Metody analizy danych

Regresja

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wersja 1
19 listopada 2021

Regresja liniowa

- W zadaniu regresji liniowej dopasowujemy parametry (β, α), aby przy pomocy zmiennych x_i wyjaśnić wartość y:

$$y = \sum_{i=1}^{n} \beta_i x_i + \alpha.$$

- Liczba niezerowych współczynników β_i oznacza ile zmiennych będzie użytych w budowanym modelu.

Wyliczenie współczynników

- Dla regresji liniowej estymator współczynników ma postać:

$$\hat{\beta} = \mathrm{argmin}_{(\alpha,\beta)}(\alpha + \mathbf{X}\beta - Y)^T(\alpha + \mathbf{X}\beta - Y),$$

- zadanie polega na zminimalizowaniu odległości między wartościami wyliczanymi przez równanie prostej $\hat{y} = \mathbf{X}\beta + \alpha$, a wartościami ze zbioru uczącego Y.
- Dzięki zastosowaniu we wzorze iloczynu skalarnego zadanie sprowadza się do minimalizacji jednej wartości, wyliczanej jako kwadrat różnicy między poprawnymi, a szacowanymi wynikami.

Iloczyn skalarny

- I;xoczyn skalarny definiuje się wzorem

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + \cdots + a_n b_n$$

- W zapisie macierzowym ma on postać $\mathbf{A} \cdot \mathbf{B} = \mathbf{A}^{\mathrm{T}}\mathbf{B}$, gdzie $\mathbf{A}^{\mathrm{T}}$ jest transpozycją macierzy.
- Transpozycja $[x_1, x_2, \ldots, x_n]^{\mathrm{T}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
- Czyli:

$$\mathbf{A} \cdot \mathbf{A} = \sum_{i=1}^{n} a_i^2.$$

Estymator najmniejszych kwadratów

- Przyjmijmy zapis $\alpha = \beta_0$.
- W celu minimalizacji funkcji $(\beta_0 + \mathbf{X}\beta - Y)^T(\beta_0 + \mathbf{X}\beta - Y)$ stosujemy estymator najmniejszych kwadratów.
- Estymator wyliczamy jako $\hat{\beta} = (X^T X)^{-1} XY$, gdzie $(X^T X)^{-1}$ jest odwrotnością macierzy $(X^T X)$, czyli $(X^T X) \cdot (X^T X)^{-1} = I$.
- Można udowodnić (Twierdzenie Gaussa-Markowa), że estymator najmniejszych kwadratów jest najlepszym spośród liniowych, nieobciążonych[1] estymatorów liniowego modelu regresji [Bingham and Fry, 2010].

[1] Wartość oczekiwana estymatora jest równa wartości estymowanego parametru.

Wpływ liczby zmiennych na regresję

- Duża liczba zmiennych użytych do budowy modelu grozi nadmiernym dopasowaniem modelu do zbioru treningowego.
- Większa liczba współczynników niezerowych utrudnia zrozumienie modelu i wnioskowanie na jego podstawie.
- Z powyższych powodów dążymy do budowy modeli rzadkich, z małą liczbą współczynników niezerowych.

Regularyzacja

- Regularyzacja jest techniką polegającą na dodawaniu do czynnika błędu kary, która jest zwiększana wraz ze wzrostem wartości β.
- Następnie próbujemy zminimalizować sumę błędów i kar.
- Technika zabezpiecza przed powstawaniem dużej liczby niezerowych współczynników.

Metody regularyzacji

- Ograniczenie wartości współczynników modelu regresyjnego można przeprowadzać na różne sposoby.
    - Ograniczenie normy L_2 wektora współczynników - regresja grzbietowa.
    - Ograniczenie normy L_1 wektora współczynników - regresja LASSO.
    - Zastosowanie mieszaniny norm - sieci elastyczne.

Normy

- Norma L_1:

$$\|\mathbf{X}\|_1 = \sum_{i=1}^{n} |x_i|.$$

- Norma L_2:

$$\|\mathbf{X}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Regresja grzbietowa

- Dla regresji grzbietowej wzór na estymator współczynników ma postać:

$$\hat{\beta}_{ridge} = \text{argmin}_{(\alpha,\beta)}(\alpha + \mathbf{X}\beta - y)^T(\alpha + \mathbf{X}\beta - y) + \lambda\beta^T\beta,$$

gdzie λ jest współczynnikiem funkcji kary regularyzacji

Regresja LASSO

- LASSO (*Least Absolute Shrinkage and Selection Operator*).
- Dla regresji LASSO wzór na estymator współczynników ma postać:

$$\hat{\beta}_{LASSO} = \mathrm{argmin}_{(\alpha,\beta)} \frac{1}{N} \|\alpha + \mathbf{X}\beta - y\|_2^2 + \lambda \|\beta\|_1,$$

  gdzie λ jest współczynnikiem funkcji kary regularyzacji, a N wielkością próbki uczącej.
- Regresja LASSO pozwala uzyskać $\beta_i = 0$.

Sieci elastyczne

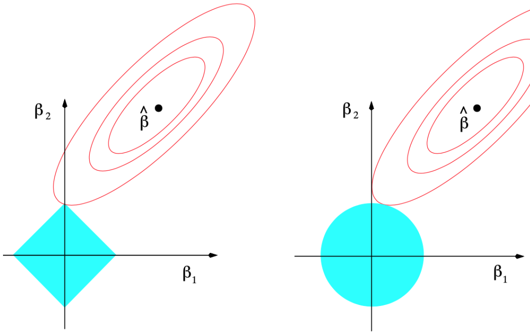
- Dla regresji siecią elastyczną wzór na estymator współczynników ma postać:

$$\hat{\beta}_{net} = \operatorname{argmin}_{(\alpha,\beta)} \frac{1}{N}\|\alpha + \mathbf{X}\beta - y\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2,$$

gdzie λ jest współczynnikiem funkcji kary regularyzacji, a N wielkością próbki uczącej.

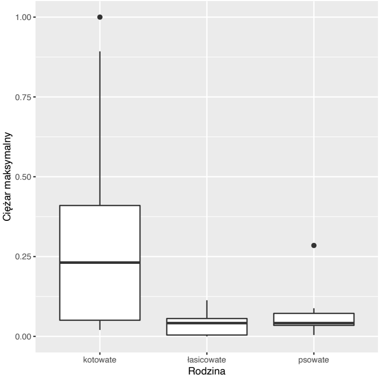
Selekcja zmiennych

- Zwiększając współczynniki λ dowolnej metody regularyzacji będziemy zmniejszać współczynniki β.
- Jednakże tylko metoda Lasso może, dzięki zastosowaniu normy L_1, zerować współczynniki.



Rysunek 1: Norma kulista regresji grzbietowej nie osiągnie $\beta_i = 0$ na co pozwala norma kwadratowa regresji Lasso [Hastie et al., 2001]

Przykład estymacji ciężaru

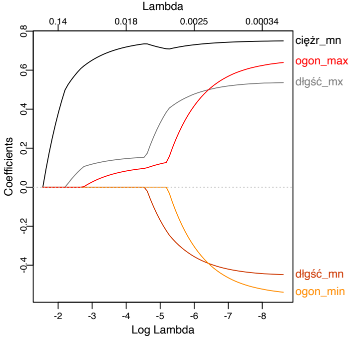


Rysunek 2: Rozkład ciężaru maksymalnego w rodzinach [Wydawnictwo Rebel, 2019]

- Analizowano trzy rodziny zwierząt
  - łasicowate, kotowate, psowate
- Zadaniem jest estymacja maksymalnego ciężaru w ramach gatunków znając
  - minimalny ciężar,
  - maksymalną i minimalną długość,
  - maksymalną i minimalną długość ogona.

Estymacja ciężaru

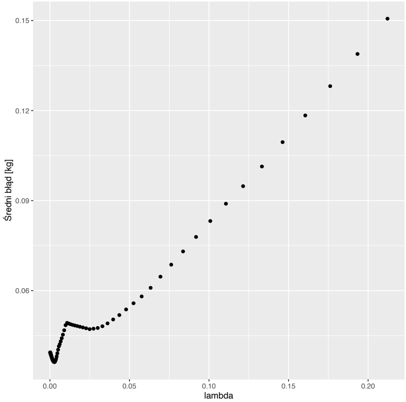
- Używając metody lasso, zbudujmy modele wyjaśniające maksymalny ciężar zwierzęcia.



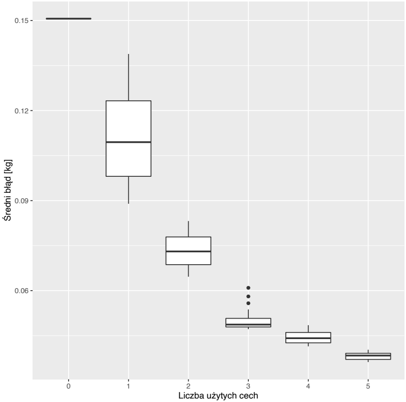
Rysunek 3: Zmiana współczynników β ze wzrostem współczynnika λ

- Istotność zmiennych wynikająca z kolejności zerowania współczynników
  - ciężar_min
  - długość_max
  - ogon_max
  - długość_min
  - ogon_min
- Wykres ilustrujący zmianę współczynników β względem współczynnika λ nazywamy ścieżkami lasso.

Wpływ liczby zmiennych na wyniki



Rysunek 4: Wpływ współczynnika λ na wyniki



Rysunek 5: Wpływ liczby zmiennych na wyniki

Regresja logistyczna

- Regresja logistyczna jest używana, gdy zmienna zależna przyjmuje tylko dwie wartości (najczęściej kodowane jako 0 i 1).
- Regresja logistyczna wyraża prawdopodobieństwo jako szansę, czyli stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki.

Szansa

- Szansę (ang. *odds*) można wyliczyć z prawdopodobieństwa p jako:

$$Odds = \frac{p}{1-p} = e^{\alpha} e^{\beta x}$$

gdzie,

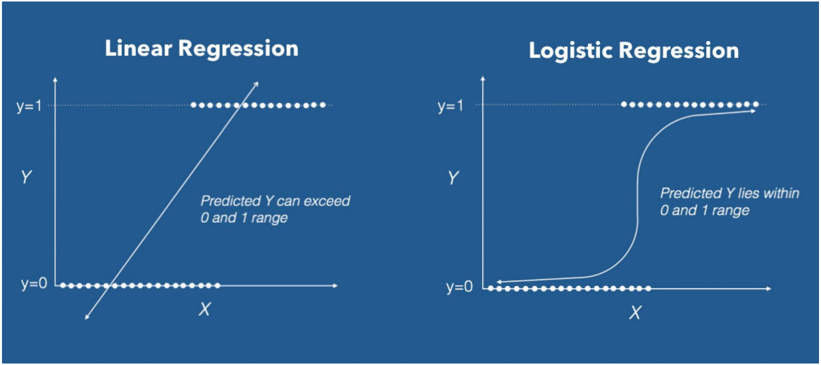
α stała regresji,
β współczynnik regresji logistycznej,
x zmienna niezależna.

Właściwości szansy

- Szansa przekształca prawdopodobieństwo z zakresu $0 < p < 1$ na wartość z zakresu $(0, +\infty)$.
- Jej logarytm przyjmuje wartości z zakresu $(-\infty, +\infty)$.
- Dzięki temu można szacować logarytm szansy metodami regresji nieograniczonymi do przedziału $[0, 1]$.
- Funkcja logit przekształca prawdopodobieństwo na logarytm szansy:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p).$$

Porównanie regresji liniowej i logistycznej



Rysunek 6: Regresja liniowa a logistyczna [Prabhakaran, 2017]

Regresja logistyczna jako klasyfikator

- Regresja logistyczna może być wykorzystywana jako klasyfikator binarny.
- Prawdopodobieństwo przynależności obserwacji x_i do klasy $y = 1$ określimy jako

$$P(y = 1|x_i) = \frac{e^{\alpha + \beta_i x_i}}{1 + e^{\alpha + \beta_i x_i}},$$

lub dla wielu zmiennych

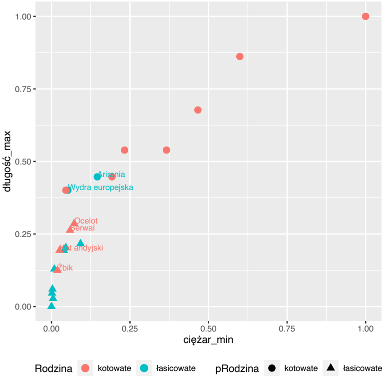
$$P(y = 1|x_1, \ldots, x_n) = \frac{e^{\alpha + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\alpha + \sum_{i=1}^{n} \beta_i x_i}}.$$

Trenowanie klasyfikatora

- Klasyfikator trenujemy algorytmem iteracyjnej ważonej metody najmniejszych kwadratów.
- Wybieramy $\hat{\beta}_0 = 0$, obliczamy p_i^0 maksymalizując funkcję wiarygodności, świadczącą o dobrym przyjęciu parametrów modelu.
- Następnie:
  1. Wyliczamy
  $$Z_i = logit(p_i^s) + \frac{Y_i - p_i^s}{p_i^s \cdot (1 - p_i^s)},$$
  2. Niech W będzie macierzą diagonalną o wartościach $p_i^s \cdot (1 - p_i^s)$ na przekątnej.
  3. Przyjmujemy
  $$\hat{\beta}^s = (X^T WX)^{-1} X^T WZ,$$
  4. Podstawiamy $s = s + 1$ i wracamy do kroku 1.

Estymacja ciężaru

- Używając regresji logistycznej, zbudujmy model rozróżniający dwie rodziny najbardziej różniące się maksymalnym ciężarem zwierzęcia.



Rysunek 7: Klasyfikacja rodzin

- Za zmienne opisowe posłużyły dwie zmienne najlepiej modelujące ciężarmaksymalny:
  - ciężar_min
  - długość_max
- Sześć źle zidentyfikowanych gatunków, podpisane na wykresie, ma nietypową dla swoich rodzin długość.

Bibliografia I

[Bingham and Fry, 2010] Bingham, N. and Fry, J. (2010).
Regression: Linear Models in Statistics.
Springer Undergraduate Mathematics Series.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001).
The Elements of Statistical Learning.
Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Prabhakaran, 2017] Prabhakaran, S. (2017).
Logistic regression – a complete tutorial with examples in r.

[Wydawnictwo Rebel, 2019] Wydawnictwo Rebel (2019).
Fauna.