Projekt

# MSO 2021/2022

## Informacje ogólne

* Projekty wykonujemy w zespołach 1-osobowych
* Termin wysłania projektu: 16 stycznia 2022 godzina 23:59:59
* Proszę wysłać projekt na adres e-mail prowadzącego grupę
* Wysyłamy plik zip o nazwie Imie\_Nazwisko.zip (bez polskich znaków) zawierający 2 pliki: kod programu (plik imie\_nazwisko.R) oraz plik w wynikami (plik: imie\_nazwisko.txt).

## Dane

1. Projekt dotyczy analizy zbioru danych „Communities and crime” które zawierają informacje o przestępczości w poszczególnych hrabstwach w USA.
2. Dane znajdują się w pliku: communities.data. Wczytywanie danych do R:

d = read.table("communities.data",h=F,sep=",")

1. Więcej informacji o poszczególnych zmiennych znajduje się w pliku: communities.names lub na stronie: <https://archive.ics.uci.edu/ml/datasets/communities+and+crime>
2. Celem jest prognozowanie zmiennej ViolentCrimesPerPop (liczba przestępstw na 100 K mieszkańców)- ostatnia kolumna w zbiorze danych.

## Polecenia:

1. Wczytaj dane do programu R.
2. Usuń pierwsze 5 kolumn (zmienne: state, county, community, communityname, fold).
3. Usuń wszystkie kolumny które zawierają braki danych (znak „?”).
4. Dokonaj podziału danych na dwa podzbiory: zbiór uczący U (obserwacje o parzystych indeksach) oraz zbiór testowy T (obserwacje o nieparzystych indeksach).
5. Dopasuj model regresji liniowej na danych U (model m).
6. Wykonaj diagnostykę modelu (na danych U), zaproponuj odpowiednie procedury diagnostyczne. Na ich podstawie dokonaj ewentualnie modyfikacji zbioru danych (usunięcie obserwacji, transformacje, itp.). Dopasuj nowy model, nazwijmy go m\_mod.
7. Dokonaj selekcji zmiennych bazując na modelu m oraz m\_mod. Zastosuje 2 podejścia: wybór zmiennych w oparciu o kryterium BIC oraz wybór zmiennych w oparciu o kryterium Mallowsa. Zastosuj metodę eliminacji wstecznej. Nazwijmy otrzymane modele: m\_BIC, m\_Mallows oraz m\_mod\_BIC, m\_mod\_Mallows).
8. Dokonaj selekcji zmiennych używając metody lasso. Parametr kary lambda wybierz używając walidacji krzyżowej. Otrzymany model nazwij m\_lasso.
9. Zaproponuj własny model (model nie omawiany na zajęciach) do prognozy. Preferowane jest użycie modelu nieliniowego. Nazwij go m\_myidea.
10. Dokonaj prognozy zmiennej celu na danych testowych T używając modeli:
    1. m
    2. m\_mod
    3. m\_BIC, m\_Mallows
    4. m\_mod\_BIC, m\_mod\_Mallows
    5. m\_lasso
    6. m\_myidea
    7. m\_empty (model zawierający tylko wyraz wolny)
11. Jakość prognozy oceń obliczając RMSE (root of mean squared error): .
12. Zapisz wyniki do pliku o nazwie: imie\_nazwisko.txt. Wyniki powinny być zapisane do tabeli, poniżej przykład:

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Liczba zmiennych w modelu** |
| m\_empty | 2.65 | 0 |
| m1 | 1.34 | 5 |
| m2 | 1.21 | 12 |
| … | … | … |