Regularyzacja w modelu regresji

Martyna Śpiewak
Bootcamp Data Science

Model regresji liniowej

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i,$$

gdzie $\epsilon_i \sim \mathcal{N}(0, \sigma)$.

Zapis macierzowy:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

gdzie

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \ldots & x_{2(p-1)} \\ 1 & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n(p-1)} \end{bmatrix},$$

$\beta^T = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ oraz $\epsilon^T = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$.

Model regresji liniowej

Cel: Przy użyciu $(x_{11}, x_{12}, \ldots, x_{1(p-1)}, y_1), \ldots, (x_{n1}, x_{n2}, \ldots, x_{n(p-1)} y_n)$, wyznaczyć współczynniki $b_0, b_1, \ldots, b_{p-1}$ tak, aby

$$y_i \approx b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_{p-1} x_{i(p-1)}$$

Funkcja kryterialna:

$$\begin{aligned} \mathbf{b} &= \operatorname*{argmin}_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \operatorname*{argmin}_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \mathbf{X}_i \beta)^2, \end{aligned}$$

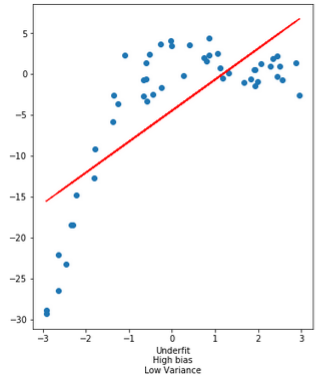
wówczas

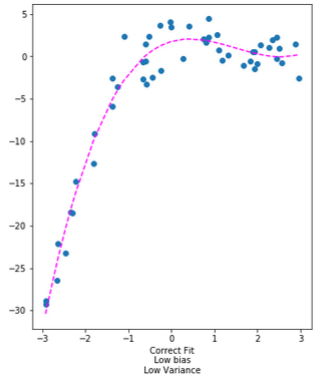
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

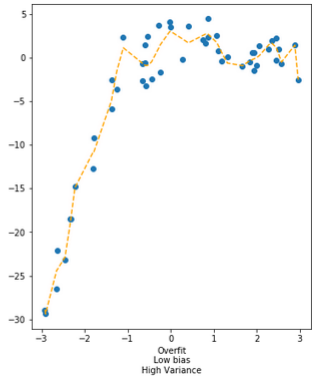
Kompromis między obciążeniem a wariancją (ang. *bias–variance tradeoff*)

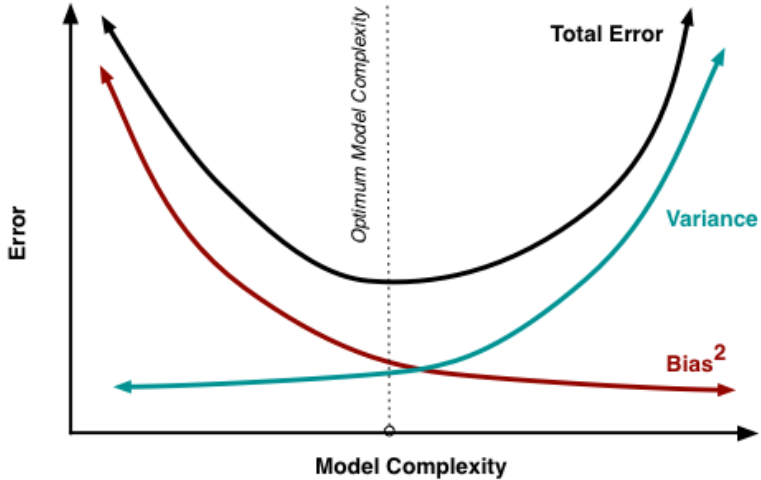
Obciążenie (ang. *bias*) — odnosi się do błędu wynikającego z uproszczonych założeń modelu dotyczących dopasowania danych. Wysokie obciążenie oznacza, że model nie jest w stanie uchwycić wzorców w danych, co powoduje niedopasowanie/niedouczenie (ang. *under-fitting*).

Wariancja (ang. *variance*) - odnosi się do błędu spowodowanego złożonym modelem próbującym dopasować dane. Duża wariancja oznacza, że model przechodzi przez większość punktów danych i powoduje nadmierne dopasowanie do danych (ang. *over-fitting*).


Underfit
High bias
Low Variance


Correct Fit
Low bias
Low Variance


Overfit
Low bias
High Variance



Źródło: http://scott.fortmann-roe.com/docs/BiasVariance.html

Przeuczenie modelu

Zbyt duża liczba zmiennych w modelu może prowadzić do **przeuczenia modelu**, co w terminologii statystycznej prowadzi do dużej wartości wariancji estymatora przy małym błędzie dopasowania.

Co zrobić aby uniknąć przeuczenia modelu?

- **regularyzacja** oparta o karę za wielkość współczynników w modelu, m.in. regularyzacja grzbietowa, Lasso i Elastic Net.
- **wybór zmiennych**, m.in. z użyciem kryteriów AIC i BIC,
- **redukcja wymiaru** zmiennych niezależnych przez zastosowanie techniki PCA lub PCR.

Regularyzacja

Regularyzacja jest techniką, która jest wykorzystywana do budowania **oszczędnych** modeli, w rozumieniu obecności zbyt dużej liczby predyktorów. Przez dużą liczbę predyktorów rozumiemy:

- duża liczba predyktorów to taka, która prowadzi do przeuczenia modelu – nawet tak niewielka liczba jak 10 zmiennych może prowadzić do przeuczenia,
- duża liczba predyktorów to taka, która może prowadzić do problemów z wydajnością obliczeniową – przy obecnych możliwościach komputerów, taka sytuacja może mieć miejsce przy występowaniu milionów lub miliardów cech.

Techniki regularyzacyjne działają poprzez

- karanie wielkości współczynników cech,
- minimalizowanie błędu między przewidywanymi a rzeczywistymi obserwacjami.

Regularyzacja grzbietowa (ang. *Ridge regression*)

Metoda najmniejszych kwadratów z **regularyzacją grzbietową** (inaczej $l2$), minimalizuje funkcję kryterialną:

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2 + \alpha\sum_{i=1}^{p-1}\beta_i^2,$$

gdzie α jest parametrem sterującym metody (siłą regularyzacji), $\alpha \geq 0$ oraz $\sum_{i=1}^{p-1}\beta_i^2 = ||\beta||_2^2$ jest normą $l2$.

Parametr α kontroluje jak silnie współczynniki modelu są ściągane do 0.

- dla $\alpha = 0$: problem uprasza się do zwykłej regresji,
- dla $\alpha \to +\infty$: współczynniki β_i „kurczą" się do zera.

Uwaga: Regresja grzbietowa tworzy modele zawierające wszystkie $p-1$ predyktorów. Współczynniki β_i kurczą się ze wzrostem α, ale nie muszą osiągać zera.

Regularyzacja Lasso (ang. *Lasso regression*)

Metoda najmniejszych kwadratów z **regularyzacją Lasso** (inaczej $l1$), minimalizuje funkcję kryterialną:

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2 + \alpha \sum_{i=1}^{p-1} |\beta_i|,$$

gdzie α jest parametrem sterującym metody (siłą regularyzacji), $\alpha \geq 0$ oraz $\sum_{i=1}^{p-1} |\beta_i| = ||\beta||_1^2$ jest normą $l1$.

Uwaga: Dla dostatecznie dużych wartości α estymacje niektórych parametrów β_i przyjmują wartości równe 0, stąd regularyzację Lasso często wykorzystuje się do selekcji zmiennych niezależnych.

Porównanie regularyzacji grzbietowej z regularyzacją Lasso

Ridge

- zawiera wszystkie (lub żadne) cechy w modelu, główną zaletą tej regularyzacji jest **ściągniecie współczynników** (ang. *shrinkage coefficient*),
- używana się głownie do **uniknięcia przeuczenia** modelu, ale z racji, że zawiera wszystkie zmienne z modelu nie jest użyteczna w przypadku, gdy liczbę predyktorów szacuje się milionach/miliardach – zbyt duża złożoność obliczeniowa,
- działa dobrze w obecności **silnie skorelowanych cech**.

Lasso

- regularyzacja Lasso poza **ściągniecie współczynników**, dokonuje również **selekcji zmiennych**
- często wykorzystywana się do **selekcji zmiennych** w przypadku gdy liczba cech jest rzędu milionów/miliardów.

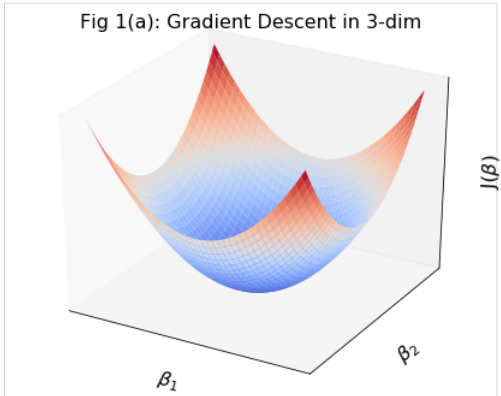
Inny sposób przedstawienia regularyzacji grzbietowej i Lasso

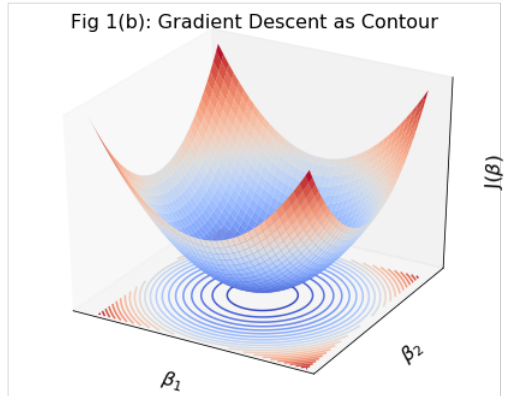
Metoda najmniejszych kwadratów z **regularyzacją grzbietową**, minimalizuje funkcję kryterialną:

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2 + \alpha\sum_{i=1}^{p-1}\beta_i^2 \quad \text{gdy} \quad \sum_{i=1}^{p-1}\beta_i^2 < s.$$

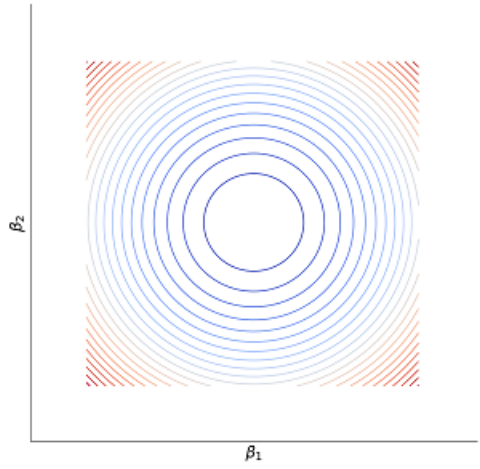
Metoda najmniejszych kwadratów z **regularyzacją Lasso**, minimalizuje funkcję kryterialną:

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2 + \alpha\sum_{i=1}^{p-1}|\beta_i| \quad \text{gdy} \quad \sum_{i=1}^{p-1}|\beta_i| < s.$$





Źródło: https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd



Źródło: https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd

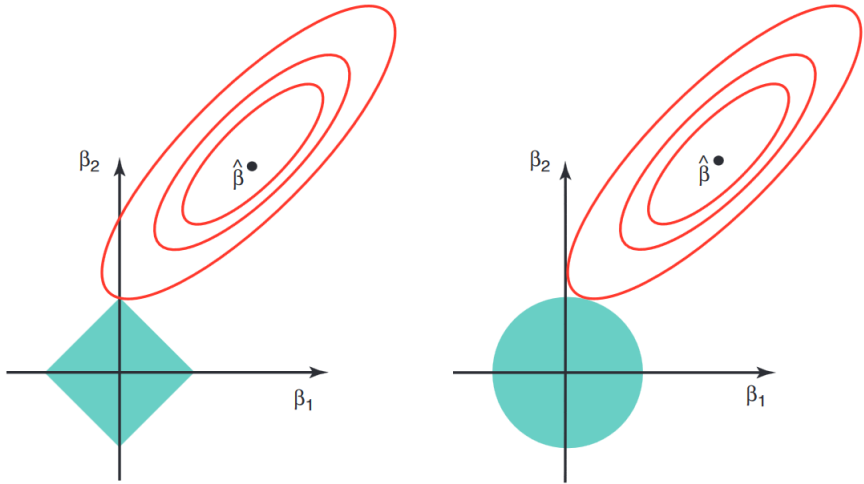


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Regularyzacja Elastic Net (ang. *Elastic Net regression*)

Metoda najmniejszych kwadratów z **regularyzacją Elastic Net**, minimalizuje funkcję kryterialną:

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2 + \alpha \cdot l1_{ratio} \cdot \sum_{i=1}^{p-1}|\beta_i| + \frac{1}{2}\cdot\alpha\cdot(1 - l1_{ratio})\cdot\sum_{i=1}^{p-1}\beta_i^2.$$

Aby kontrolować karę $l1$ i $l2$ powyższe jest równoważne z

$$a \cdot ||b||_1^2 + b \cdot ||b||_2^2,$$

gdzie $\alpha = a + b$ i $l1_{ratio} = \frac{a}{a+b}$.

Wybór podzbioru zmiennych

Jednym ze sposobów na uniknięcie przeuczenia modelu jest zbudowanie go na mniejszej liczbie zmiennych.

Jak zredukować liczbę zmiennych w modelu?

- wybór modelu o najniższej wartości kryterium AIC,
- wybór modelu o najniższej wartości kryterium BIC.

Kryteria AIC i BIC

Kryterium Akaike (ang. *Akaike Information Criterion*):

$$\mathrm{AIC} = n \cdot \ln\left(\frac{\mathrm{SSE}}{n}\right) + 2p$$

Kryterium Schwarza (ang. *Bayesian Information Criterion*):

$$\mathrm{BIC} = n \cdot \ln\left(\frac{\mathrm{SSE}}{n}\right) + \ln(n)p,$$

gdzie p jest liczbą parametrów w modelu, n jest liczbą obserwacji, a L jest funkcją wiarygodności.

Metody krokowe

Mając dany układ cech dodajemy lub usuwamy jedną cechę, tj. dodajemy cechę nie występującą obecnie w modelu którą w danym momencie uważamy za właściwą, lub usuwamy cechę występującą w modelu, jeżeli uznamy ją w danym momencie za niewskazaną.

- **forward selection** — jest to metoda, która polega na stopniowym dołączaniu do modelu kolejnych zmiennych;
- **backward selection** — jest to metoda, która polega na stopniowym usuwaniu z modelu kolejnych zmiennych.

Materiały na podstawie

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R , Springer 2014.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer 2009.