

Regularyzacja i dobór stopnia złożoności modelu

Joanna Jaworek-Korjakowska

WEAIIB, Katedra Automatyki i Robotyki, ISS

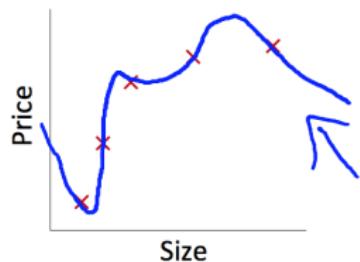
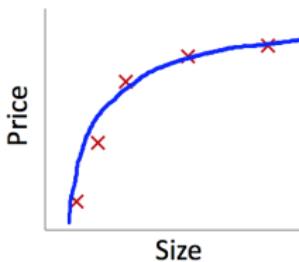
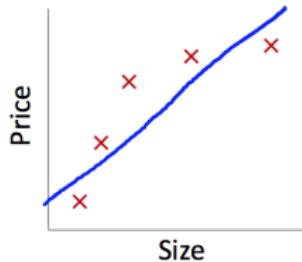
2018/2019

Zdolność do uogólniania

Naszym celem jest znalezienie funkcji $f_S : X \rightarrow Y$, która jest przybliżeniem funkcji $f : X \rightarrow Y$ opisującej rzeczywistą zależność między X i Y

$$\frac{d}{d\theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Example: Linear regression (housing prices)



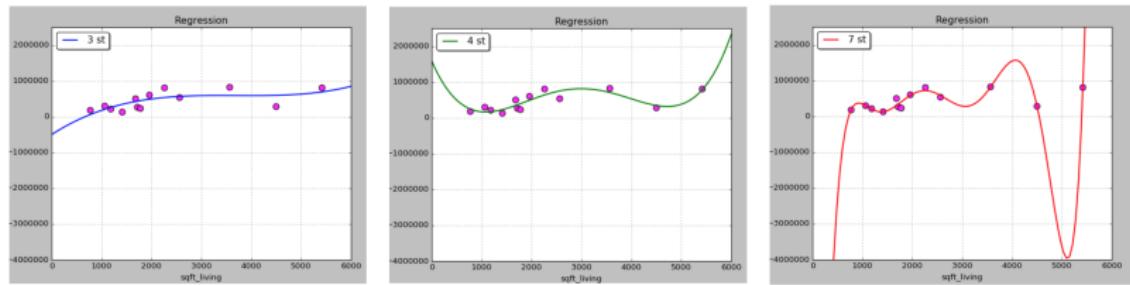
Jednym z podstawowych wymogów dotyczących funkcji f_S jest zdolność do uogólniania (generalization). Rozumiemy przez to, że różnica między błędem empirycznym a oczekiwany dąży do zera, kiedy rozmiar zbioru uczącego dąży do nieskończoności.

Przetrenowanie i regularyzacja

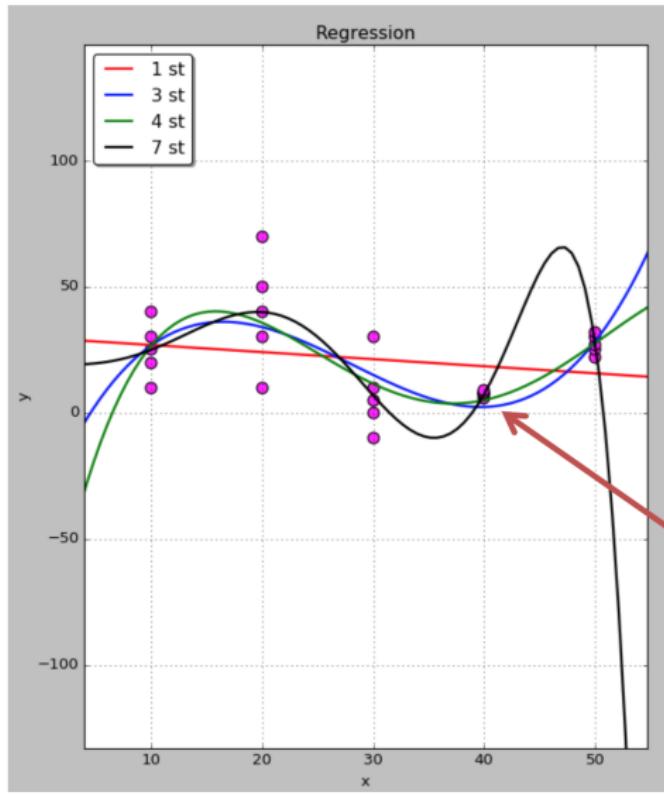
Minimalizując błąd empiryczny na zbiorze uczącym S może się pojawić problem przetrenowania (ang. *overfitting*). Zjawisko to polega na tym, że znaleziona funkcja f_S bardzo dobrze pasuje do danych S , ale kosztem zdolności do uogólniania.

Regularyzacja rozwiązuje ten problem. Mówiąc obrazowo, dodajemy karę za skomplikowaną postać funkcji f_S .

Porównanie – która krzywa jest najlepsza?



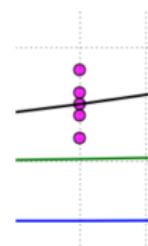
Czy błąd uczenia można zredukować do zera?



Model	MSE
1 stopnia	280.74
3 stopnia	164.23
4 stopnia	147.38
7 stopnia	137.91

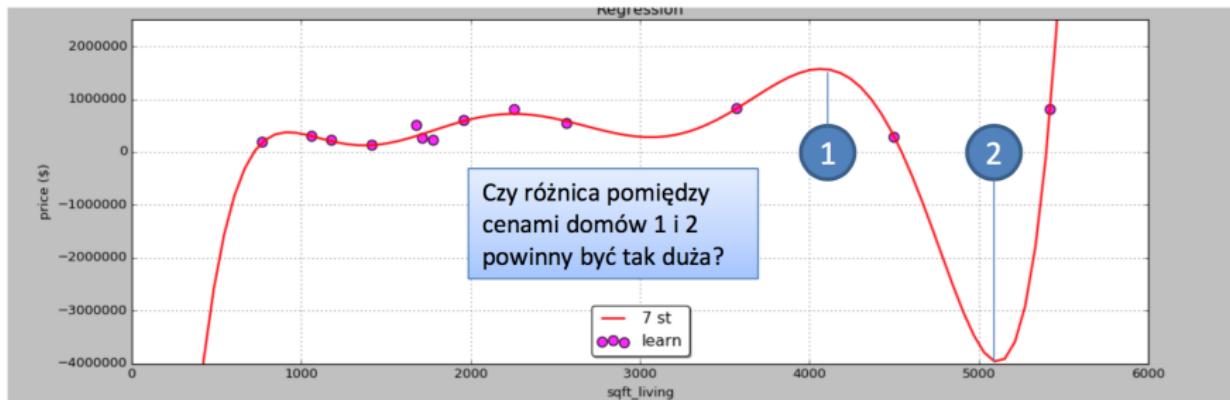
W ogólnym przypadku nie

Jeżeli zbiór uczący zawiera punkty przypisujące tym samym x różne wartości wyjściowe y , błąd uczenia zawsze pozostanie



Zastosowanie modelu do predykcji

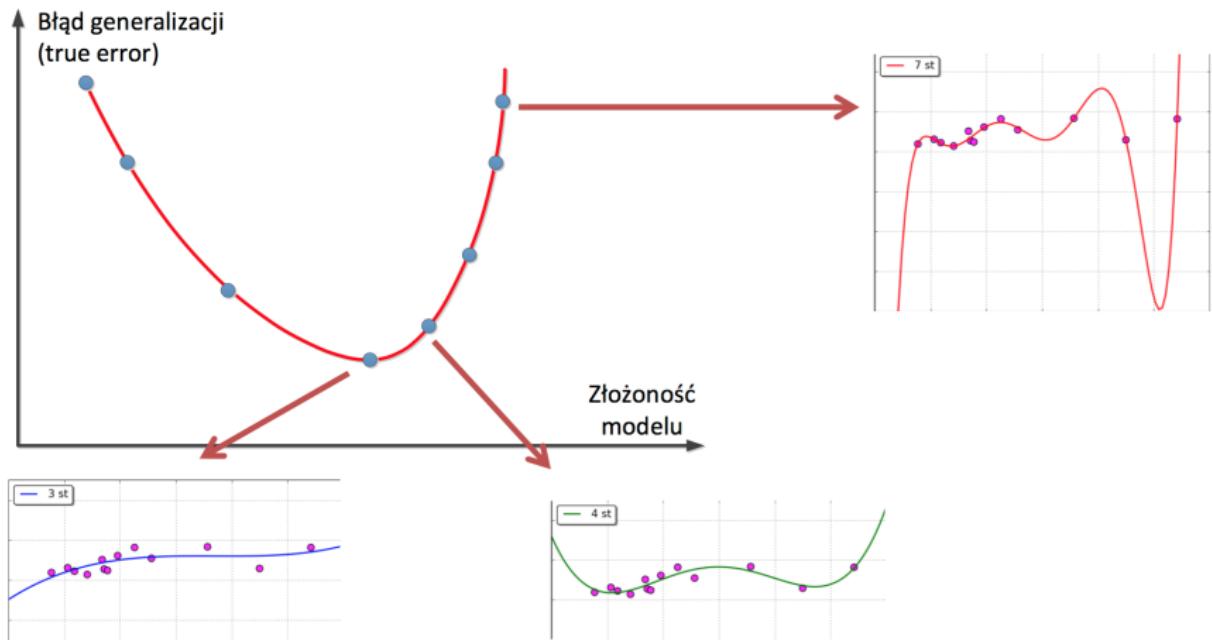
- Podczas predykcji model jest stosowany, aby przewidzieć wartość wyjściową dla nieznanych danych.
- Czy błąd uczenia pozwala ocenić jakość modelu podczas predykcji?



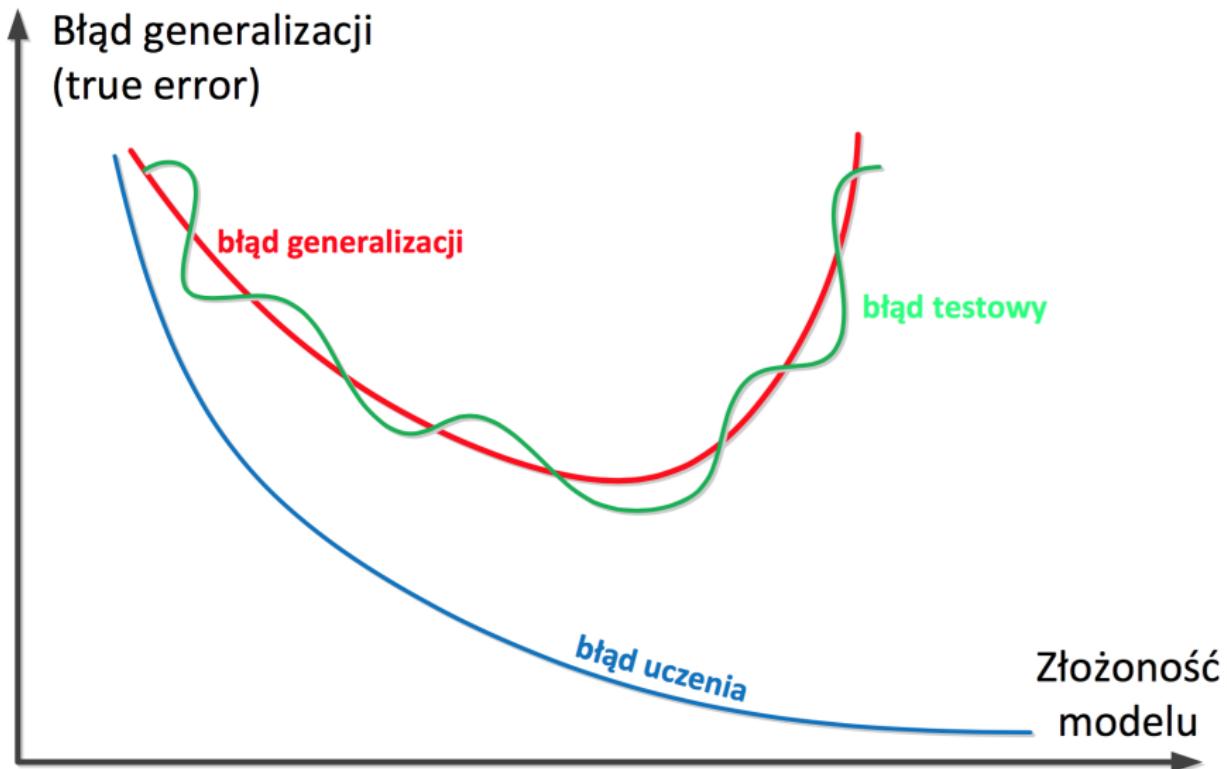
- Błąd uczenia jest bardzo optymistycznym oszacowaniem, ponieważ w wyniku optymalizacji krzywa została dopasowana do danych uczących.
- **Niska wartość błędu uczenia nie pociąga za sobą dobrych własności predykcyjnych**
(chyba, że model był uczyony na wszystkich kombinacjach danych wejściowych)

Błąd generalizacji

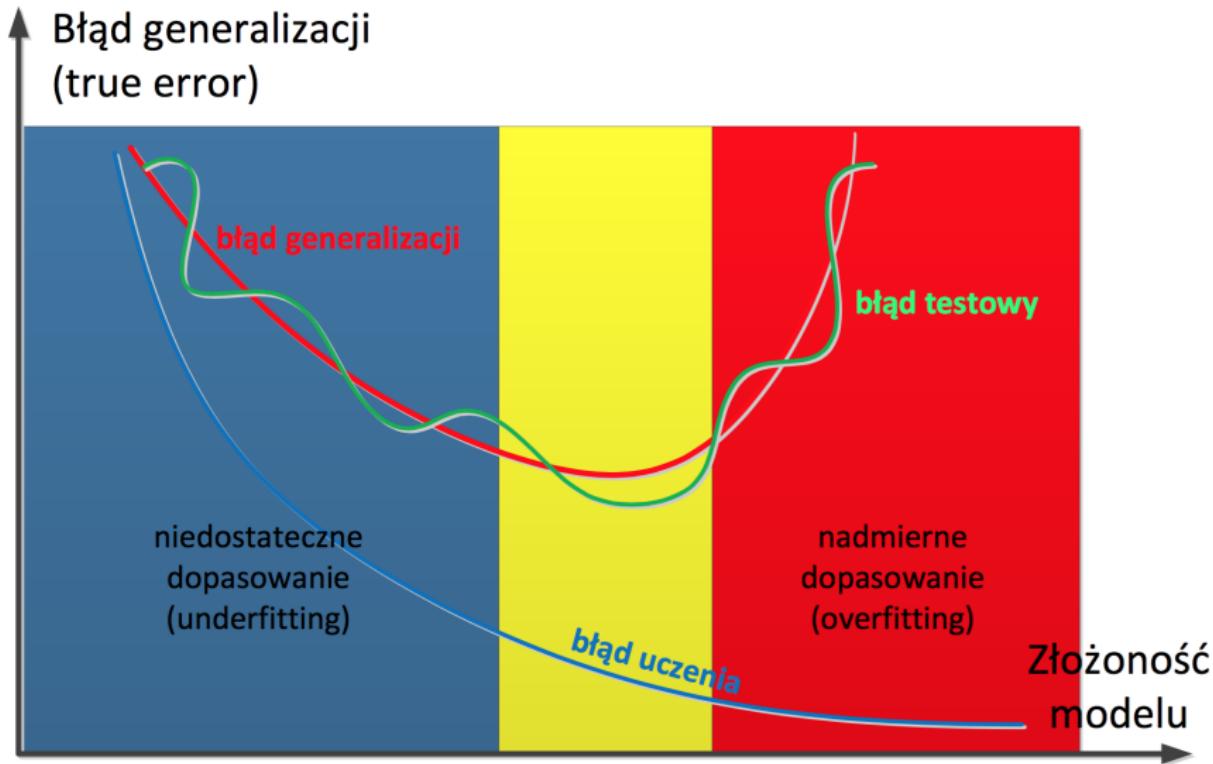
Na ogół wraz ze wzrostem złożoności „prawdziwy błąd” modelu początkowo maleje, a następnie rośnie.



Błąd uczenia, testowy i generalizacji



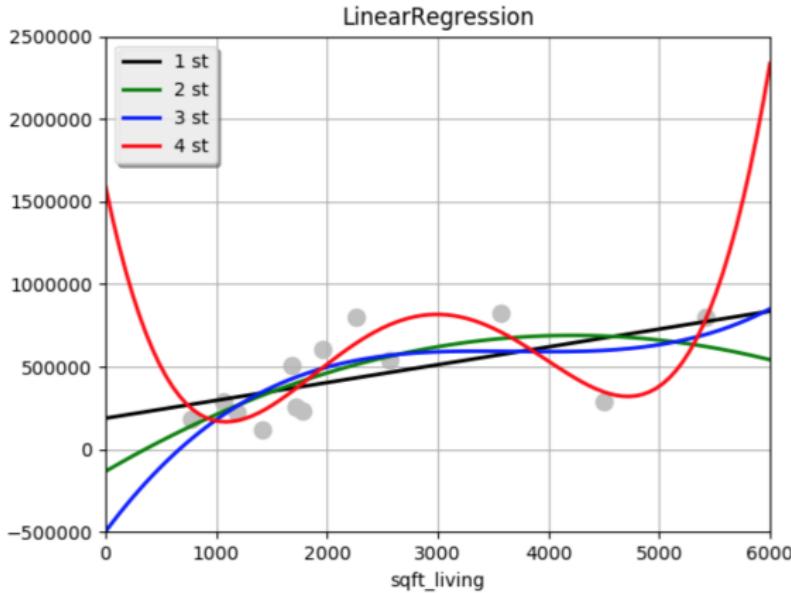
Niedostateczne i nadmierne dopasowanie



Kompromis pomiędzy bias i variance

- Wybór modelu w problemach uczenia nadzorowanego (nie tylko regresji) wiąże się z realizacją dwóch sprzecznych celów:
 - Model powinien być dobrze dopasowany do danych uczących, aby uchwycić zależności pomiędzy danymi
 - Model powinien też dobrze przybliżać nieznane dane (zapewniać mały błąd generalizacji)
- Modele złożone dobrze dopasowują się do danych wyjściowych, ale charakteryzują się dużą zmiennością (**variance**) wartości wyjściowych. Ryzykiem jest nadmierne dopasowanie (**overfitting**)
- Modele prostsze są obciążone dużym błędem systematycznym (**bias**) i ich zastosowanie niesie ryzyko niewystarczającego dopasowania (**underfitting**).
- Trzecim składnikiem błędów generalizacji jest nieredukowalny błąd związany ze **zmiennością danych**

Regularyzacja



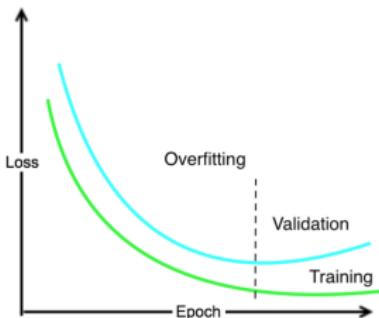
Dla regresji wielomianowej algorytm będzie starał się dobrze dopasować krzywą do obserwacji.

Im większy stopień wielomianu tym więcej jest możliwych zagięć krzywej
Równocześnie krzywizny stają się coraz „chudsze”, co z reguły odpowiada wysokim wartościom wag

Dlaczego duże wagi są złe?

- Dla regresji wielomianowej jednej zmiennej nie jest to oczywiste – cechy są dość mocno skorelowane, więc wpływ dużej wagi w_i może skorygowany przez wpływ innej dużej wagi w_j
- Dla niezależnych cech – duża wartość wagi w_i oznacza dużą wrażliwość funkcji regresji na drobne fluktuacje i -tej cechy $h_i(x)$
 - Dla blisko położonych obserwacji x^k oraz x^l różnice wartości funkcji $f(x^k)$ i $f(x^l)$ mogą być bardzo duże.
 - Model bardzo dobrze dopasowany do danych uczących może nie sprawdzić się dla nieznanych danych
- Lepszym rozwiązaniem jest gorsze dopasowanie do danych uczących przy równoczesnym ograniczeniu parametrów świadczących o potencjalnie dużym błędzie generalizacji – czyli zmniejszeniu wartości wag.

Bias vs. variance



Źródło: [1]

Bias

Duży błąd zarówno na zbiorze testowym jak i walidacyjnym

- zbyt prosty model
- zbyt mało cech

Variance

Błąd na zbiorze walidacyjnym wyraźnie większy niż na zbiorze uczącym

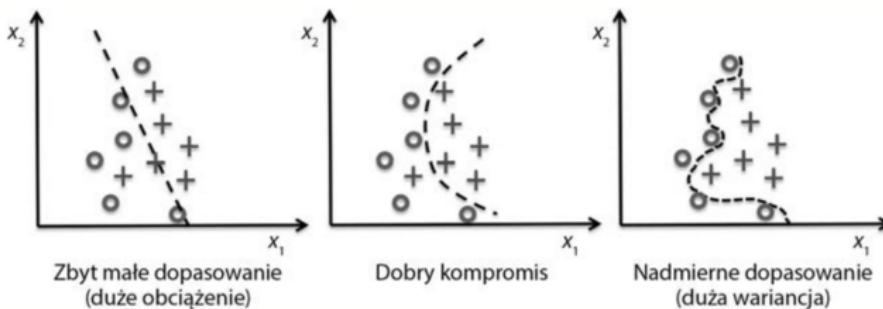
- nadmierne dopasowanie modelu
- zbyt złożony model
- za wiele cech

Jak regularyzować modele regresyjne?

Co zrobić by nie przeuczać?

Standardowo dla modeli regresyjnych stosuje się trzy strategie:

- Wybór zmiennych (subset selection), np. z użyciem kryteriów jednowymiarowego filtrowania.
- Stosowanie regularizacji opartej o karę za wielkość współczynników w modelu,
- Redukcja wymiaru zmiennych predykcyjnych przez zastosowanie techniki PCA/PCR.



Problem dopasowania

- zbyt wiele cech
- zbyt złożony model

Efekt

Dobry wynik na danych uczących i wyraźnie gorszy dla nowych przykładów

Rozwiązań:

- ① Ograniczenie złożoności modelu
 - ▶ redukcja liczby cech
 - ▶ zmiana typu lub meta-parametrów modelu
- ② Regularyzacja
 - ▶ ℓ_2
 - ▶ ℓ_1
 - ▶ Elastic net

Regularyzacja

Innym sposobem ograniczenia wariancji modelu regresyjnego jest ograniczenie wartości współczynników tego modelu.

Ograniczać można normę L_2 wektora współczynników (rozwiązywanie znane pod nazwą **regresja grzbietowa** (ang. **ridge regression**))

Normę L_1 wektora współczynników (rozwiązywanie znane pod nazwą **LASSO**)

Mieszankę tych norm (rozwiązywanie znane pod nazwą **sieci elastycznych**)

Regularizacja L2 (Ridge Regression)

- W przypadku zwykłej regresji liniowej szukane są wagie minimalizujące funkcję celu postaci

$$J(w) = \sum_{i=1}^m (y_i - w^T x_i)^2$$

- Dla regresji grzbietowej dodany jest składnik (funkcja kary) ograniczający wartości wag

$$J(w) = \sum_{i=1}^m (y_i - w^T x_i)^2 + \lambda \|w\|^2$$

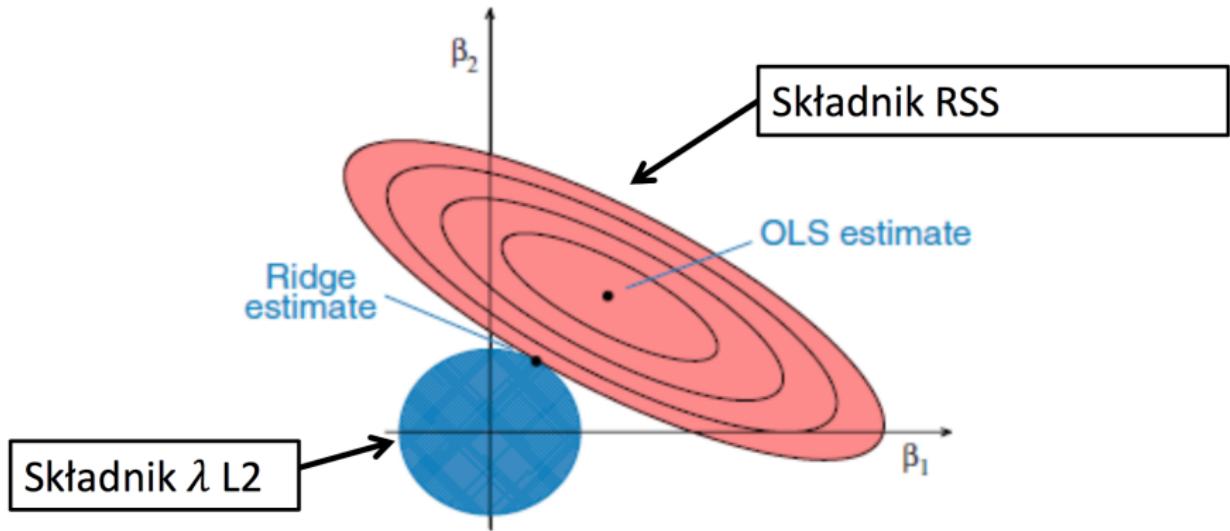
- Czynnik $\|w\|^2$ to tzw. norma L^2 , czyli po prostu:

$$w^T w = \sum_{i=1}^n w_i^2$$

- Nieujemna stała λ określa udział składnika $\|w\|^2$ w funkcji celu.

Regularizacja L2 -analiza Lambda

$$J(w) = \sum_{i=1}^m (y_i - w^T x_i)^2 + \lambda \|w\|^2$$



Źródło: <https://onlinecourses.science.psu.edu/stat857/node/155>

Tutaj: β_1 i β_2 to wagи w

Nazywane również **ridge**

Funkcja kosztu:

$$J(w) = \text{błąd jak dotychczas} + \lambda \sum_{i=1}^n w_i^2$$

Metoda gradientu:

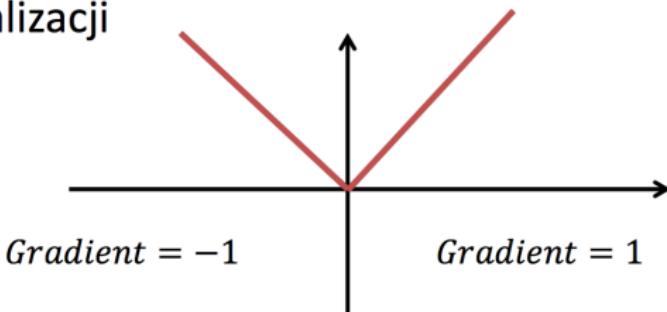
$$\frac{\partial J}{\partial w_i} = \text{gradient jak dotychczas} + 2w_i$$

Uzyskujemy mniejsze wartości bezwzględne parametrów modelu

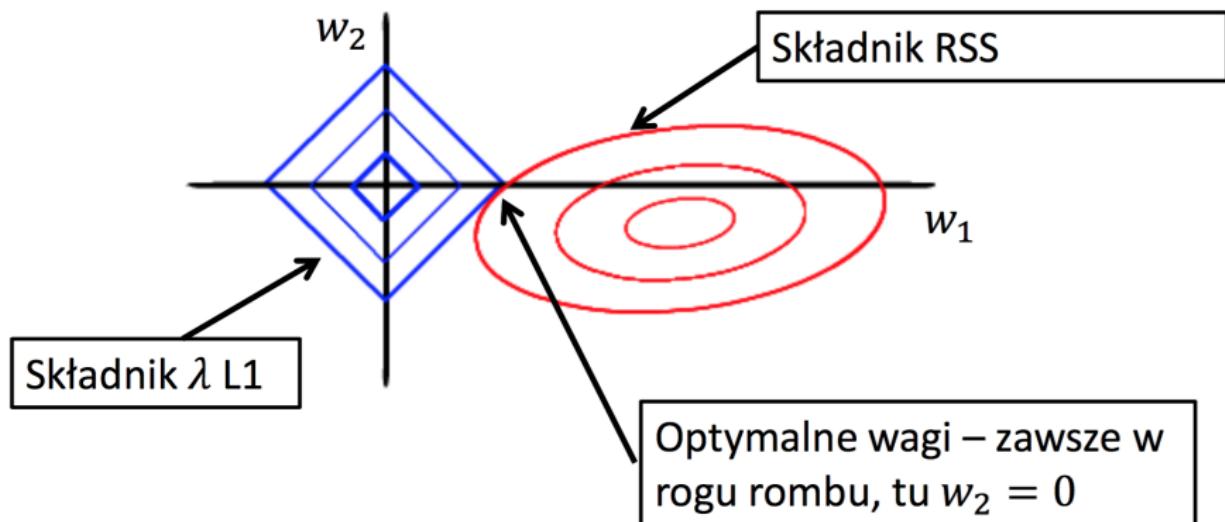
- W przypadku regularizacji L1 składnikiem regularyzującym jest norma L1 (czyli suma wartości bezwzględnych wag)

$$J(w) = \sum_{i=1}^m (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^n |w_i|$$

- Funkcja ta nie jest różniczkowalna w zerze, więc nie istnieje rozwiązanie analityczne
- Mimo tego można zastosować gradientowe metody optymalizacji



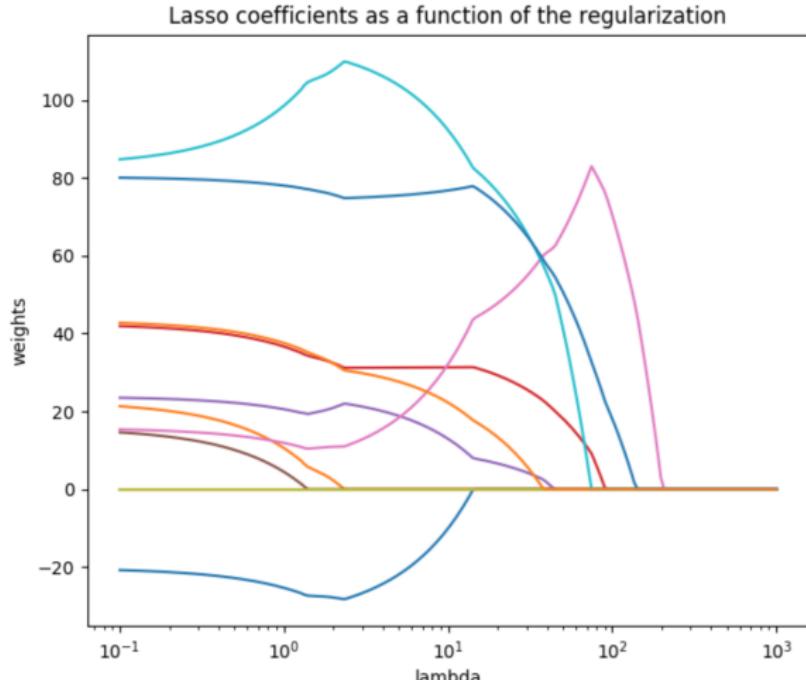
$$J(w) = \sum_{i=1}^m (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^n |w_i|$$



Źródło: <https://stats.stackexchange.com/questions/30456/geometric-interpretation-of-penalized-linear-regression>

Lasso

- Dla L1 wraz ze wzrostem lambda kolejne wagi będą znikać (przyjmować wartość 0)
- Dla L2 wagi będą stawały się dowolnie małe, ale nie zanikają zupełnie



Lasso działa więc jak mechanizm wyboru cech (ang. feature selection).

Stopniowo odrzuca współliniowe atrybuty, pozostawia zbiór najbardziej istotnych (tych, które najlepiej „objaśniają” zmienność wartości wyjściowych).

Lasso

Nazywane również **LASSO** (least absolute shrinkage and selection operator)

Funkcja kosztu:

$$J(w) = \text{błąd jak dotychczas} + \lambda \sum_{i=1}^n |w_i|$$

Metoda gradientu:

$$\frac{\partial J}{\partial w_i} = \text{gradient jak dotychczas} + \text{sign}(w_i)$$

Uzyskujemy część wartości parametrów równe **dokładnie 0**

Połączenie l1 i l2

Funkcja kosztu:

$$J(w) = \text{błąd jak dotychczas} + \lambda\beta\sum_{i=1}^n |w_i| + 0.5(1 - \beta)\lambda\sum_{i=1}^n w_i^2$$

Inne metody:

- dropout
- batch normalization
- wcześniejsze przerywanie uczenia
- ograniczanie rozmiaru drzew decyzyjnych