Metody analizy danych

Agnieszka Sołtys

O mnie

- Absolwentka UW: Matematyka na MIM i Informatyka i Ekonometria na WNE
- Doktorat w IPI PAN
- Staże: Columbus Ohio, ICM, WUM
- Zajęcia: UW, PW, uczelnie prywatne
- Data Scientist: modele predykcyjne, segmentacja, pozyskanie i przygotowanie danych

Sztuczna inteligencja

Sztuczna inteligencja, SI (ang. artificial intelligence – AI)[1] – inteligencja wykazywana przez urządzenia sztuczne. W informatyce oznacza także tworzenie modeli i programów symulujących choć częściowo zachowania inteligentne.
`https://openai.com/blog/chatgpt`

[1] Źródło: Wikipedia

Uczenie maszynowe

Uczenie maszynowe (ang. machine learning)[2]:

- Obszar sztucznej inteligencji poświęcony algorytmom, które poprawiają się automatycznie poprzez doświadczenie, czyli ekspozycję na dane.
- Algorytmy uczenia maszynowego budują model matematyczny na podstawie przykładowych danych, zwanych zbiorem uczącym, w celu prognozowania lub podejmowania decyzji bez bycia zaprogramowanym explicite przez człowieka do tego celu.

`https://www.youtube.com/watch?v=V1eYniJ0Rnk`

[2] Źródło: Wikipedia

Uczenie maszynowe - cykl pracy

- Zdefiniowanie zadania. Jaki model wybrać?
- Jak ocenić sukces?
- Przygotowanie danych!

Źród\OT4\lo:https://jonwood.co/blog/2021/5/17/how-the-machine-learning-process-is-like-cooking

Uczenie maszynowe - dwa główne nurty

1. **Uczenie nadzorowane (ang. supervised learning)** - dane zawierają etykiety (ang. labels):
   - zmienne objaśniające $x_1, \ldots, x_p$,
   - objaśniane y (etykiety).

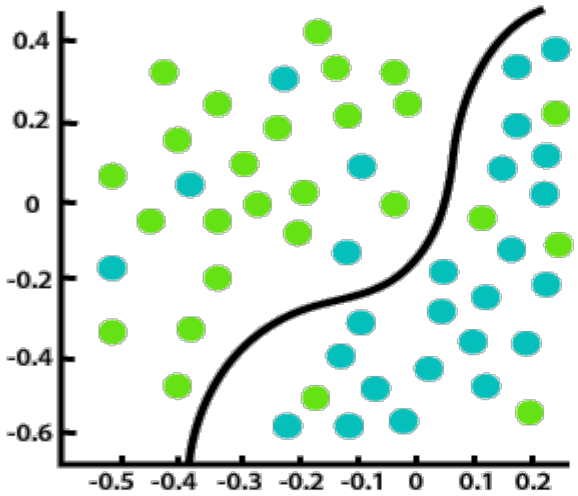
   Np. scoring kredytowy - przewidywanie, czy klient ubiegający się o kredyt spłaci go w terminie.
2. **Uczenie nienadzorowane (ang. unsupervised learning)**- dane nie zawierają etykiet:
   - odkrywanie wzorców w danych $x_1, \ldots, x_p$.

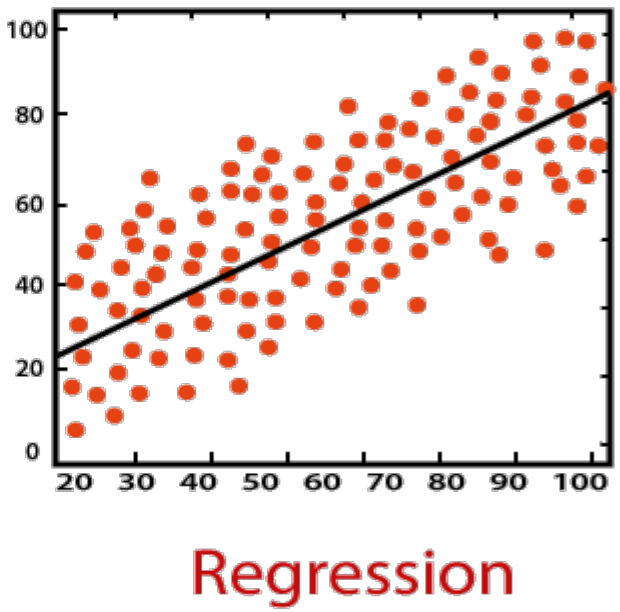
   Np. segmentacja klientów w celu dopasownia odpowiedniej oferty.

Klasyfikacja i regresja

Dwa najważniejsze zadania uczenia ndzorowanego to:

1. klasyfikacja: $y \in \{0, 1\}$
2. regresja: $y \in \mathbb{R}$





Źród\OT4\lo:https://www.javatpoint.com/regression-vs-classification-in-machine-learning

Plan zajęć

1. Uczenie nadzorowane:
   - modele liniowe - regresja liniowa i logistyczna, regularyzacja
   - drzewa decyzyjne i lasy losowe
   - SVM
   - sieci neuronowe
2. Uczenie nienadzorowane:
   - transformacje danych, np. skalowanie, analiza składowych głównych (PCA)
   - analiza skupień

Problemy uczenia maszynowego

Twierdzenie o nieistnieniu darmowych obiadów (ang. no free lunch theorem)
Żaden algorytm uczenia maszynowego nie będzie działał najlepiej dla wszystkich zbiorów danych.

Dekompozycja wariancja-obciążenie (ang. bias-variance tradeoff)
Najlpeszy pod względem predykcji na nowych danych jest model, który jest ani zbyt skomplikowany, ani zbyt prosty (trzeba znaleźć kompromis pomiędzy złożonością a dopasowaniem).

Problemy uczenia maszynowego

1. niedobór danych - np. w zastosowaniach medycznych
2. niereprezentatywne dane - np. budujemy model scoringowy dla banku, ale liczba obserwacji, gdzie klient spłacił kredyt jest dużo większa niż tych, gdzie klient kredytu nie spłacił
3. dane kiepskiej jakości - zawierające dużo błędów, obserwacji odstających czy szumu
4. często potrzebna jest wiedza dziedzinowa - szczególnie przy zbieraniu danych, np. jakie cechy mogą wpływać na zachorowanie na daną chorobę?

Przede wszystkim JAKOŚĆ DANYCH!

Zaliczenie

- Dwa zestawy zadań, pierwszy 28. maja, drugi 18. czerwca.
- Na każdy zestaw zadań będzie tydzień, omówienie na zajęciach.
- Każdy zestaw punktowany po 10 punktów, zaliczenie od 11 punktów.

Regresja

Przykłady zastosowań

- predykcja ceny wynajmu mieszkania w zależności od jego cech (wielkości, położenia, wyposażenia),
- predykcja ilości nawozu w zależności od położenia i cech pola uprawnego,
- predykcja wydatków gospodarstw domowych,
- predykcja dawki leku w zależności od cech pacjenta
- i wiele, wiele innych!

Regresja liniowa

- Za pomocą zmiennych objaśniających $x_1, x_2, \ldots, x_p$, chcemy wyjaśnić wartość zmiennej objaśnianej y.
- Robimy to za pomocą linii prostej (hiperpłaszczyzny).
- Wzór do prognozowania dla modelu liniowego:

$$\widehat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_p \cdot x_p = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_j,$$

gdzie β_0, β_1, $\ldots$, β_p są parametrami modelu.

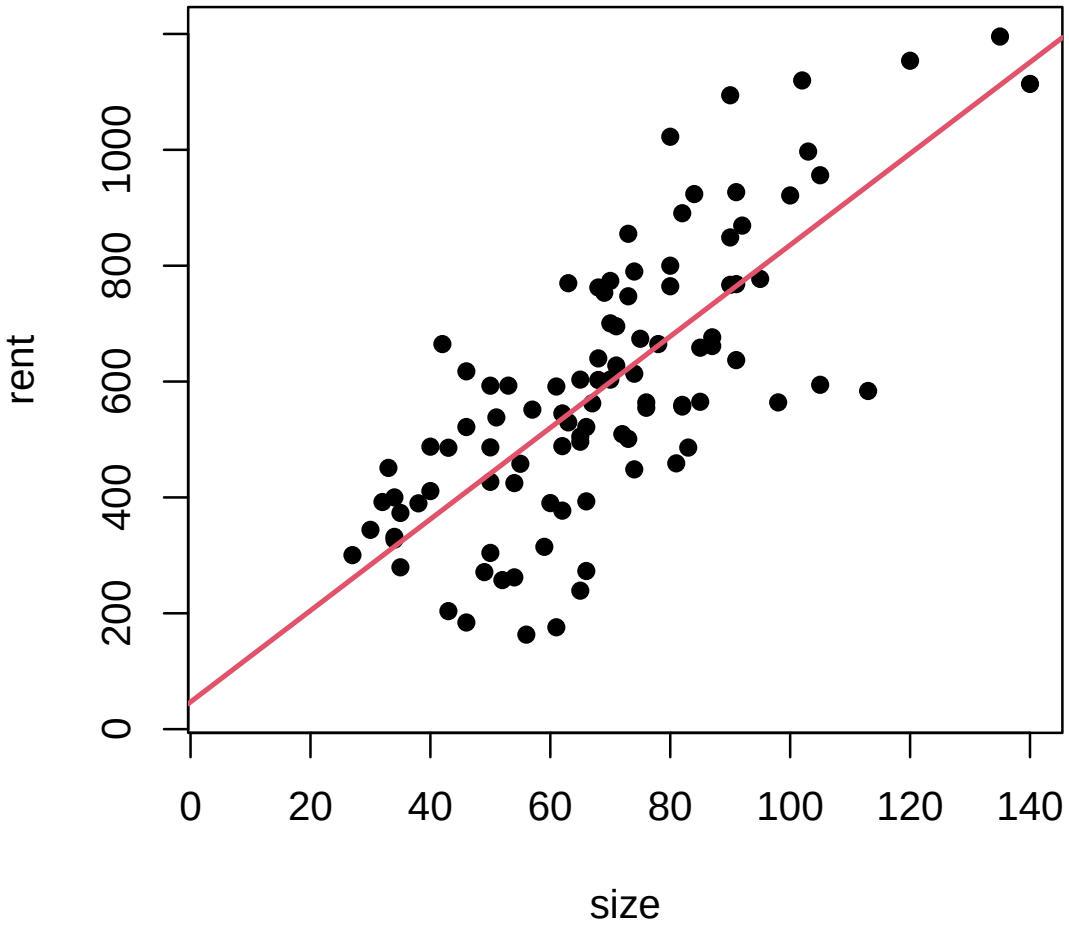
- Dla dwóch wymiarów mamy równanie prostej:

$$\widehat{y} = \beta_0 + \beta_1 \cdot x_1$$

Regresja liniowa

$$\widehat{rent} = \beta_0 + \beta_1 \cdot size = 47 + 7.9 \cdot size$$

	rent	size
1	741.39	68
2	715.82	65
3	528.25	63
4	553.99	65
5	698.21	100
6	935.65	81
...	...	...



Regresja liniowa - metoda najmniejszych kwadratów

n - liczba obserwacji

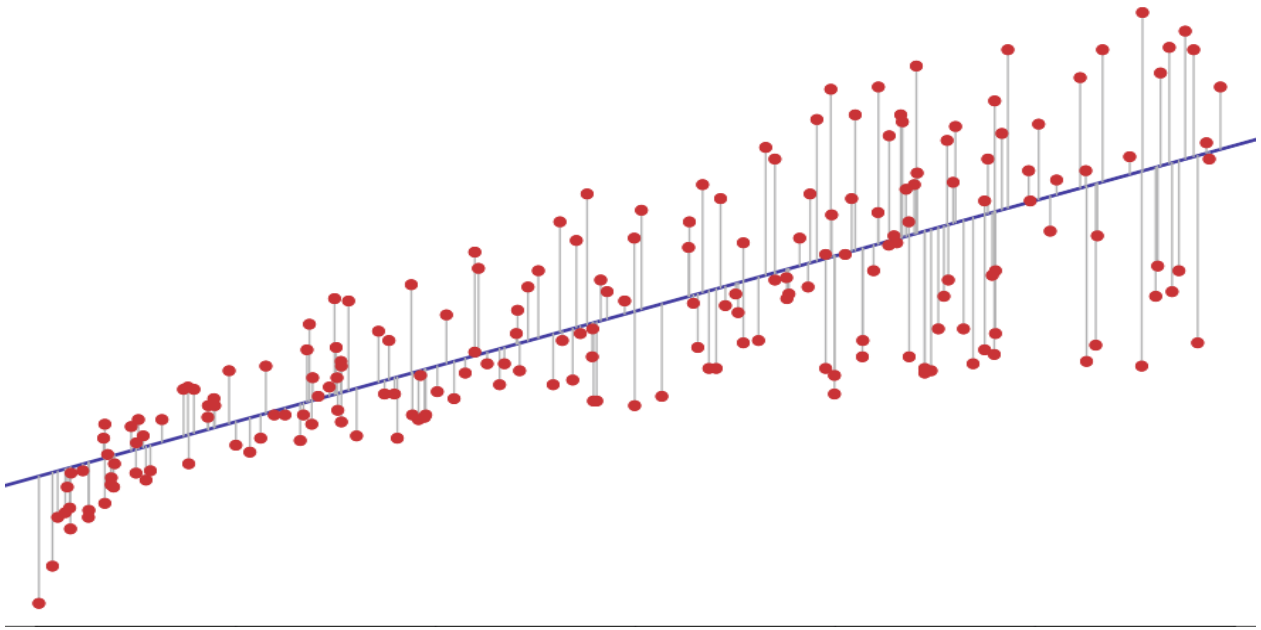
$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}, x_1 = \begin{pmatrix} x_{11} \\ \dots \\ x_{1n} \end{pmatrix}, \dots, x_p = \begin{pmatrix} x_{p1} \\ \dots \\ x_{pn} \end{pmatrix}.$$

Regresja liniowa - metoda najmniejszych kwadratów

Współczynniki β znajdujemy tak żeby minimalizować *RSS* (residual sum of squares):

$$RSS = (y_1 - \hat{y}_1)^2 + \ldots + (y_n - \hat{y}_n)^2 =$$

$$= (y_1 - \beta_0 - \beta_1 \cdot x_{11} - \ldots - \beta_p \cdot x_{p1})^2 + \ldots + (y_n - \beta_0 - \beta_1 \cdot x_{1n} - \ldots - \beta_p \cdot x_{pn})^2$$



Ocena jakości dopasowania

Współczynnik determinacji R^2:

- jest miarą jakości prognozy dla modelu regresji
- proporcja wariancji zmiennej objaśnianej wyjaśnionej przez model
- daje wynik od 0 do 1, wartość 1 odpowiada doskonałej prognozie, a 0 - modelowi stałemu, który tylko prognozuje średnią z obserwacji.

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS},$$

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

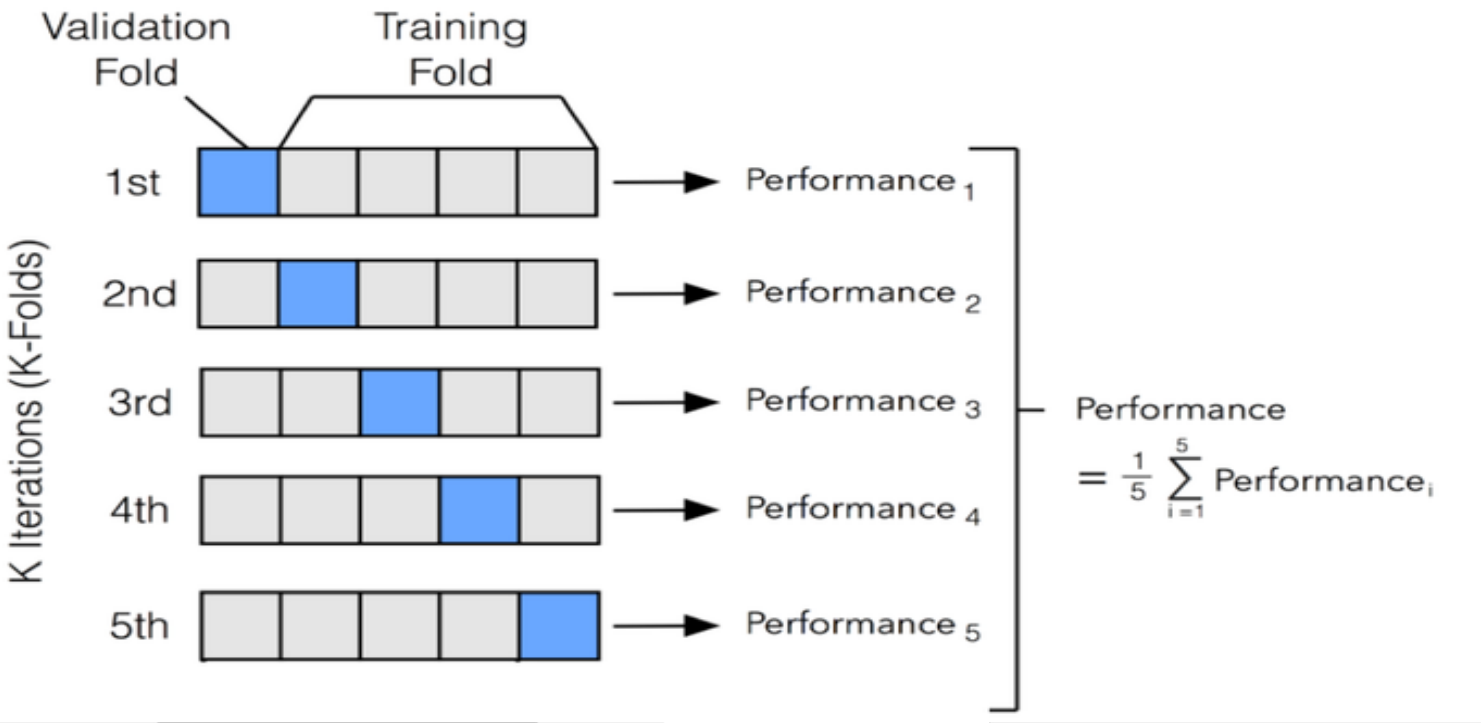
Zbiory treningowy i testowy

- Nie możemy do oceny modelu użyć tych samych danych, których użyliśmy do jego zbudowania!
- Model dopasowuje się do danych i ocena na tym samym zbiorze byłaby zbyt optymistyczna.
- Chcemy zmierzyć jak model się uogólnia (jak działa na nowych danych).

Podział zbioru na **treningowy** i **testowy**: na treningowym uczymy model, na testowym testujemy (oceniamy jakość dopasowania).

Kroswalidacja

- Kroswalidacja jest bardziej stablina i dokładna niż użycie pojedynczego podziału na zestaw uczący i testowy.
- Dane dzielone są wielokrotnie (k-krotnie) i budowanych jest wiele (k) modeli.



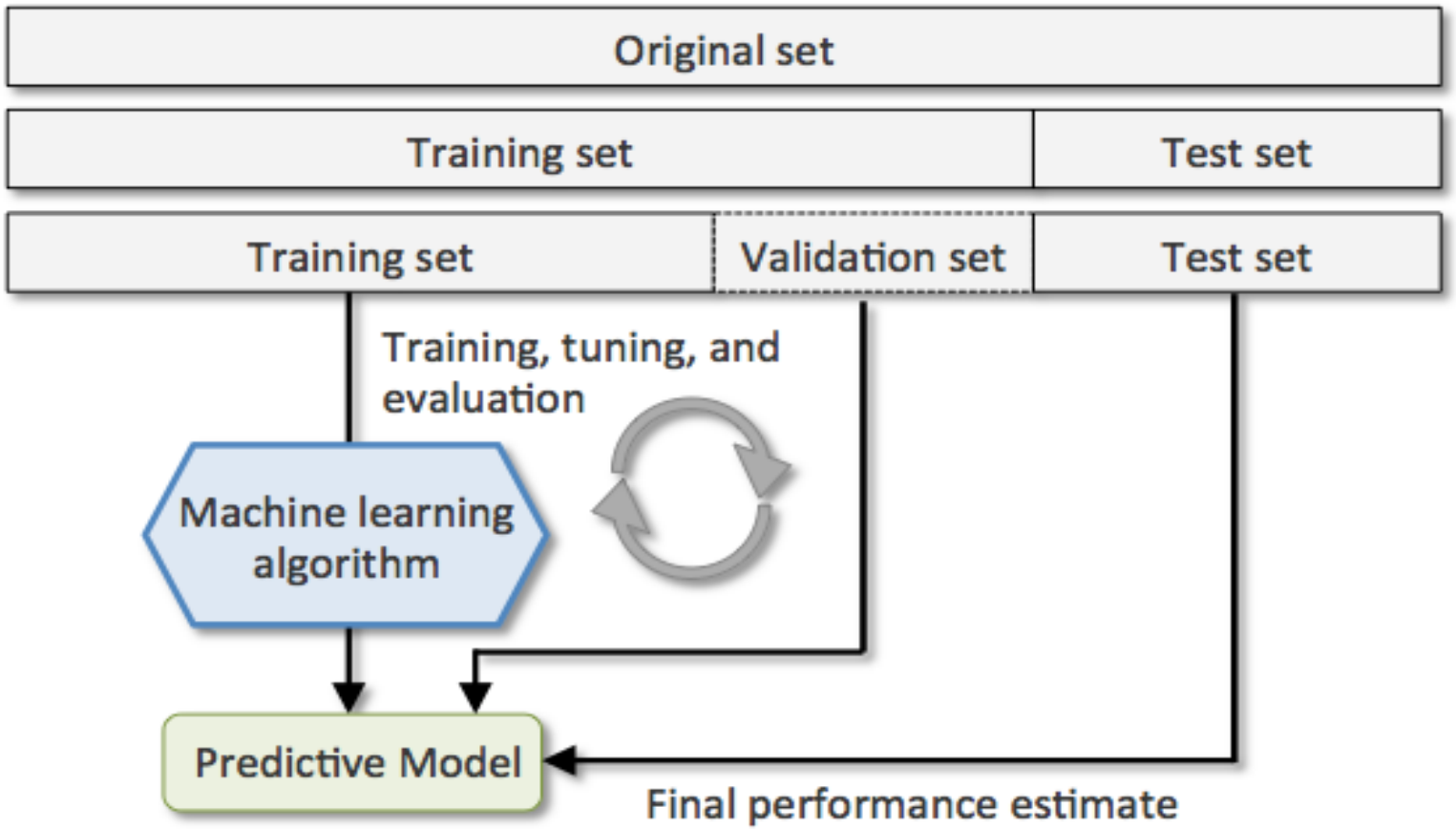
https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/

Kroswalidacja - uwagi

- Policzone błędy dla każdego k są na koniec uśredniane, dając ostateczne oszacownie.
- Najczęściej wybierane jest $k = 5$ lub $k = 10$.
- Każda obserwacja z danych znajdzie się w zestawie testowym dokładnie raz - uodporniamy się na „szczęśliwy" albo „pechowy" dobór próbki testowej.

Zestaw walidacyjny

Wybór najlepszego algorytmu i parametrów (najlepszego zestawu cech, regularyzacji).



https://stats.stackexchange.com/questions/410118/cross-validation-vs-train-validation-test

Regularyzacja

Dodanie ograniczeń dla parametrów β (ściąganie).

- Regresja grzbietowa: problem ze zwyczajną regresją liniową, gdy $x_1, \ldots, x_p$ są mocno skorelowane.

$$\widehat{\beta}^{ridge} = \arg\min_{\beta} RSS, \text{ pod warunkiem } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

- LASSO (least absolute shrinkage and selection operator): selekcja zmiennych

$$\widehat{\beta}^{LASSO} = \arg\min_{\beta} RSS, \text{ pod warunkiem } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- Elastic Net: połączenie regresji grzbietowej i LASSO

Typy zmiennych objaśniających

- ilościowe (ciągłe) - mierzalne, da się je uszeregować według określonej skali, np. wiek, wzrost, masa ciała,
- jakościowe - niemierzalne:
  - nominalne, np. kolor oczu (niebieskie, zielone, brązowe),
  - porządkowe, np. wykształcenie (podstawowe, średnie, wyższe)