SMDA 2018/19 – Exercise 5, Lecture L9 - 07/12/2018

Exercise 5: Analysis of Prostate Cancer dataset – regression methods with derived input directions (Principal Component Regression) and model selection
Please, execute the following tasks and provide answers to the proposed questions.

1. Open your kernel SMDA_EX2(L5)_ProstateCancer_Surname in Kaggle

2. Generate a copy called SMDA_EX4(L8)_PCR_Surname by the Fork button

3. Perform the following import from sklearn:
- **from sklearn.decomposition import PCA**
- **from sklearn.linear_model import LinearRegression**
- **from sklearn.model_selection import KFold**

4. Starting from the end of the kernel, compute a new model for the prostate cancer dataset by Principal Component Regression
- Hint: first compute Principal Component decomposition on the standardized matrix of independent variables (training set only)
- Hint: consider all (i.e., 8) components at first, then you will try to consider only the first j components, for j<8, and compare the results
- Hint: transform the standardized matrix of the training set according to the principal components computed at the first step. We'll call the obtained matrix "matrix of Transformed Training Set"
- Hint: transform the standardized matrix of the test set according to the principal components computed at the first step. We'll call the obtained matrix "matrix of Transformed Test Set"
- Compute an OLS model using the "matrix of Transformed Training Set" for independent variables

5. Show model's coefficients, intercept and score (on both the training and the test set)
- Hint: coefficients computed on the transformed space refer to principal components, not to original variables, hence an inverse transformation has to be performed. This inverse transformation involves the multiplication by the matrix of principal components

6. Compute the PCR model with all possible numbers of principal components (from 1 to 8) and show related model coefficients, intercepts and scores

7. Plot the scores on training and test sets depending on the number of components used
- Hint: put in the x-axis the number of principal components and in the y axis the scores

8. Find the best number of principal components by 10-folds cross-validation
- Perform cross-validation on the entire dataset

9. Plot the cross-validation mean performance against the number of components
- Use functions from sklearn.model_selection