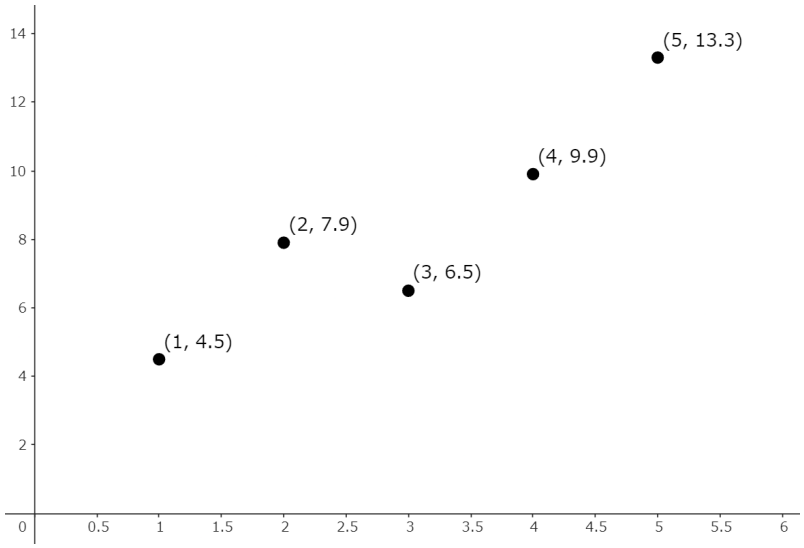
Regresja liniowa - skrypt

Zachariasz Jażdżewski

25 marca 2024

1 Czym jest dopasowanie prostej

Załóżmy, że jesteśmy badaczami zajmującymi się analizą wzrostu roślin w różnych warunkach środowiskowych. Przez kilka miesięcy zbieraliśmy dane dotyczące wysokości roślin oraz ilości dostarczanego im światła słonecznego w różnych dniach.



Światło słoneczne (kiloluks)	1	2	3	4	5
Wzrost roślin (cm)	4.5	7.9	6.5	9.9	13.3

Po przejrzeniu tych danych i rozrysowaniu ich na wykresie zauważyliśmy pewne trendy, które sugerują, że istnieje związek między ilością światła słonecznego a wzrostem roślin. W związku z tym zastanawiamy się, w jaki sposób możemy stworzyć model, który pozwoliłby nam przewidzieć wzrost roślin na podstawie dostarczanego im światła słonecznego.

Metodę znajdowania takiego modelu nazywa się regresją liniową. Regresja liniowa polega na badaniu zależności między zmiennymi; w najprostszej formie jest to technika modelowania zależności między tylko dwiema zmiennymi (jak na przykład nasze dane rozłożone w czasie). My zajmiemy się prostym przykładem regresji liniowej, jakim jest dopasowanie prostej do zestawu danych, wykorzystując metodę najmniejszych kwadratów.

2 Metoda najmniejszych kwadratów i najlepsze rozwiązanie układu równań

Przedstawimy twierdzenie, które bardzo przyda się przy omawianiu zagadnienia regresji liniowej.

Twierdzenie. Zbiór najlepszych rozwiązań układu równań $Ax = b$ jest identyczny ze zbiorem rozwiązań układu $A^TAx = A^Tb$.

$$Ax = b \iff A^TAx = A^Tb$$

Komentarz. Taki sposób wyznaczania najlepszego rozwiązania układu równań liniowych nazywamy *metodą najmniejszych kwadratów*, a sam układ równań $A^TAx = A^Tb$ nazywamy *normalnym*

3 Dopasowanie prostej

Niech $(x_1, y_2), (x_2, y_2), \ldots, (x_n, y_n)$ będą punktami z płaszczyzny $\mathbb{R}^2$ takimi, że nie wszystkie liczby $x_1, x_2, \ldots, x_n$ są równe. Wyznaczymy prostą $y = ax + b$, która najlepiej pasuje do danych punktów. Jej współczynniki a i b dobieramy w taki sposób, aby suma

$$\sum_{i=1}^{n}(ax_i + b - y_i)^2$$

w której $|ax_i + b - y_i|$ jest odległością pomiędzy punktami (x_i, y_i) i $(x_i, ax_i + b)$ była najmniejsza z możliwych.

Suma ta jest najmniejsza wtedy i tylko wtedy, gdy (a, b) jest *najlepszym rozwiązaniem* układu równań liniowych:

$$\begin{cases} ax_1 + b = y_1 \\ ax_2 + b = y_2 \\ \quad\vdots \\ ax_n + b = y_n \end{cases}$$

Taki układ możemy zapisać w postaci macierzowej jako $Ax = b$, gdzie

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Korzystając z poznanego twierdzenia, najlepsze rozwiązanie tego układu wyznaczamy za pomocą *normalnego układu równań* $A^TAx = A^Tb$:

$$A^TA = \begin{bmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}$$

$$A^Tb = \begin{bmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$$

Możemy zauważyć, że wyznacznik macierzy A^TA jest równy

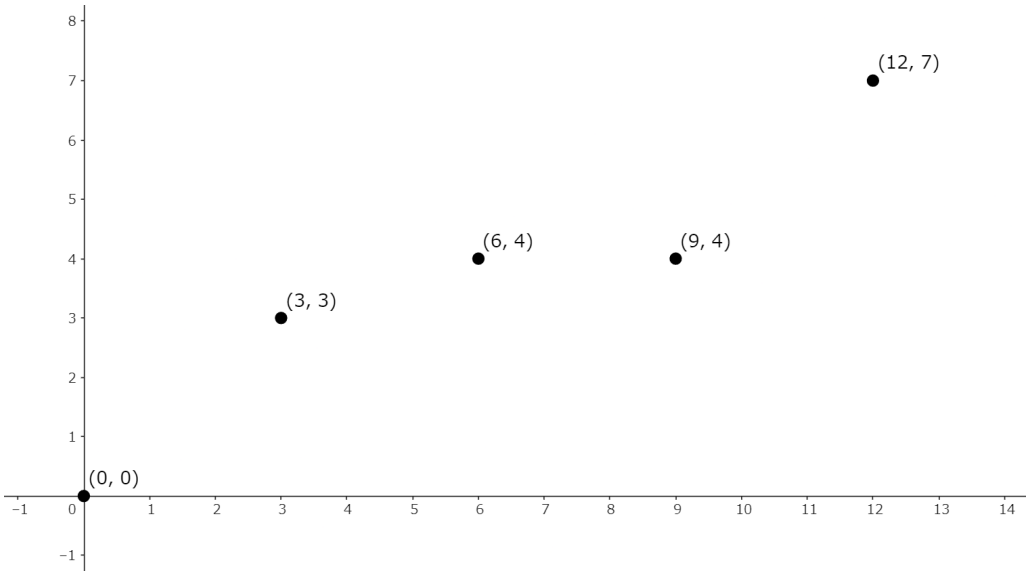
$$\sum_{1 \le i < j \le n} (x_i - x_j)^2$$

Zatem ostateczna suma jest niezerowa tylko wtedy, gdy nie wszystkie liczby $x_1, \dots, x_n$ są równe. Stąd też układ $A^TAx = A^Tb$ ma dokładnie jedno rozwiązanie (a co za tym idzie układ $Ax = b$ ma dokładnie jedno *najlepsze* rozwiązanie), jeśli tylko nie wszystkie liczby $x_1, \dots, x_n$ są równe.

4 Jak to się liczy

4.1 Prosty przykład

Wyznaczymy najlepszą liniową zależność $y = ax + b$ między współrzędnymi x_i i y_i punktów $(0, 0), (3, 3), (6, 4), (9, 4)$ i $(12, 7)$.



Szukane współczynniki a i b są najlepszym rozwiązaniem układu równań liniowych $Ax = b$, gdzie

$$A = \begin{bmatrix} 0 & 1 \\ 3 & 1 \\ 6 & 1 \\ 9 & 1 \\ 12 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 3 \\ 4 \\ 4 \\ 7 \end{bmatrix}$$

Za pomocą poznanej metody najmniejszych kwadratów, aby rozwiązać takie równanie, musimy rozwiązać normalny układ równań $A^T Ax = A^T b$. Mamy zatem

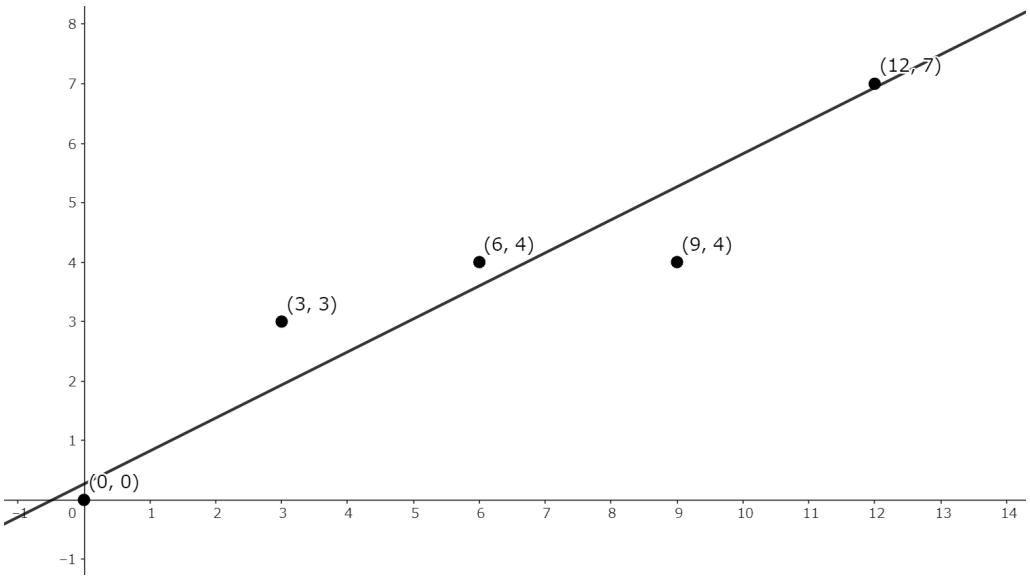
$$A^T A = \begin{bmatrix} 0 & 3 & 6 & 9 & 12 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 3 & 1 \\ 6 & 1 \\ 9 & 1 \\ 12 & 1 \end{bmatrix} = \begin{bmatrix} 270 & 30 \\ 30 & 5 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 0 & 3 & 6 & 9 & 12 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 4 \\ 4 \\ 7 \end{bmatrix} = \begin{bmatrix} 158 \\ 18 \end{bmatrix}$$

Rozwiązujemy zatem układ równań

$$A^T Ax = A^T b$$
$$(A^T A)^{-1}(A^T A)x = (A^T A)^{-1}(A^T b)$$
$$x = (A^T A)^{-1}(A^T b)$$
$$x = \begin{bmatrix} 270 & 30 \\ 30 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 158 \\ 18 \end{bmatrix} = \begin{bmatrix} 5/9 \\ 4/15 \end{bmatrix}$$

Otrzymaliśmy zatem, że prosta $y = \frac{5}{9}x + \frac{4}{15}$ jest najlepszą liniową zależnością pomiędzy współrzędnymi punktów $(0, 0), (3, 3), (6, 4), (9, 4)$ i $(12, 7)$.



4.2 Zastosowanie

Skoro poznaliśmy już narzędzie dopasowania prostej, spróbujmy je wykorzystać przy problemie ze wstępu. Znajdźmy prostą która będzie najlepiej dopasowana do danych o wzroście roślin.

Przypomnijmy, badaliśmy wzrost roślin przy różnych ilościach światła. Dane mamy podane w tabeli:

Światło słoneczne (kiloluks)	1	2	3	4	5
Wzrost roślin (cm)	4.5	7.9	6.5	9.9	13.3

Mamy zatem układ równań $Ax = b$, gdzie:

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}, \quad b = \begin{bmatrix} 4.5 \\ 7.9 \\ 6.5 \\ 9.9 \\ 13.3 \end{bmatrix}$$

Konstruujemy *układ normalny* $A^T Ax = A^T b$. Policzmy macierze $A^T A$ i $A^T b$:

$$A^T A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix} = \begin{bmatrix} 55 & 15 \\ 15 & 5 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4.5 \\ 7.9 \\ 6.5 \\ 9.9 \\ 13.3 \end{bmatrix} = \begin{bmatrix} 145.9 \\ 42.1 \end{bmatrix}$$

Mając macierze $A^T A$ i $A^T b$ wystarczy rozwiązać równanie macierzowe. Tym razem skorzystamy z metody eliminacji Gaussa.

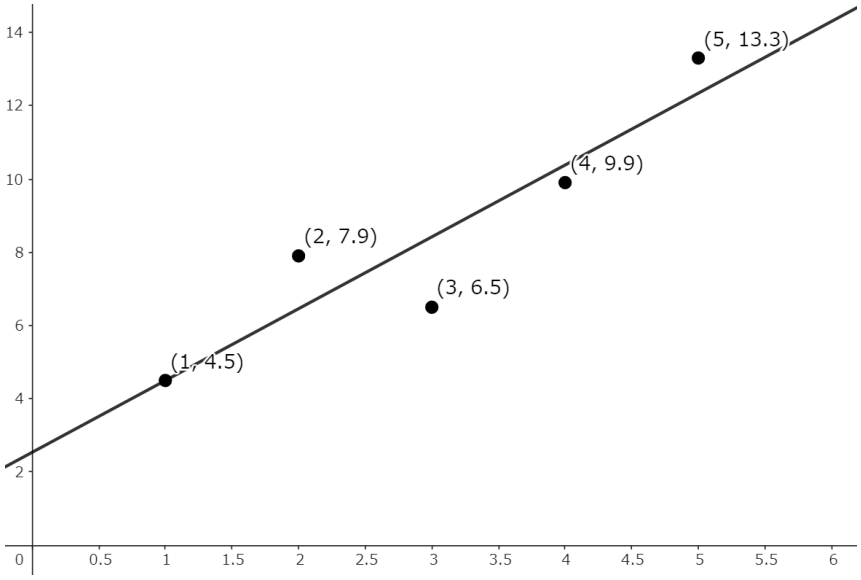
$$[\, A \,|\, b \,] \quad \xrightarrow{\text{operacje elementarne}} \quad [\, I \,|\, x \,]$$

Podstawiając wartości otrzymujemy:

$$\left[\begin{array}{cc|c} 55 & 15 & 145.9 \\ 15 & 5 & 42.1 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & \frac{3}{11} & \frac{1459}{550} \\ 15 & 5 & \frac{421}{10} \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & \frac{3}{11} & \frac{1459}{550} \\ 0 & \frac{10}{11} & \frac{127}{55} \end{array}\right] \sim$$

$$\left[\begin{array}{cc|c} 1 & \frac{3}{11} & \frac{1459}{550} \\ 0 & 1 & \frac{125}{50} \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & \frac{49}{25} \\ 0 & 1 & \frac{127}{50} \end{array}\right] = \left[\begin{array}{cc|c} 1 & 0 & 1.96 \\ 0 & 1 & 2.54 \end{array}\right]$$

Rozwiązaniem naszego układu równań jest więc $a = 1.96$ i $b = 2.54$, zatem prosta $y = 1.96x + 2.54$ jest najlepszą liniową zależnością między naszymi punktami. Narysujmy ją zatem na wykresie.



5 Wnioski

Jak zobaczyliśmy, regresja liniowa jest bardzo przydatnym narzędziem używanym na codzień w przeróżnych dziedzinach nauki, szczególnie gdy mamy potrzebę analizy i interpretacji danych. Dzięki niemu możemy skutecznie modelować zależności między zmiennymi oraz prognozować przyszłe zachowania. Przy wyższym poziomie zaawansowania możemy nawet dopasowywać różnorakie funkcje wielomianowe co pozwala na jeszcze bardziej precyzyjne predykcje.