Zadanie projektowe musi być wykonane najpóźniej na tydzień przed końcem zajęć

Realizacja jest samodzielna - w domu – taka jest formuła projektu.

Do realizacji zadania **wystarczają** moje wykłady –

lub książka :

G. James, D. Witten, T. Hastie, R. Tibshirani: „***An Introduction to Statistical Learning***", Springer - rozdziały 3,4,6 - książka ogólnie dostępna w pdf w sieci

Jako rezultat uzyskam **na emaila** :

a) **Sprawozdanie (w języku angielskim) w pliku pdf o nazwie IGT-nr sekcji :**
- na początku lista osób wykonujących z opisem co kto zrobił
- Temat zadania
- Odpowiedzi na pytania – wraz z uzasadnieniem – jeśli się o to pyta

b) **kod w R w pliku o nazwie tej samej co sprawozdanie z rozsz. R**

Zadanie

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

(b) Generate a response vector Y of length $n = 100$ according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where β_0, β_1, β_2, and β_3 are constants of your choice.

(c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) Now fit a lasso model to the simulated data, again using $X, X^2, \ldots, X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

(f) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

and perform best subset selection and the lasso. Discuss the results obtained.