Regresja wielowymiarowa na danych MPG

Patryk Michalak

13 stycznia 2025

1 Wprowadzenie

1.1 Regresja

Regresja jest techniką statystyczną polegającą na wyliczeniu współczynników wyrażenia na podstawie pewnego wektora argumentów $\overrightarrow{x}$ i wyniku y, do postaci $y = f(\overrightarrow{x}, \overrightarrow{\beta}) + \epsilon$. Wymienia się dwa typy regresji

- Regresja liniowa - $f(\overrightarrow{x}, \overrightarrow{\beta}) = x_1\beta_1 + x_2\beta_2 + ... + x_n\beta_n$
- Regresja logistyczna

1.2 Metoda najmniejszych kwadratów wielowymiarowa

Do policzenia współczynników można zastosować poniższy wzór

$$\overrightarrow{\beta} = (XX^T)^{-1}X^T * \overrightarrow{y}$$

Gdzie X to macierz argumentów z dodaną kolumną jedynek, której każda nowy wiersz jest kolejną daną do regresji ,X^T jest macierzą transponowaną X a wektor $\rightarrow$

2 Przebieg

2.1 Parametry

Opracujemy dane MPG (miles per gallon), które są spisem różnych pojazdów które mają MPG w zależności od parametrów pojazdów.

Ze względu na mnogą liczbę różnych wartości wskaźnika, możemy przyjąć że mamy tu doczynienia z regresją liniową zestawu danych. Na wskaźnik mil na gal wpływa siedem różnych parametrów:

- Ilość cylindrów silnika (cylinders)
- Objętość Skokowa (displacement)
- Konie mechaniczne (horsepower)
- Masa pojazdu (weight)
- Akceleracja (accelaration)
- Rok produkcji (model_year)
- Miejsce produkcji (origin)

Dodatkowo jest także podana nazwa pojazdu dla identyfikacji. Część pojazdów jednakże nie posiada znanej ilości koni mechaniczne, oznaczone w pliku wejściowym jak '?' - te pojazdy nie będą brane pod uwagę podczas regresji. Otrzymujemy macierz parametrów o wymiarze Liczba wpisów (n) * Ilość Parametrów (m).

2.2 Regresja liniowa własnoręczna używając metodę najmniejszych kwadratów

Tworzymy macierz X którą pierwszą kolumnę wypełniamy jedynkami a kolejne wypełniamy naszymi parametrami dla każdego pojazdu. Także tworzymy macierz X^T poprzez transpozycję macierzy X.

$$X = \begin{bmatrix} 1 & cylinders_1 & displacement_1 & horsepower_1 & weight_1 & acceleration_1 & model\_y_1 & origin_1 \\ 1 & cylinders_2 & displacement_2 & horsepower_2 & weight_2 & acceleration_2 & model\_y_2 & origin_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & cylinders_n & displacement_n & horsepower_n & weight_n & acceleration_n & model\_y_n & origin_n \end{bmatrix}$$

Także przyjmujemy wektor n-wymiarowy który przyjmuje kolejne wartości mil na gal.

$\overrightarrow{y} = \begin{bmatrix} mpg_1 & mpg_2 & \dots & mpg_n \end{bmatrix}$

Stosując wzór (Wzór na współczynniki) otrzymujemy wektor współczynników dla każdego parametrów $\overrightarrow{\beta} = \begin{bmatrix} \beta_\phi & \beta_0 & \beta_1 & \dots & \beta_m \end{bmatrix}$

gdzie β_ϕ jest wyrazem wolnym.

$$\overrightarrow{\beta}_{squares} = [-17.2184346220103, -0.4933, 0.0198, -0.0169, -0.0064, , 0.0805, 0.7507, 1.4261]$$

2.3 Regresja liniowa posługując się biblioteką sklearn

Przy użyciu biblioteki sklearn, możemy załadować nasze dane do modelu regresji liniarnej. Otrzymujemy współczynniki

$$\overrightarrow{\beta}_{lib} = [-0.4933, 0.0198, -0.0169, -0.0064, , 0.0805, 0.7507, 1.4261]$$

2.4 Porównanie wyników

Różnica pomiędzy wynikami regresji z biblioteki a własnoręcznie napisanej regresji metodą najmniejszych jest rzędu 10^{-14} - możemy przyjąć że obie te metody dają te same wyniki, przyjmując że ignorujemy wolny współczynnik