553.740 Project 2; Penalized Regression.
Due on Monday March 25.

- **Please type your answers (using, e.g., LateX or MS Word.)**
- *The solution must be written with sufficient explanations and your results must be commented. Returning a series of numerical results and figures is not enough. A solution in the form of a program output is not acceptable either.*
- *Please return your program sources. They will not graded (so no direct credit for them), but they will be useful in order to understand why results are not correct (and decide whether partial credit can be given) and to ensure that your work is original. You may use any programming language, although Python, Matlab or R are recommended.*
- *The ".csv" files associated with this project are available on Blackboard.*

Question 1.

We consider a multivariate version of ridge regression where $X : \Omega \to \mathbb{R}^d$, $Y : \Omega \to \mathbb{R}^q$, $\beta_0 \in \mathbb{R}^q$ and b a $d \times q$ matrix, with predictor $f : \mathbb{R}^d \to \mathbb{R}^q$ given by:

$$f(x) = \beta_0 + b^T X \, .$$

Introducing a $d \times d$ positive semi-definite symmetric matrix D, we consider penalized least-square estimators $\hat{\beta}_0$, $\hat{b}$ given by minimizers of

$$F(\beta_0, b) = \sum_{k=1}^{N} |y_k - \beta_0 - b^T x_k|^2 + \lambda \mathrm{trace}(b^T D b) \, ,$$

where $(x_1, y_1), \ldots, (x_N, y_N)$ are training data.

(1.1) Letting $\bar{x}$, $\bar{y}$ denote the average of $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ respectively, and denoting by $\mathcal{Y}_c$ and $\mathcal{X}_c$ the matrices

$$\mathcal{Y}_c = \begin{pmatrix} (y_1 - \overline{y})^T \\ \vdots \\ (y_N - \overline{y})^T \end{pmatrix}, \mathcal{X}_c = \begin{pmatrix} (x_1 - \overline{x})^T \\ \vdots \\ (x_N - \overline{x})^T \end{pmatrix}$$

prove that

$$\hat{b} = (\mathcal{X}_c^T \mathcal{X}_c + \lambda D)^{-1} \mathcal{X}_c^T \mathcal{Y}_c$$

and $\hat{\beta}_0 = \overline{y} - \hat{b}^T \overline{x}$.

(1.2) Justify the fact that this multivariate problem can be seen as solving independently q separate univariate linear problems.

(1.3) Write a program that takes into input a matrix $\mathbf{Y}$ containing each y_k as row vectors, and $\mathbf{X}$ containing each x_k as row vectors, the matrix D and $\lambda > 0$ and returns the estimated parameters $\hat{\beta}_0$ and $\hat{b}$.

Use this program to fit a multivariate regression model on the data in the file "project2_S2019_Q1Train.csv" for all values on $\lambda = 0.01, 0.02, \ldots, 2.00$. For each of these values, compute the prediction error evaluated on the test set "project2_S2019_Q1Test.csv", and plot these errors as functions of λ. This dataset is such that $d = q = 4$ and the csv files contain the coordinates of X on the first four columns and those of Y on the last four.

Provide the estimated values of β_0 and b for $\lambda = 1$.

(1.4) Keeping the same notation, we slightly modify the problem by minimizing

$$F(\beta_0, b) = \sum_{k=1}^{N} |y_k - \beta_0 - b^T x_k|^2 + \lambda \text{trace}(\beta D \beta^T)\,,$$

where D now is a positive semi-definite $q \times q$ symmetric matrix and $\beta = (\beta_0, b^T)^T$. Prove that the optimal solution $\hat{\beta}$ must satisfy the " Sylvester equation"

$$\mathcal{X}^T \mathcal{X} b + \lambda b D = \mathcal{X}^T \mathcal{Y}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are defined in the lecture notes.

(1.5) Python and Matlab provide functions for the solution of Sylvester equations and it is recommended to use one of these languages to solve the next question. Write a program that computes the solution of the multivariate problem in question (1.4), taking as input the $\mathbf{X}$ and $\mathbf{Y}$ arrays, the parameter λ and the penalty matrix D, while returning the optimal β_0 and b.

Test this program with the dataset provided in "project2_S2019_1.2.csv," for which $d = 3$ and $q = 200$, and with

1. $\lambda = 10$, $D = \text{Id}$.
2. $\lambda = 1000$, D is a tridiagonal matrix with -1 above and below the diagonal, and 2 on the diagonal, except $D(1,1) = D(q,q) = 1$.

In each case, plot the values of $\beta_0(j)$, $b(1,j)$, $b(2,j)$ and $b(3,j)$ as functions of j (on the same chart).

Question 2.

(2.1) The files "project2_S2019_Q2Train.csv" and "project2_S2019_Q2Test.csv" contain samples of a 10-dimensional variable $\mathbf{X}$ and a one-dimensional $\mathbf{Y}$, with respectively $N = 250$ samples (training) and $M = 1000$ samples (test). Use the ridge regression program written in the question (1.3) to evaluate regression parameters for $\lambda = 1, 2, \ldots, 100$. Compute the prediction error on the test set and plot it as a function of λ. Provide also the minimum value of the error and the parameter λ at which it is attained.

(2.2) Write a program that computes the kernel version of ridge regression in two cases

1. Gaussian kernel $K(x, y) = \exp(-|x - y|^2/2\sigma^2)$.
2. Polynomial kernel of order h: $K(x, y) = (x^T y) + \cdots + (x^T y)^h$

The function should take as input the $\mathbf{X}$ matrix, the $\mathbf{Y}$ matrix, the kernel parameter and the coefficient λ of ridge regression.

Answer the same questions as in (2.1) for a Gaussian kernel with $\sigma = 2.5$ and for polynomial kernels with $h = 1, 2, 3, 4$. For the Gaussian kernel, use $\lambda = 0.001, 0.002, \ldots, 0.1$ instead of $1, 2, \ldots, 100$.

(2.3) One of the results in the kernel case should be identical to the linear one. Explain why.