DS 5200

Homework 2

Directions: Each part is worth 5 points. Please submit an unzipped PDF before the deadline, merging handwritten solutions with PDF output from an R-markdown file. Also submit a zipped version of the *.Rmd file.

1. *[Analytical question]* Consider the following loss functions for error terms $e_i, i = 1, \dots, N$ in linear regression. For each loss function, (i) state whether it is convex, (ii) provide a mathematical proof, and (iii) explain how it can be useful in the context of linear regression.

   a) Quadratic loss (related to mean squared error, L_2 norm) $L = \sum_{i=1}^{N} e_i^2$

   b) Mean absolute error (L_1 norm) $L = \sum_{i=1}^{N} |e_i|$

   c) Huber loss (smooth mean absolute error) with parameter δ

$$L = \sum_{i=1}^{N} l(e_i)\,, where\ l(e) = \begin{cases} \frac{1}{2}e^2, & if\ |e| \leq \delta \\ \delta|e| - \frac{1}{2}\delta^2, & if |e| > \delta \end{cases}$$

2. *[Analytical question]* For linear regression $Y_i = \theta_0 + \theta_1 X_i + e_i,\ i = 1, \dots, N$ minimizing squared loss (e_i is random noise variable):

   a) Calculate the gradient with respect to the parameter vector.

   b) Write down the steps of the (batch) gradient descent rule.

   c) Write down the steps of the stochastic gradient descent rule.

3. [Implementation question]

   a) Overlay graphs of the loss functions in **question 1** for a range of e (consider two different values of δ for Huber loss). Use the graph to discuss the relative advantages and disadvantages of these loss functions for linear regression.

   b) Implement gradient descent for the loss functions above

   c) Implement stochastic gradient descent for the loss functions above

4. [Implementation question] In this question we will revisit JW Figure 3.3, and empirically evaluate various approaches to fitting linear regression.

    a) Simulate N=50 values of X_i, distributed Uniformly on interval $(-2, 2)$. Simulate the values of $Y_i = 3 + 2X_i + e_i$, where e_i is drawn from $\mathcal{N}(0,4)$. Fit linear regression with squared loss to the simulated data using (i) analytical solution, (ii) batch gradient descent, and (iii) stochastic gradient descent implemented in Question 3. Set learning rate α to a small value (say, $\alpha = 0.01$).

    b) Repeat (a) 1,000 times, overlay the histograms of the estimates of the slopes, and overlay the true value. Comment on how the **choice of the algorithm** affects the estimates of the slope parameter.

    c) Simulate N=50 values of X_i, distributed Uniformly on interval $(-2, 2)$. Simulate the values of $Y_i = 3 + 2X_i + e_i$, where e_i is drawn from $\mathcal{N}(0,4)$. Fit linear regression with (i) squared loss with the analytical solution, (ii) mean absolute error with batch gradient descent, and (iii) Huber loss with batch gradient descent implemented in Question 3. Set learning rate α to a small value (say, $\alpha = 0.01$).

    d) Repeat (c) 1,000 times, overlay the histograms of the estimates of the slopes, and overlay the true value. Comment on how the **choice of the loss function** in the case of Normal distribution affects the estimates of the slope parameter.

    e) Simulate N=50 values of X_i, distributed Uniformly on interval $(-2, 2)$. Simulate the values of $Y_i = 3 + 2X_i + e_i$, where e_i is drawn from $\mathcal{N}(0,4)$. Modify the simulated values of Y to introduce outliers, as follows. With probability 0.1, select an observation for modification. If it is selected, increase its value by 200% with probability 0.5, and decrease its value by 200% with probability 0.5. Fit linear regression to the modified data, with (i) squared loss with the analytical solution, (ii) mean absolute error with batch gradient descent, and (iii) Huber loss with batch gradient descent implemented in Question 3. Set learning rate α to a very small value (say, $\alpha = 0.01$).

    f) Repeat (e) 1,000 times, overlay the histograms of the estimates of the slopes, and overlay the true value. Comment on how the **choice of the loss function in presence of outliers** affects the estimates of the slope parameter.