Projekt zaliczeniowy II
Statystyczna Analiza Danych

Termin oddania 12 czerwca.
Oddanie do 2 dni po terminie powoduje przyznanie co najwyżej 10 pkt.
Oddanie po 15 czerwca powoduje przyznanie 0 pkt.

Łącznie można zdobyć 15 punktów.

Dane
Dane do analizy należy pobrać z katalogu
https://www.mimuw.edu.pl/~szczurek/SAD1/ZadanieZaliczeniowe/Dane/
ściągając plik „GrupaX.zip" gdzie X to nr grupy laboratoryjnej.
Grupy 2 i 3 prowadzi Michał Ciach.
Grupę 4 prowadzi Grzegorz Głowienko.
Grupę 5 prowadzi Ania Macioszek.
Grupy 1 i 6 prowadzi Piotr Radziński.
Grupę 9 prowadzi Dorota Celińska-Kopczyńska.

Plik GrupaX.zip należy rozpakować. Dane czytamy w R poleceniem load("cancer.RData").
Wczytają się data.train i data.test.

i) Zbiór treningowy data.train zawiera, oprócz kolumn dla predyktorów, kolumnę Y oznaczającą działanie leku na nowotworowe linie komórkowe (zmienna o wartościach w przedziale [0,1]). Im niższa wartość Y, tym silniej lek działa na komórki danego typu raka (mniej komórek przeżywa po podaniu leku). Pozostałe kolumny to zmienne objaśniające (ekspresja genów w liniach komórkowych).
ii) Zbiór testowy data.test nie ma zmiennej Y. Na nim należy zastosować nauczony model.

Cel
Wybrać najlepszy model danych tak aby błąd testowy był jak najmniejszy.

Zadania
Zadania należy rozwiązać w języku R.

Zadanie 1 (1 pkt) (Analiza zmiennych objaśniających)

a) Podsumuj, ile zmiennych jest jakiego typu?
b) Wybierz 500 kolumn o największej zmienności. Policz korelację dla każdej z par tych kolumn. Zilustruj poziom współliniowości między tymi wybranymi kolumnami, rysując wykres rozkładu policzonych korelacji (wynikiem tego punktu ma być jeden wykres **geom_violin()** w ggplot2).

Zadanie 2 (2 pkt)
Poczytaj w podręczniku (*Elements of Statistical Learning* *https://web.stanford.edu/~hastie/Papers/ESLII.pdf*) na temat modelu elastic net (łącznie z Sekcją 18.4). Opisz jak działa ta metoda w oparciu o wprowadzone na wykładzie metody regresji grzbietowej i lasso. Podaj jakie ma parametry, które parametry są estymowane, a które *tuningowe,* i jaką funkcję ta metoda optymalizuje.

Zadanie 3 (2 pkt)
Opisz swój pomysł na sposób dokonania wyboru modelu (zbioru najlepszych predyktorów) w metodach: elastic net oraz random forest. Pomysł może być autorski lub znaleziony w podręcznikach lub artykułach naukowych. W tym drugim przypadku podaj źródła.

Zadanie 4 (2 pkt)
Zbuduj modele: elastic net (z paczki glmnet) oraz random forest. Zastosuj walidację krzyżową, aby

- dobrać parametry tuningowe,
- wyestymować błąd testowy dla swoich modeli.

Zrób podsumowanie tabelaryczne wyników, jakie otrzymywały metody w walidacji krzyżowej. Określ, który model wydaje Ci się najlepszy i dlaczego.

Zadanie 5 (8 pkt) (Przygotowanie predykcji dla danych testowych)
Naucz wybrany model na całych danych treningowych, dla zbioru cancer.RData. Zastosuj nauczony model do danych testowych i przewidź zmienną objaśnianą dla tych danych.

Co należy oddać

- Raport napisany w markdown i skompilowany do pliku o nazwie [nazwisko].pdf,
- Kod z implementacją w pliku [nazwisko].Rmd,
- Dane z predykcji w pliku [nazwisko].RData,

gdzie nazwisko to nazwisko autora prac nad zadaniem.

Po załadowaniu pliku [nazwisko].RData sprawdzający powinien znaleźć wektor o nazwie pred, czyli wektor predykcji dla danych testowych. Kolejność elementów w wektorze z predykcjami powinna odpowiadać kolejności danych testowych (czyli i-ta predykcja jest dla i-tej danej testowej).

Sposób oceniania predykcji
Miarą poprawności predykcji będzie mean squared error (MSE; błąd średniokwadratowy).