DS 5220 - Fall 2018 Olga Vitek

Homework 1

Analytical questions are best prepared using Latex/word, however photos of handwritten notes are also acceptable. For the questions involving programming, use a single notebook (Jupyter or R notebook) to answer all the programming questions. Run the code, explain the findings through markdown or visualizations and export it to PDF. Merge the notebook PDF with the rest of the files (latex or photos) of the homework. Submit the single PDF file through Blackboard.

Due on Blackboard before midnight on Tuesday September 18, 2018.
Each part of the problems 5 points

1. *[Analytical question]* Consider two Normally distributed random variables $Y_1$and Y_2 with expected values μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ.
   (a) State the joint probability distribution of these random variables
   (b) Use Bayes theorem to derive the conditional probability distribution of $Y_1|Y_2$
   (c) Explain why and how the form of the conditional probability distribution is relevant to the context of linear regression with one predictor
2. *[Analytical question]* Consider the following loss functions for error terms e_i, $i = 1, \ldots, N$ in linear regression. For each loss function, state whether it is convex, and provide a mathematical proof.
   (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

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

3. *[Analytical question]* For linear regression $Y_i = \theta_0 + \theta_1 X_i + e_i,\ i = 1, \ldots, N$ minimizing quadratic loss:
   (a) Derive the analytical solution
   (b) State the steps of gradient descent to estimate the slope and the intercept.
   (c) State the steps of batch gradient descent to estimate the slope and the intercept.
4. *[Implementation question]*
   (a) Overlay graphs of the loss functions in question 2 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

5. *[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 $Y_i = 2 + 3X_i + e_i$, 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) gradient descent, and (iii) batch gradient descent implemented in Question 4. 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 $Y_i = 2+3X_i+e_i$, 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 gradient descent, and (iii) Huber loss with gradient descent implemented in Question 4. 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 $Y_i = 2 + 3X_i + e_i$, 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 50% with probability 0.5, and decrease its value by 50% with probability 0.5. Fit linear regression to the modified data, with (i) squared loss with the analytical solution, (ii) mean absolute error with gradient descent, and (iii) Huber loss with gradient descent implemented in Question 4. Set learning rate α to a small value (say, $\alpha = 0.01$).

   (f) 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 presence of outliers affects the estimates of the slope parameter.