, p) - y^j) \cdot \frac{\partial}{\partial p_0}(p_0 x_0^j + p_1 x_1^j)$$

$$= \frac{1}{N} \sum_{j=0}^{N-1} (f(x^j, p) - y^j) \cdot x_0^j$$

Następnie, przejdźmy do wzoru na p_1:

$$\frac{\partial Q}{\partial p_1} = \frac{\partial}{\partial p_1} \frac{1}{2N} \sum_{j=0}^{N-1} (f(x^j, p) - y^j)^2 \tag{7}$$

$$= \frac{1}{2N} \sum_{j=0}^{N-1} \frac{\partial}{\partial p_1} (f(x^j, p) - y^j)^2$$

$$= \frac{1}{2N} \sum_{j=0}^{N-1} 2 \cdot (f(x^j, p) - y^j) \cdot \frac{\partial}{\partial p_1}(p_0 x_0^j + p_1 x_1^j)$$

$$= \frac{1}{N} \sum_{j=0}^{N-1} (f(x^j, p) - y^j) \cdot x_1^j$$

Po analizie wzorów (6) i (7) łatwo zauważyć, że w tym przypadku, ogólny wzór na p_i przyjmuje postać:

$$p_i(t+1) = p_i(t) - \alpha \cdot \left(\frac{1}{N} \sum_{j=0}^{N-1} (f(x^j, p) - y^j) \cdot x_i^j \right), \tag{8}$$

gdzie α to krok pojedynczej zmiany $\alpha \in (0, 1)$.

Powyższe obliczenia będą również poprawne dla dowolnej liczby parametrów p. Rozważmy więc przybliżenie bardziej złożonej funkcji:

$$f(x, p) = a + bx^2 + c\sin(x), \tag{9}$$

przy pomocy funkcji (10). Zapiszmy zatem ją jako:

$$f(x, p) = p_0 x_0 + p_1 x_1 + p_2 x_2, \tag{10}$$

gdzie $x_0 = 1$, $x_1 = x^2$, $x_2 = \sin(x)$, a p - jest listą poszukiwanych parametrów. Modyfikacja p w tym przypadku również będzie przebiegać zgodnie ze wzorem (8).

2 Ocena przydatności wyników

Jeśli $Z = f(x, p)$ jest sekwencją proponowanych wyników (czyli Z jest krotką wypełnioną obserwacjami ze znalezionej funkcji), to krotkę błędów można oznaczyć jako:

$$Err = Y - Z. \tag{11}$$

Dla optymalnego modelu na pewno $Var\ Err \leq Var\ Y$ – równość występowałaby dla modelu stałego $Z = 0$. Iloraz

$$FUV = \frac{Var\ Err}{Var\ Y} \tag{12}$$

określa się zatem jako „współczynnik niewyjaśnionej wariancji". Jest to liczba z przedziału $[0; 1]$ – im mniejsza tym lepiej.

Jeszcze popularniejszym w literaturze jest „współczynnik determinacji" lub po prostu „R kwadrat", czyli

$$R^2 = 1 - FUV \tag{13}$$

W przypadku R^2 interpretacja jest nieco podobna, co w przypadku współczynnika korelacji – 0 oznacza zupełny brak zależności, a 1 oznacza zależność bezbłędną (bez uwzględniania takich szczegółów jak monotoniczność zależności liniowej). Wartość tę często podaje się procentowo, jako „odsetek wyjaśnionej wariancji". W zależności od dziedziny zastosowań, wyboru modelu i wyboru źródła bibliograficznego, różne wartości mogą być interpretowane w różny sposób. Wartości poniżej 70% sugerują, że zależność występuje, ale model może nie być najlepszym z możliwych dla wybranych danych. Wartości powyżej 85% często są oceniane jako sukces w dopasowywaniu modelu do danych z pewnym szumem.

Satysfakcjonująca wartość R^2 może zostać wskazana jako warunek konieczny uznania regresji za udaną, ale nie powinien to być warunek wystarczający. W tym celu warto zbadać rozkład wartości Err. W przypadku regresji liniowej zawsze $Err = 0$, ale to jeszcze nawet nie znaczy, że rozkład jest symetryczny. Jeśli błędy nie przystają do rozkładu normalnego, to trudno je nazwać przypadkowym szumem, a wskazane jest poszukiwanie modelu, który wyjaśniłby dane lepiej. Ocena, czy dane „przystają do rozkładu normalnego" może obejmować narysowanie histogramu i ocenę „na oko", ale powinna w sobie zawierać również formalne podejście takie jak test Shapiro-Wilka albo test zgodności $\mathcal{X}^2$ Pearsona.