EPFL | MGT-418 : Convex Optimization | Project 2

Quantile Regression

(Graded)

Description

Given training samples $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, m$, consisting of inputs $\boldsymbol{x}_i \in \mathbb{R}^n$ (*e.g.*, the last n electricity prices before hour i) and outputs $y_i \in \mathbb{R}$ (*e.g.*, the price during hour i), the goal of linear regression is to find a coefficient vector $\boldsymbol{w} \in \mathbb{R}^n$ and a threshold $b \in \mathbb{R}$ such that

$$y_i \approx \boldsymbol{w}^\top \boldsymbol{x}_i + b \quad \forall i = 1, \ldots, m.$$

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

$$\underset{\boldsymbol{w} \in \mathbb{R}^n,\, b \in \mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{m} L(r_i(\boldsymbol{w}, b)), \tag{1}$$

where $r_i(\boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i$ is the prediction error of the i-th data point, and L is a convex loss function such as the absolute loss ($L(z) = |z|$) or the squared loss ($L(z) = z^2$). The squared loss yields a maximum likelihood estimate of $\boldsymbol{w}$ and b if the prediction errors are normally distributed. Unfortunately, it is quite sensitive to outliers as it grows quadratically with the absolute value of the prediction errors. The absolute loss, on the other hand, yields a maximum likelihood estimate of $\boldsymbol{w}$ and b if the prediction errors follow a Laplace distribution and is less sensitive to outliers as it grows linearly with the absolute value of the prediction errors. Both loss functions serve to predict an output y_i given an input $\boldsymbol{x}_i$. In some cases, it may be desireable to complement the forecast for y_i with a confidence interval. Quantile regression yields such an interval by considering a variety of penalty factors $\tau \in (0, 1)$ for over- and underestimates of y_i. In fact, it uses the pinball loss function defined as

$$L(z) = \tau[z]_+ + (1 - \tau)[z]_-,$$

where $[x]_+ = \max\{0, x\}$ and $[x]_- = \max\{0, -x\}$. From now on, we assume that L is the pinball loss function unless stated otherwise.

Questions

1. **(20 points)** Show that for any optimal regressor with parameters $\boldsymbol{w}^\star$ and $b^\star$, at least $\lceil \tau m \rceil$ training samples lie on or above the hyperplane $H = \{(\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R} \,|\, y = \boldsymbol{w}^{\star\top} \boldsymbol{x} + b^\star\}$, on which the prediction error vanishes, and at least $\lceil (1 - \tau)m \rceil$ training samples lie on or below H.

2. **(10 points, Bonus Question)**: Show that for any optimal regressor with parameters $\boldsymbol{w}^\star$ and $b^\star$ such that $b^\star$ is the smallest among all the optimal values of b, 0 is the τ-quantile of the uniform distribution on the prediction errors $r_i(\boldsymbol{w}^\star, b^\star)$, $i = 1, \ldots, m$, where the τ-quantile is defined as

$$\inf_{\eta} \left\{ \eta \;\middle|\; \frac{1}{m} |\{i = 1, \ldots, m \,|\, r_i(\boldsymbol{w}^\star, b^\star) \le \eta\}| \ge \tau \right\}.$$

3. **Linear Programming Formulation (10 points):** Verify that Problem (1) is equivalent to

$$\begin{array}{ll} \underset{\boldsymbol{w}\in\mathbb{R}^n,\, b\in\mathbb{R},\, \boldsymbol{t}\in\mathbb{R}^m_+}{\text{minimize}} & \sum_{i=1}^m \frac{1}{2} t_i + (\tau - 1/2)(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) \\ \text{subject to} & +\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \leq t_i \quad \forall i = 1, \ldots, m \\ & -\boldsymbol{w}^\top \boldsymbol{x}_i - b + y_i \leq t_i \quad \forall i = 1, \ldots, m. \end{array} \tag{2}$$

Hint: Show first that the pinball loss can be written as $L(z) = 1/2|z| + (\tau - 1/2)z$.

4. **Quantile Regression vs. Least-Squares Regression (10 points):** Use the skeleton code *quantile_regression.m* to implement Problem (2) as a function in Matlab with inputs $\boldsymbol{x}, y, \tau$ and outputs $\boldsymbol{w}$ and b. Use this function to solve Problem (2) for the training samples in *p2q4data.mat* for $\tau = 0.1, 0.5, 0.9$. Compare your results to the solution of the least-squares problem obtained with $L(z) = z^2$. Use the skeleton code *p2q4.m* to do so. Plot the predicted outputs $\boldsymbol{w}^\top \boldsymbol{x} + b$ as a function of the inputs $\boldsymbol{x}$, and comment on the performance of the two methods.

5. **Empirical Quantile Estimation (10 points):** Using the results from the previous question, plot the cumulative distribution function of the uniform distribution over the prediction errors $r_i(\boldsymbol{w}, b)$ for $\tau = 0.1, 0.5, 0.9$, and verify that 0 is a τ-quantile of this distribution.

6. **Regularization and Kernel Trick (40 points):** Let $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$ be a feature map that lifts the inputs $\boldsymbol{x}_i$ to a higher dimensional space $\mathbb{R}^N$, $N \geq n$. In addition, introduce a Tikhonov regularization term with weight $\rho > 0$. The resulting regularized lifted regression problem is

$$\begin{array}{ll} \underset{\boldsymbol{w}\in\mathbb{R}^N,\, b\in\mathbb{R},\, \boldsymbol{t}\in\mathbb{R}^m_+}{\text{minimize}} & \sum_{i=1}^m \frac{1}{2} t_i + (\tau - 1/2)(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + b - y_i) + \frac{\rho}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & +\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + b - y_i \leq t_i \qquad \forall i = 1, \ldots, m \\ & -\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b + y_i \leq t_i \qquad \forall i = 1, \ldots, m. \end{array} \tag{3}$$

   (a) **(5 points)** Let λ_i^+ and λ_i^- be the Lagrange multipliers of the constraints $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + b - y_i \leq t_i$ and $-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b + y_i \leq t_i$, respectively, and construct the Lagrangian for problem (3).

   (b) **(10 points)** Show that the dual problem of (3) is

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}^+, \boldsymbol{\lambda}^- \in \mathbb{R}^m_+}{\text{maximize}} & -\sum_{i=1}^m (\lambda_i^+ - \lambda_i^- + \tau - 1/2) \left(y_i + \frac{1}{2\rho} \sum_{j=1}^m K_{ij} (\lambda_j^+ - \lambda_j^- + \tau - 1/2) \right) \\ \text{subject to} & \sum_{i=1}^m \lambda_i^+ - \lambda_i^- = m\,(1/2 - \tau) \\ & \lambda_i^+ + \lambda_i^- \leq 1/2 \quad \forall i = 1, \ldots, m, \end{array} \tag{4}$$

   where $K_{ij} = \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j)$.

   (c) **(15 points)**Use the KKT conditions to show that

$$\boldsymbol{w}^\star = \frac{1}{\rho} \sum_{i=1}^m \Big((\lambda_i^-)^\star - (\lambda_i^+)^\star + 1/2 - \tau \Big) \boldsymbol{\phi}(\boldsymbol{x}_i) \quad \text{and} \quad b^\star = 1/\rho \sum_{i=1}^m K_{ki} \Big((\lambda_i^+)^\star - (\lambda_i^-)^\star + \tau - 1/2 \Big) + y_k$$

   for any $k = 1, \ldots, m$ such that $(\lambda_k^+)^\star + (\lambda_k^-)^\star < 1/2$.

   Hint: Examine the Lagrangian for Problem (4)*, and note that the Lagrange multiplier associated with the constraint* $\sum_{i=1}^m \lambda_i^+ - \lambda_i^- = m(1/2 - \tau)$ *is the primal variable* b.

(d) **(10 points)** As opposed to the primal problem (3), the dual problem (4) can be solved without explicit knowledge of the feature map $\boldsymbol{\phi}$. In fact, it suffices to know the kernel matrix $\boldsymbol{K}$. In the following, we investigate the performance of the Gaussian kernel

$$\boldsymbol{K}_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma}\right)$$

with $\sigma = 1/2$ by solving the dual problem (4) and the primal problem (3) without kernelization ($\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$). In both problems, we set $\rho = 10^{-4}$, $\tau = 1/2$, and use the training samples *p2q6.m*. To solve Problem (3), adapt your implementation of *quantile_regression.m* to include the regularization term $\rho\|\boldsymbol{w}\|_2^2/2$. Use the skeleton code *p2q6.m* for the remainder of your implementation. Plot the predicted outputs $\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}) + b$ as function of the inputs $\boldsymbol{x}$, and comment briefly on the performance of the two methods.

7. **Electricity Price Prediction (10 points):** Now, we use quantile regression to predict electricity prices. By fitting a time series x_t, $t = 0, 1, 2, \ldots$ with an auto-regressive predictor of memory n, we predict $x_{t+1} \approx \boldsymbol{w}^\top(x_t, \ldots, x_{t-n+1})^\top + b$.

   (a) **(5 points)** Use your implementation of *quantile_regression.m* and the skeleton code *p2q7.m* to solve Problem (3) for the electricity price dataset *electricity.mat* with $n = 10$, $m = 100$, $\tau = 0.1, 0.5, 0.9$, and $\rho = 0.01$. Plot the predictions on the first 50 samples, *i.e.*, for $t = 50, \ldots, 99$.

   (b) **(5 points)** Interpret the effect of τ in terms of over- and under-prediction.