Regresje

Zaawansowane metody obliczeniowe

Autor:
Oskar Swat

14 stycznia 2024 r.

Wstęp

Celem tego sprawozdania jest zaprezentowanie pracy związanej z regresją logistyczną oraz wyjaśnienie, czym jest regresja logistyczna i liniowa. W ramach pracy zostanie użyta regresja logistyczna na konkretnych danych oraz przeprowadzone zostaną analizy i wizualizacje wyników.

Cel pracy

Celem pracy jest zrozumienie i zastosowanie regresji logistycznej na danych dotyczących pacjentów, aby przewidzieć, czy pacjent przeżyje 5 lat lub dłużej po operacji, czy też zmarł w ciągu 5 lat.

Zakres pracy

Dane, na których będziemy pracować, pochodzą z pliku haberman.csv i zawierają informacje o pacjentach, takie jak wiek, rok operacji, liczba wykrytych pozytywnych węzłów pachowych oraz stan przeżycia. Wybór regresji logistycznej do analizy danych wynika z charakteru problemu klasyfikacji binarnej (przewidzenie jednej z dwóch kategorii) oraz ze względu na istnienie nieliniowych zależności między atrybutami a stanem przeżycia.

Metodyka

Praca została przeprowadzona w języku Python, wykorzystując biblioteki takie jak numpy, pandas, scikit-learn, seaborn, oraz matplotlib. Dane zostały wczytane z pliku haberman.csv do ramki danych pandas. Po zaimplementowaniu regresji logistycznej, przeprowadzono również analizę za pomocą gotowego modelu regresji logistycznej z biblioteki scikit-learn.

Wstęp do regresji

Regresja logistyczna i regresja liniowa to dwa podstawowe rodzaje analizy regresji wykorzystywane w statystyce i uczeniu maszynowym. Oba modele służą do modelowania zależności między zmiennymi niezależnymi a zmiennymi zależnymi, jednak różnią się w swoich założeniach i matematycznych podstawach.

Regresja Liniowa

Regresja liniowa jest używana do modelowania liniowych zależności między zmiennymi niezależnymi a zmienną zależną. W regresji liniowej zakłada się, że zmienna zależna jest liniową kombinacją zmiennych niezależnych, co można zapisać w postaci wzoru matematycznego:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{1}$$

gdzie:

Y to zmienna zależna,
β_0 to wyraz wolny (przesunięcie),
$\beta_1, \beta_2, \ldots, \beta_n$ to współczynniki regresji dla zmiennych niezależnych $X_1, X_2, \ldots, X_n$,
$X_1, X_2, \ldots, X_n$ to zmienne niezależne,
ϵ to błąd losowy.

Celem regresji liniowej jest znalezienie optymalnych wartości współczynników β_i, które minimalizują sumę kwadratów różnic między wartościami przewidywanymi a rzeczywistymi.

Jednym z ważnych założeń regresji liniowej jest liniowość relacji między zmiennymi oraz brak autokorelacji i homoskedastyczności błędów.

Regresja Logistyczna

Regresja logistyczna, w przeciwieństwie do regresji liniowej, jest używana w problemach klasyfikacji binarnej, gdzie celem jest przewidzenie przynależności do jednej z dwóch klas. Regresja logistyczna wykorzystuje funkcję logistyczną do modelowania prawdopodobieństwa przynależności do danej klasy. Wzór matematyczny regresji logistycznej można zapisać następująco:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{2}$$

gdzie:

$P(Y = 1)$ to prawdopodobieństwo przynależności do klasy 1,
β_0 to wyraz wolny (przesunięcie),
$\beta_1, \beta_2, \ldots, \beta_n$ to współczynniki regresji dla zmiennych niezależnych $X_1, X_2, \ldots, X_n$,
$X_1, X_2, \ldots, X_n$ to zmienne niezależne,
e to liczba Eulera (około 2.71828).

Funkcja logistyczna $\frac{1}{1+e^{-z}}$ pozwala na ograniczenie wyników do przedziału (0, 1), co można interpretować jako prawdopodobieństwo przynależności do klasy 1.

Funkcja logistyczna (sigmoidalna):

$$p(x, b) = \frac{1}{1 + e^{-bz}} \tag{3}$$

Ta funkcja pozwala mapować liniową kombinację bx na zakres (0, 1) i jest używana do modelowania prawdopodobieństwa przynależności do klasy 1 w regresji logistycznej. x to wektor zmiennych niezależnych, a b to wektor współczynników.

Funkcja wiarygodności (likelihood):

$$L(b) = \prod_i p(x_i, b)^{y_i} \cdot (1 - p(x_i, b))^{1-y_i} \tag{4}$$

Funkcja wiarygodności jest używana w procesie maksymalizacji podczas dopasowywania modelu regresji logistycznej do danych. Jest to iloczyn prawdopodobieństw przynależności do odpowiednich klas (1 lub 0) dla każdej obserwacji i.

Funkcja log-wiarygodności (log-likelihood):

$$L(b) = \sum_i y_i \log(p(x_i, b)) + (1 - y_i)(1 - p(x_i, b)) \tag{5}$$

Funkcja log-wiarygodności jest logarytmem funkcji wiarygodności. Jest używana, aby uprościć obliczenia, ponieważ iloczyn prawdopodobieństw w funkcji wiarygodności może prowadzić do małych wartości i utraty dokładności numerycznej. Dzięki zastosowaniu logarytmu można przekształcić iloczyn w sumę, co ułatwia obliczenia.

Współczynnik Uczenia

W regresji logistycznej, współczynnik uczenia (α) kontroluje, jak szybko model dostosowuje się do danych podczas procesu uczenia. Odpowiedni wybór współczynnika uczenia jest istotny, aby uniknąć przeuczenia lub niedouczenia modelu.

Wartości Początkowe

Wartości początkowe współczynników (β_i) w regresji logistycznej mogą mieć wpływ na proces uczenia. Zazwyczaj są one inicjowane losowo lub przy użyciu pewnych heurystyk. Wybór odpowiednich wartości początkowych może przyspieszyć proces zbieżności algorytmu.

Podsumowując, regresja logistyczna jest potężnym narzędziem do modelowania zależności w problemach klasyfikacji binarnej, wykorzystując funkcję logistyczną do przewidywania prawdopodobieństwa przynależności do danej klasy. Ma ona swoje unikalne cechy i założenia, które odróżniają ją od regresji liniowej, co sprawia, że jest bardziej odpowiednia dla takich zadań.

Sposób realizacji

Działanie Programu

Program został zaimplementowany w języku Python i korzysta z różnych bibliotek, takich jak NumPy, pandas, scikit-learn, seaborn i matplotlib. Główne kroki programu to:

1. Wczytanie danych z pliku "haberman.csv"do ramki danych pandas.
2. Przeprowadzenie preprocessingu danych, w tym przekształcenie wartości stanu przeżycia na wartości binarne (0 lub 1).
3. Implementacja regresji logistycznej ręcznie w funkcji `logistic_regression`, która używa gradientowego spadku do dostosowania wag modelu.
4. Użycie gotowego modelu regresji logistycznej z biblioteki scikit-learn do porównania wyników.
5. Ocena modelu przy użyciu macierzy konfuzji, wygenerowanie raportu klasyfikacji oraz wyświetlenie heatmapy macierzy konfuzji.

Dane

Dane pochodzące z pliku *"haberman.csv"* zawierają informacje o pacjentach poddanych operacjom związanym z rakiem piersi. Atrybuty danych to:

- Wiek pacjenta w momencie operacji (numeryczny).
- Rok operacji pacjenta (rok - 1900, numeryczny).
- Liczba wykrytych pozytywnych węzłów pachowych (numeryczna).
- Stan przeżycia (atrybut klasy):
  - 1: pacjent przeżył 5 lat lub dłużej.
  - 2: pacjent zmarł w ciągu 5 lat.

Wyniki

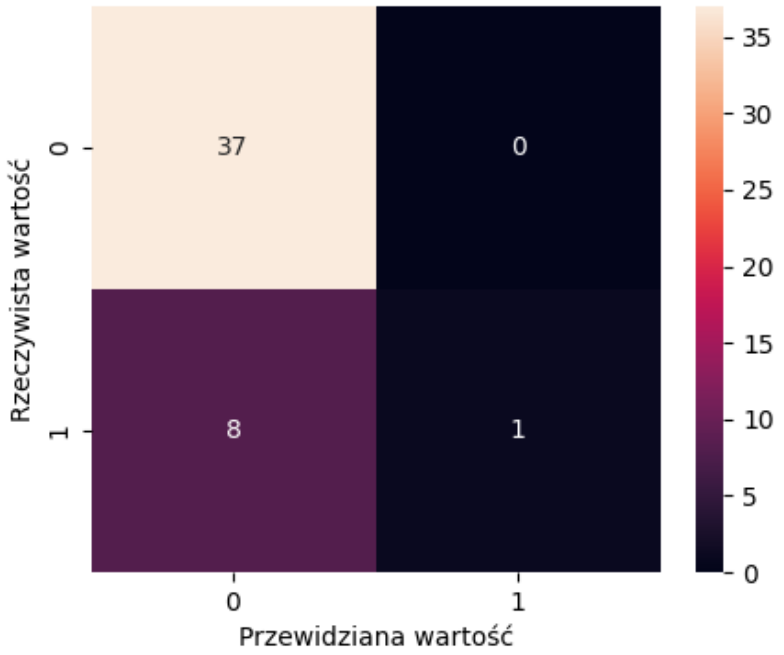
Wyniki treningu regresji logistycznej

- Wyniki treningu regresji logistycznej przedstawione są w kolejnych iteracjach.
- Wartości B reprezentują współczynniki modelu regresji logistycznej dla zmiennych niezależnych. W ostatniej iteracji (iteracja 100) otrzymane współczynniki to: **[-1.22903907, 0.93935797, -1.1608563]** .
- Log likelihood to wartość funkcji log-wiarygodności dla danego zestawu współczynników. Im wyższa ta wartość, tym lepiej model pasuje do danych. W ostatniej iteracji wynosi ona **-2517.26002076** .

Wyniki klasyfikacji

- Raport klasyfikacji prezentuje miary oceny jakości klasyfikacji dla modelu regresji logistycznej.
- Precision to stosunek prawdziwie pozytywnych przypadków do wszystkich przypadków pozytywnych, wynosi 0.74 dla klasy 1 i 0.50 dla klasy 2.
- Recall to stosunek prawdziwie pozytywnych przypadków do wszystkich przypadków, które powinny być pozytywne, wynosi 0.96 dla klasy 1 i 0.12 dla klasy 2.
- F1-score to miara równowagi między precision a recall. Wynosi 0.84 dla klasy 1 i 0.19 dla klasy 2.
- Support to liczba próbek w danej klasie.

Heatmapa macierzy konfuzji



Rysunek 1: Macierz konfuzji (dla danego uruchomienia programu)

- Rysunek 1 prezentuje wyniki klasyfikacji w formie tabeli. Na diagonalnej znajdują się wartości prawidłowo sklasyfikowanych przypadków, poza nią błędnie sklasyfikowane przypadki.
- W macierzy konfuzji widzimy, że dla klasy 1 (przetrwałych pacjentów) model poprawnie sklasyfikował 37 przypadki, ale błędnie sklasyfikował 0 przypadków jako klasy 2 (zmarli).
- Dla klasy 2 (zmarli pacjenci) model poprawnie sklasyfikował 1 przypadek, ale błędnie sklasyfikował 8 przypadki jako klasy 1

Podsumowanie

Dlaczego Regresja Logistyczna jest odpowiednia dla tych danych

Dane pochodzące z pliku "haberman.csv"obejmują informacje medyczne o pacjentach, a celem było przewidzenie, czy pacjent przeżyje 5 lat lub dłużej po operacji. To jest problem klasyfikacji binarnej, który jest odpowiedni do rozwiązania przy użyciu regresji logistycznej, ponieważ chcemy modelować prawdopodobieństwo przynależności do jednej z dwóch kategorii.

Celem pracy było zrozumienie i zastosowanie regresji logistycznej w analizie danych medycznych z pliku "haberman.csv". W pierwszej części pracy porównane zostały, obydiwe regresje. Następnie dokonałem analizę danych i zainplementowałem regresję logistyczną zarówno ręcznie, korzystając z algorytmu gradientowego spadku, jak i za pomocą gotowego modelu z biblioteki scikit-learn. Wyniki treningu modelu obejmowały obserwacje z kolejnych iteracji, wartości współczynników modelu oraz log-wiarygodność. Dzięki tym wynikom można monitorować proces uczenia modelu.

Po uzyskaniu wyników treningu, można ocenić jakość modelu przy użyciu macierzy konfuzji oraz raportu klasyfikacji, który zawierał miary takie jak precision, recall i F1-score. Wyniki klasyfikacji pokazały, że nasz model dobrze radził sobie z przewidywaniem pacjentów, którzy przeżyli, ale miał trudności w klasyfikacji pacjentów, którzy zmarli.

Literatura

- codebasics, "Machine Learning Tutorial",
  Link do YouTube
- Mirosław Mamaczur, "Jak działa regresja logistyczna?",
  Link do strony
- Maher Maalouf, "Logistic regression in data analysis: An overview",
  Link do ResearchGate