METODY EKSPLORACJI DANYCH

Laboratorium. Analiza regresji.
Dobór zmiennych do modelu wieloczynnikowego
Regresja z wykorzystaniem tzw. regularyzacji

WPROWADZENIE TEORETYCZNE

Regularyzacja to technika, która pomaga ograniczyć przeuczenie (ang. overfitting) modelu poprzez dodanie dodatkowego składnika kary do funkcji kosztu. Dzięki temu model jest zmuszony do wyboru prostszych parametrów, co zazwyczaj prowadzi do lepszej zdolności modelu do generalizacji na nowych danych.

Dwie popularne formy regularyzacji stosowane w regresji to:

1) LASSO (ang. Least Absolute Shrinkage and Selection) LASSO), która jest znana jako regularyzacja typu L1 (ang. L1 regularization)
2) Ridge (pol. grzbietowa) – regularyzacja typu L2 (ang. L2 regularization)

W przypadku regresji liniowej dla regularyzacji L1 lub L2 znajdujemy rozwiązanie następującego problemu optymalizacji:

Szukamy takiego wektora $\boldsymbol{a} = (a_0, a_1, \dots, a_k)$, że

$$\min_{\boldsymbol{a}=(a_0,a_1,\dots,a_k)} \left\{ \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \lambda_L \sum_{j=1}^{k} |a_j|^L \right\}$$

gdzie
n jest liczbą obserwacji (w zbiorze danych $D = \{(\boldsymbol{x_i}, y_i) : i = 1, \dots n;\ \boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots x_{ik})\}$
$\widehat{y_i} = a_o + a_1 x_{i1} + a_2 x_{i2}, \dots, a_k x_{ik}$ – wartością modelu dla i-tej obserwacji $\boldsymbol{x_i}$,
dla $L = 1$ mamy regularyzację LASSO,
dla $L = 2$ – grzbietową (ang. Ridge).

Innymi słowy.

Dla LASSO rozwiązujemy zadanie:

$$\min_{\boldsymbol{a}=(a_0,a_1,\ldots,a_k)} \left\{ \sum_{i=1}^{n} (y_i - a_o - a_1 x_{i1} - a_2 x_{i2} - \cdots - a_k x_{ik})^2 + \lambda_1 (|a_1| + |a_2| + \cdots + |a_k|) \right\}$$

Dla regresji grzbietowej – zadanie:

$$\min_{\boldsymbol{a}=(a_0,a_1,\ldots,a_k)} \left\{ \sum_{i=1}^{n} (y_i - a_o - a_1 x_{i1} - a_2 x_{i2} - \cdots - a_k x_{ik})^2 + \lambda_2 (|a_1|^2 + |a_2|^2 + \cdots + |a_k|^2) \right\}$$

Do czego i dlaczego regularyzacja jest potrzebna?
W klasycznej regresji liniowej minimalizujemy sumę kwadratów błędów. Jeśli liczba cech jest duża w stosunku do liczby obserwacji lub cechy są silnie skorelowane, model może dopasować się „idealnie" do danych treningowych, ale będzie bardzo wrażliwy na tzw. szum – czyli będzie słabo prognozował nowe przypadki. Regularyzacja wprowadza kontrolę nad wielkością współczynników, co redukuje tę niestabilność.

Parametr λ ($\lambda 1$ lub $\lambda 2$) to hiperparametr regularyzacji, który określa siłę kary.

Małe λ to mały wpływ regularyzacji (model bliski zwykłej regresji), duże λ - silne „ściskanie" współczynników.

Jak wybrać λ?
Zazwyczaj używa się walidacji krzyżowej (np. k-fold CV) do przetestowania różnych wartości λ i wybrania tej, która minimalizuje błąd predykcji na zestawie walidacyjnym. Biblioteki takie jak scikit-learn oferują klasy Ridge, Lasso i ElasticNet, które automatycznie przeprowadzają ten proces.

Co z interpretacją współczynników?
Ridge: współczynniki są nadal wszystkie różne od zera, więc interpretacja wymaga uwzględnienia ich zmniejszonej wielkości.
LASSO: niektóre współczynniki mogą stać się dokładnie zerowe, co ułatwia identyfikację najważniejszych cech.

Elastic Net: kompromis między dwoma podejściami; przy dużej liczbie skorelowanych cech może wybrać grupę cech zamiast jednej dominującej. Wtedy rozwiązujemy zadanie:

$$\min_{\boldsymbol{a}=(a_0,a_1,\ldots,a_k)} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{k} |a_j| + \lambda_2 \sum_{j=1}^{k} |a_j|^2 \right\}$$

Podsumowanie

Regularyzacja jest kluczowym narzędziem w regresji, które pozwala:

- zmniejszyć ryzyko przeuczenia,
- poprawić stabilność numeryczną (szczególnie przy wysokiej współliniowości),
- w niektórych przypadkach wykonać automatyczną selekcję cech (LASSO, Elastic Net).

Dobór odpowiedniej metody i wartości λ zależy od charakterystyki danych (liczba cech, ich korelacje) oraz od tego, czy priorytetem jest interpretowalność (LASSO) czy jedynie stabilność predykcji (Ridge).

Regularne stosowanie walidacji krzyżowej zapewnia, że wybrany poziom regularyzacji rzeczywiście poprawia zdolność modelu do generalizacji.