# Homework 2 (60p)

- the same rules about submission apply as for Homework 1

## Question 1 (30p) [Classification]

Go to `https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied/data` and download the dataset along with the PDF file. It contains information about small businesses in US that obtained loans and which subsequently defaulted. The description of the dataset, along with ideas on how to analyze it are in the PDF file. There is additional code at the above link which you may use as long as you don't copy large chunks of it.

**Data pre-processing:**

- feel free to pre-process the dataset, such as dropping values with missing observations, or dropping some covariates but keep enough of them so that using model selection makes sense. You are allowed to drop a part of the dataset as long as observations are missing or you have good reasons to do so. Explain clearly each of your data pre-processing steps and motivate them.

**You are interested in predicting whether a small business will default (outcome is default yes/no) based on covariates in the dataset**. Split the data randomly into two (approximately equally big) parts: training and test sample. Choose 5 of the following methods for prediction: *logistic regression, logistic regression with forward selection, logistic regression with lasso, logistic regression with adaptive lasso, KNN, a single tree with pruning, bagging with prunning, bagging without pruning.* For lasso methods, make sure you compare the lasso solutions to the post-lasso solution (logistic regression with coefficient selected to be non-zero by lasso).

a) Plot the data and argue verbally or by plots which variables you think should explain best default.

b) Choose tuning parameters (when necessary) by methods we learned in class (you fit the model always on the training data, and can select the tuning parameters in the training or test data. For example, cross-validated error in the test sample, 1 std away from the minimum; test error in the test sample. You may choose another criterion but explain carefully your choice of method for picking the tuning parameter).

c) Compare the methods above in terms of the metric of your choice, but always on the test set and not the training set. Explain which method performs best, and why you think that is the case.

d) Highlight for each method which **(maximum 5)** covariates are best in terms of prediction and why you think that is (compare complexity, degrees of freedom, tradeoffs).

e) Pick a method based on d) and compare this to a). Explain why your results are (not) as you expected in a) and what may be the reason behind this.

*Hint: Here, there is no one solution that is perfect, and each group may consider their own strategy. There are many choices that you can and should make yourself. For any questions you may have, please ask the teaching assistant. The clearer the analysis and the explanations, the higher the grade on this section. Note that longer answers do not lead to a better grade, but clearer answers do.*

## Question 2 (30p) [Ridge regression: when does it dominate OLS ?]

Let the true model be $y_i = X_i'\beta + \epsilon_i, i = 1, \ldots, n$, with $y_i$ a scalar, $X_i$ a $p \times 1$ vector of observed covariates, and $\beta$ a $p \times 1$ vector of unknown parameters. In matrix form, this model is $y = X\beta + \epsilon$, where $y$ is $n \times 1$ and $X$ is $n \times p$, and $p$ is fixed. Let the ridge estimator for $\lambda > 0$ be $\hat{\beta}(\lambda) = (X'X + \lambda I_p)^{-1} X'y$. Below, we hold $\lambda$ fixed, unless stated otherwise. You may use a) to solve b); a) and b) to solve c) etc. Assume throughout the exercise that:

**Assumption 1** *(i) $X'X$ is invertible; (ii) $E(\epsilon|X) = 0$; (iii) $Var(\epsilon|X) = \sigma^2 I_n$.*

Let $W(\lambda) = (X'X + \lambda I_p)^{-1}$.

a) Show that $E[\hat{\beta}(\lambda)|X] - \beta = -\lambda W(\lambda)\beta$, therefore that the ridge estimator is biased.

b) Show that $Var(\hat{\beta}(\lambda)|X) = W(\lambda)(\sigma^2 X'X)W(\lambda)$ and that $Var(\hat{\beta}(0)|X) - Var(\hat{\beta}(\lambda)|X)$ is positive definite (pd), therefore that the ridge estimator is more efficient than the OLS estimator. *(Hint: use $A^{-1} + B^{-1} = B^{-1}(B + A)A^{-1}$; try to simplify operations as much as possible)*

c) Let the in-sample predictive mean square error be: $PMSE(\lambda|X) = E[\|X\hat{\beta}(\lambda) - X\beta\|^2 \,|\, X]$. Show that

$$PMSE(0|X) - PMSE(\lambda|X) = \lambda \, \text{trace} \, \{W^2(\lambda)(2\sigma^2(X'X) + \lambda(\sigma^2 I_p - \beta\beta' X'X))\}.$$

d) Show that if $\sigma^2 I_p - \beta\beta' X'X$ is positive definite, then the ridge estimator dominates the OLS estimator in predictive mean squared error, in the sense that $PMSE(0|X) - PMSE(\lambda|X) > 0$.

e) Assume that $n^{-1}X'X \xrightarrow{p} \Sigma$, a positive definite matrix of constants, and $\lambda = an^{\alpha}$, where $a > 0$ is a constant. If $\alpha \in (0, 1/2)$, show that $PMSE(0|X) - PMSE(\lambda|X) \xrightarrow{p} 0$, but if $\alpha = 1/2$, then show that $PMSE(0|X) - PMSE(\lambda|X) \xrightarrow{p} c$, where $c = -a^2(\beta'\Sigma^{-1}\beta) < 0$.