Zadanie zaliczeniowe
SAD1

TERMIN ODDANIA to 13 czerwca 2019 r. Łącznie można zdobyć 20 punktów. Oddanie do 2 dni po terminie powoduje przyznanie co najwyżej 10 pkt. Oddanie po 16 czerwca powoduje przyznanie 0 pkt.

UWAGA: 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.**
Zadanie należy rozwiązać w języku R.

1) W pliku `protein.RData`.
   Dane czytamy w `R` poleceniem `load("protein.RData")`.
   i) Zbiór treningowy `data.train` zawiera kolumnę `Y` oznaczającą zmienną objaśnianą - poziom pewnego białka w grupie pacjentów. Pozostałe kolumny to zmienne objaśniające. Na nim należy wytrenować odpowiedni model.
   ii) Zbiór testowy `data.test` nie ma zmiennej Y. Na nim należy zastosować nauczony model.

Cel: zastosować model z selekcją cech i wybrać najlepszy podzbiór zmiennych objaśniających, o dowolnym rozmiarze, tak aby błąd testowy był jak najmniejszy. Zidentyfikować zbiór najważniejszych predyktorów.

2) W pliku `cancer.RData`.
   Dane czytamy w `R` poleceniem `load("cancer.RData")`.
   i) Zbiór treningowy `data.train` zawiera, oprócz kolumn dla predyktorów, kolumnę `Y` oznaczającą działanie leku na nowotworowe linie komórkowe. Pozostałe kolumny to zmienne objaśniające (ekspresja genów w liniach).
   ii) Zbiór testowy `data.test` nie ma zmiennej `Y`. Na nim należy zastosować nauczony model.

Cel: Wybrać najlepszy model danych i wybrać najlepszy podzbiór zmiennych objaśniających, tak aby błąd testowy był jak najmniejszy.

Zadania. Dla obu zbiorów danych:

1) (Łącznie 10pkt) **Przygotuj raport z analizy danych:** Powinien on zawierać:
   a) (2 pkt) Analizę zmiennych objaśniających. Opisz, jakiego są one typu? Dla podzbioru najlepszych wybranych predyktorów zilustruj poziom współliniowości między kolumnami na histogramie używając statystyki VIF (*Variance Inflation Factor*).
   b) (4 pkt) Opis co najmniej dwóch różnych metod uczenia statystycznego danych, których użyłeś/-aś, aby zbudować model dla danych `protein.RData` i co najmniej trzech metod, których użyłeś/-aś dla danych

`cancer.RData`. Podaj uzasadnienie, dlaczego te metody wydały Ci się adekwatne.

c) (3 pkt) Zastosuj walidację krzyżową, aby dokonać estymacji błędu testowego dla swoich modeli. Zrób podsumowanie tabelaryczne wyników, jakie otrzymywały metody opisane w punkcie b) w walidacji krzyżowej na obu zbiorach danych (łącznie co najmniej 5 wyników). Określ, który model wydaje Ci się najlepszy dla zbioru `protein.RData` i dla zbioru `cancer.RData` i dlaczego.

d) (1 pkt) Podaj listę (w tabelce) najważniejszych predyktorów dla zbioru `protein.RData` i dla zbioru `cancer.RData`. Opisz, dlaczego wybrałeś/-aś właśnie te predyktory i jaką metodą. Dla zbioru `protein.RData` należy podać 5 najważniejszych predyktorów, a dla `cancer.RData` należy podać 100 najważniejszych predyktorów.

2) (Łącznie 10 pkt) **Przygotuj predykcje dla danych testowych:** Naucz wybrany typ modelu z wybraną liczbą najważniejszych predyktorów na całych danych treningowych, dla zbioru `protein.RData` i dla zbioru `cancer.RData` . Zastosuj nauczone modele do obydwu zbiorów danych testowych i przewidź zmienną objaśnianą dla tych danych.

3) Raport, kod w R z implementacją oraz dane z predykcji w podpunktach a, b, pkt 2 należy wysłać mailem do prowadzącego laboratorium. Predykcje oraz najważniejsze predyktory z zadania 1 powinny znajdować się w jednym pliku o nazwie `nazwisko.RData`, gdzie nazwisko to nazwisko autora prac nad zadaniem. Po załadowaniu tego pliku powinny znaleźć się trzy wektory:

   a) `pred.protein` – wektor predykcji dla modelu regresji liniowej o optymalnej liczbie zmiennych dla danych `protein.RData`.
   b) `predictors.protein` – wektor zawierający nazwy kolumn odpowiadające pięciu najważniejszym predyktorom dla danych `protein.RData.`
   c) `pred.cancer` – wektor predykcji dla modelu o optymalnej liczbie zmiennych dla danych `cancer.RData`.

   Kolejność elementów w wektorach 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).