Assignment 1: OLS+Ridge+feature engineering

1. Divide your data into 3 parts: the training set, the validation, and the test set. Decide which ratio TRAIN : VALIDATION : TEST to use and explain your motives.

2. Standardize all you features $X_1, ..., X_p$. Standardization of feature X_i means replacing its values with $\frac{X_i - \overline{X_i}}{\sqrt{\overline{X_i^2} - \overline{X_i}^2}}$, where $\overline{f}$ denotes averaging over your training sample[1].

3. Find weights of linear regression using the least square estimate (of course, using only training set):

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_p X_p + \epsilon$$

   i.e. by:

$$\mathbf{w}^* = (X^T X)^{-1} X^T Y$$

   Let us call that model as Model I.

4. Calculate Z-scores:

$$Z_i = \frac{w_i^*}{\hat{\sigma}\sqrt{v_{ii}}}$$

   where $(X^T X)^{-1} = (v_{ij})_{0 \le i,j \le p}$, and divide your variables into 3 groups: relevant, of moderate relevance, not relevant.

5. Using transformations $h_m(X_1, ..., X_p)$ from the list below

   - $h_m(\mathbf{x}) = X_j^2$ or $h_m(\overline{X}) = X_i X_j$ — 2nd order polynomials.
   - $h_m(\mathbf{x}) = \log(X_j), \sqrt{X_j}, ||\overline{X}||$
   - $h_m(\mathbf{x}) = [L_M \le X_k < U_m]$ — indicator region of X_k
   - $h_m(X) = (X - t_m)_+^{\alpha_m}$ — splines

   introduce new features to your model: $X_{p+1} = h_1(\mathbf{x}), ..., X_{p+M} = h_M(\mathbf{x})$. Explain intuition behind your new features. The number of new features should be such that $p + M$ is at least 15.

6. Before to proceed, standardize all you features $X_1, ..., X_{p+M}$.

[1] L_2-regularized models are strongly dependent on features' scale, therefore, we always standardize our variables. Remember that rule: except for some special cases, standardization of predictors is a very important prerequisite of most of models.

7. Consider a new model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_{p+M} X_{p+M} + \epsilon$$

and estimate weights β_i^* according to the least square estimate, i.e. by:

$$\beta^* = (X'^T X')^{-1} X'^T Y$$

where $X' \in \mathbb{R}^{n \times (1+p+M)}$ is a new data matrix with added features. We will call this estimate Model II.

8. Calculate Z-scores of all variables using β^* and divide your variables into 3 groups: relevant, of moderate relevance, not relevant.

9. Choose 3-5 candidates for $\lambda : \lambda_1, \lambda_2, ..., \lambda_5$ and find new weights:

$$\beta_\lambda^* = \arg\min_\beta ||X'\beta - Y||^2 + \lambda||\beta||^2 = (X'^T X' + \lambda I)^{-1} X'^T Y$$

Explain your choice of values for parameter λ. We will call this estimate Model III.

10. Calculate square loss and R^2 of all your models I-III (for different values of parameters) on the training set and on the validation. Draw a table of your results. Find the best values for λ based on the table.

11. Make a desision, which one of all models is the best one. Explain your choice.