Zaawansowane techniki tworzenia modeli

infoShareAcademy.com

Roadmapa:

- wartości odstające
- normalizacja (skalowanie / standaryzacja)
- kodowanie zmiennych
- kroswalidacja
- regularyzacja
- PCA
- optymalizacja hiperparametrów

infoShare ACADEMY

Preprocessing danych

- Usunięcie zmiennych o zerowej wariancji - jeżeli zmienna ma zerową wariancję to nie niesie żadnej wartości informacyjnej. Można rozważyć, czy nie należy także usuwać zmiennych o niezerowej, ale niskiej wariancji.

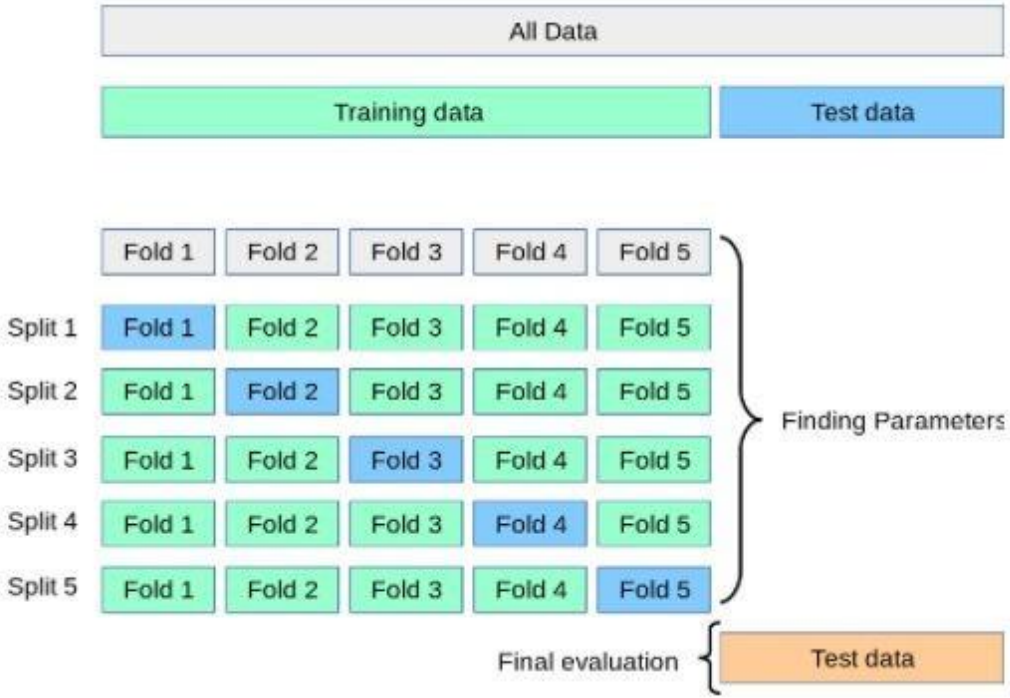
- Przekształcenie zmiennych:
  - zmienne kategoryczne:
    - one-hot encoding
    - labeling
  - zmienne liczbowe:
    - normalizacja
    - standaryzacja

- Imputacja braków danych

- Selekcja zmiennych:
  - Naturalną miarą służącą do pozytywnej selekcji jest korelacja zmiennych ze zmienną objaśnianą.
  - Istotność zmiennych
  - SHAP
  - LIME
  - inne

info Share ACADEMY

Walidacja krzyżowa (kroswalidacja)



infoShare ACADEMY

Regularyzacja

Technika stosowana w wielu algorytmach statystycznych dla eliminacji przeuczenia, które jest możliwe zwłaszcza w przypadku wielowymiarowych, ale nielicznych danych.

Regularyzacja działa poprzez wprowadzenie dodatkowego wyrażenia w formułach jako kary dla dużych wartości W, przy założeniu, że duże współczynniki występują w wysoce przeuczonych funkcjach.

info Share ACADEMY

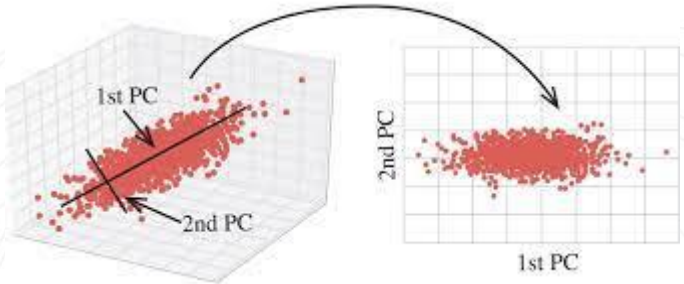
PCA - Analiza składowych głównych

- najbardziej popularny algorytm redukcji wymiarów
- polega na rzutowaniu danych do przestrzeni o mniejszej liczbie wymiarów tak, aby jak najlepiej zachować strukturę danych
- otrzymujemy zmienne nieskorelowane
- służy głównie do redukcji zmiennych opisujących dane zjawisko oraz odkrycia ewentualnych prawidłowości między cechami

info Share ACADEMY

PCA

- pozwala połączyć dużo zmiennych skorelowanych ze sobą w jedną zmienną zachowując wyjaśnianą wariancję na podobnym poziomie
- z oryginalnych zmiennych powstają nowe zmienne, które nie są skorelowane ze sobą ale są złożone z różnych zmiennych oryginalnych. Nowe zmienne są posortowane od wyjaśniającej najwięcej wariancji Y do najmniej wyjaśniającej
- w skrócie: liniowa kombinacja optymalnie zważonych zmiennych oryginalnych
- najczęściej wybierana jest taka liczba komponentów, która wyjaśnia co najmniej 80% wariancji danych
- **zmienne muszą być znormalizowane**
- **outliery muszą być rozpoznane**



info Share ACADEMY

Optymalizacja hiperparametrów

- GridSearch
- RandomSearch
- Hyperopt

infoShare ACADEMY

Dzięki!