EPFL | MGT-418 : Convex Optimization | Project 2

Questions – Fall 2023

Robust Regression with Huber Loss (graded)

This project is due on **November 22, 2023, at 23:59**. You may form teams of up to two people. Each team should upload a single zip-file containing their report and Python code to Moodle. Make sure to clearly state the team members in your report.

Description

Given a training dataset (x_i, y_i), $i = 1, \ldots, m$, consisting of inputs $x_i \in \mathbb{R}^n$ (*e.g.*, minimum and maximum temperatures on day i, *etc.*) and outputs $y_i \in \mathbb{R}$ (*e.g.*, energy consumption on day i), the goal of linear regression is to find coefficients $w \in \mathbb{R}^n$ and a threshold $b \in \mathbb{R}$ such that

$$y_i \approx w^\top x_i + b \quad \forall i = 1, \ldots, m.$$

This is usually achieved by solving an empirical loss minimization problem of the form

$$\underset{w \in \mathbb{R}^n,\, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^m L(w^\top x_i + b - y_i),$$

where L is typically the absolute loss ($L(z) = |z|$) or the squared loss ($L(z) = z^2$). The squared loss is sensitive to outliers because it grows quadratically with the absolute value of the prediction error. The absolute loss is less sensitive to the outliers but is non-smooth and therefore susceptible to numerical errors. The Huber loss function combines the statistical and numerical advantages of both the absolute and the squared loss. Specifically, for $\delta > 0$, the Huber loss function is defined as

$$L_\delta(z) = \begin{cases} \frac{1}{2} z^2 & \text{if } |z| \leq \delta \\ \delta |z| - \frac{1}{2}\delta^2 & \text{if } |z| > \delta. \end{cases}$$

In this exercise, we will solve the following regression problem, which minimizes the sum of the Huber loss of the prediction errors and a Tikhonov regularization term $\frac{\rho}{2}\|w\|_2^2$, where $\rho > 0$ is the regularization weight. The corresponding optimization problem reads as follows,

$$\underset{w \in \mathbb{R}^n,\, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^m L_\delta(w^\top x_i + b - y_i) + \frac{\rho}{2}\|w\|_2^2 \,. \tag{1}$$

Questions

1. **QP Reformulation (30 points)**: The infimal convolution of two functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is the function $h : \mathbb{R} \to [-\infty, \infty)$ defined through

   $$h(z) = \inf_{t \in \mathbb{R}} f(t) + g(z - t).$$

   Show that the infimal convolution of $f(z) = \delta|z|$ and $g(z) = \frac{1}{2}z^2$ is equal to the Huber loss function $L_\delta(z)$. *Hint:* Consider the cases $|z| \leq \delta$ and $|z| > \delta$ separately.

   Using this result, verify that problem (1) is equivalent to

   $$\underset{w \in \mathbb{R}^n,\, b \in \mathbb{R},\, t \in \mathbb{R}^m}{\text{minimize}} \quad \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^m \delta |w^\top x_i + b - y_i - t_i| + \frac{\rho}{2}\|w\|_2^2 \,,$$

which in turn can be reformulated as

$$\begin{array}{ll} \underset{\substack{w\in\mathbb{R}^n,\, b\in\mathbb{R},\, t\in\mathbb{R}^m \\ r^+,\, r^-\in\mathbb{R}^m_+}}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \delta \mathbf{1}^\top (r^+ + r^-) + \frac{\rho}{2}\|w\|_2^2 \\ \text{subject to} & w^\top x_i + b - y_i - t_i \le r_i^+ \quad \forall i = 1,\dots,m \\ & y_i - w^\top x_i - b + t_i \le r_i^- \quad \forall i = 1,\dots,m. \end{array} \tag{2}$$

2. **Robustness to Outliers (20 points)**: The data files `p2x.npy` and `p2y.npy` available from Moodle contain the x and y variables for 42 training samples. Solve problem (2) for this dataset with $\delta = 1$ and $\rho = 1$. Compare your results to the solution of the least-squares regression problem, which is obtained by setting $\delta = +\infty$ and $\rho = 1$ in problem (1). A skeleton of the code you will have to implement is provided in the Python file `p2q2.py`. Plot the predicted outputs $w^\top x + b$ as a function of the inputs x, and briefly comment on the performance of the two methods.

3. **Kernel Trick (50 points)**: As in the lecture, let $\phi : \mathbb{R}^n \to \mathbb{R}^N$ be a feature map that lifts the inputs to a higher-dimensional space $\mathbb{R}^N$, $N \ge n$. The resulting lifted regression problem is

$$\begin{array}{ll} \underset{\substack{w\in\mathbb{R}^N,\, b\in\mathbb{R},\, t\in\mathbb{R}^m \\ r^+,\, r^-\in\mathbb{R}^m_+}}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \delta \mathbf{1}^\top (r^+ + r^-) + \frac{\rho}{2}\|w\|_2^2 \\ \text{subject to} & w^\top \phi(x_i) + b - y_i - t_i \le r_i^+ \quad \forall i = 1,\dots,m \\ & y_i - w^\top \phi(x_i) - b + t_i \le r_i^- \quad \forall i = 1,\dots,m. \end{array} \tag{3}$$

3.1. Denote by λ_i^+ and λ_i^- the Lagrange multipliers corresponding to the constraints $w^\top\phi(x_i) + b - y_i - t_i \le r_i^+$ and $y_i - w^\top\phi(x_i) - b + t_i \le r_i^-$, respectively, and construct the Lagrangian function for problem (3). **(5 points)**

3.2. Show that the Lagrangian dual of problem (3) can be stated as problem (4) below,

$$\begin{array}{ll} \underset{\beta\in\mathbb{R}^m}{\text{maximize}} & -\frac{1}{2}\sum_{i=1}^m \beta_i^2 - \frac{1}{2\rho}\sum_{i=1}^m\sum_{i'=1}^m \beta_i \phi(x_i)^\top \phi(x_{i'})\beta_{i'} + \sum_{i=1}^m \beta_i y_i \\ \text{subject to} & \sum_{i=1}^m \beta_i = 0, \quad -\delta \le \beta_i \le \delta \quad \forall i = 1,\dots,m. \end{array} \tag{4}$$

Hint: Use the variable transformation $\beta_i \leftarrow \lambda_i^- - \lambda_i^+$, $i = 1,\dots,m$. **(15 points)**

3.3. Use the KKT conditions to show that

$$w_j^\star = \frac{1}{\rho}\sum_{i=1}^m \beta_i^\star \phi_j(x_i) \quad \forall j = 1,\dots,N \quad \text{and} \quad t_i^\star = -\beta_i^\star \quad \forall i = 1,\dots,m$$

at optimality. **(5 points)**

3.4. Show that $b^\star = -\beta_k^\star + y_k - \frac{1}{\rho}\sum_{i=1}^m \phi(x_k)^\top\phi(x_i)\beta_i^\star)$ for any $k \in \{1,\dots,m\}$ such that $\beta_k^\star \in (-\delta,\delta)$. *Hint:* Construct the Lagrangian function for the dual problem (4), and observe the Lagrange multiplier associated with the constraint $\sum_{i=1}^m \beta_i = 0$ corresponds to the primal variable b. Also, use the KKT conditions. **(5 points)**

3.5. When you formulate a term $x^\top Px$ in a QCQP or QP in matrix form, CVXPY will check if the matrix P is positive semidefinite. For large matrices P, which have eigenvalues close or equal to zero, numerical errors can lead to this check failing, even if the matrix is in fact positive semidefinite. For this reason you should tell CVXPY that the kernel matrix is positive semidefinite. Define the kernel matrix $\kappa \in \mathbb{S}^m$, where $\kappa_{ij} = \phi(x_i)^\top\phi(x_j)$. Show that κ is positive semidefinite. **(10 points)**

3.6. In contrast to the primal problem (3), the dual problem (4) can be solved without knowledge of the feature map ϕ. Instead, it suffices to know the kernel function $K(x, x') = \phi(x)^\top \phi(x')$. The data provided in `p3x.npy` and `p3y.npy` contains the power production of a solar power plant in India, measured in 15 minute intervals over a day in summer 2020. Solve the dual problem (4) using the Gaussian Kernel

$$K(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$$

with $\delta = 1$, $\rho = 10^{-4}$ and $\sigma = 240$. Solve also the original regression problem (2) with $\delta = 1$ and $\rho = 10^{-4}$ using the same dataset. A skeleton of the code you will have to implement is provided in the Python file `p2q3.py`. Plot the predicted outputs $w^\top \phi(x) + b$ as a function of the inputs x, and briefly comment on the performance of the two methods. *Hint: Formulate the problem in matrix-form and tell CVXPY that κ is PSD using* `cp.atoms.affine.wraps.psd_wrap` **(10 points)**