CIS 4190/5190: Applied Machine Learning — Fall 2024

Homework 1

Handed Out: September 11 — *Due: October 2, 8 p.m.*

- You are encouraged to format your solutions using LaTeX. Handwritten solutions are permitted, but remember that you bear the risk that we may not be able to read your work and grade it properly — we will not accept post hoc explanations for illegible work. You will submit your solution manuscript for written HW 1 as a single PDF file.
- The homework is **due at 8 PM** on the due date. We will be using Gradescope for collecting the homework assignments. Please submit your solution manuscript as a PDF file via Gradescope. Post on Ed Discussion and contact the TAs if you are having technical difficulties in submitting the assignment.
- Make sure to assign pages to each question when submitting homework to Gradescope. The TA may deduct 0.2 points per sub-question if a page is not assigned to a question.

1 Written Questions

Note: You do not need to show work for multiple choice questions. If formatting your answer in LaTeX, use our LaTeX template `hw1_template.tex` (This is a read-only link. You'll need to make a copy before you can edit. Make sure you make only private copies.).

1. [Bias-Variance Tradeoff] (7 pts) Suppose we have an L_2-regularized linear regression model, which has loss $L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 + \lambda\|\beta\|_2^2$. For each of the following, indicate whether it tends to increase **bias**, decrease bias, or keep bias the same, and similarly for **variance**:

   A) Decrease the number of training examples n

   B) Increase the regularization parameter λ

   C) Decrease the dimension d of the features $\phi(x) \in \mathbb{R}^d$

   D) Increase c, where we replace the features $\phi(x)$ with $c \cdot \phi(x)$, for some $c \in \mathbb{R}_{>0}$ (for this part, assume no regularization, i.e., $\lambda = 0$)

   E) Increase the gradient descent step size α (but not so much that gradient descent diverges)

   F) suppose you fit a model and find that it has low loss on the training data but high loss on the test data; for each of the above five values n, λ, d, c, and α, indicate whether you should increase or decrease it to reduce the test loss, or it has no impact on the test loss.

2. [Regularization/Sparsity] (6 pts) In class, we demonstrated the intuition behind ℓ_1 and ℓ_2 regularization. In this problem we will try to see why ℓ_1 regularization create

sparsity (i.e. reduce β to zero) from the perspective of gradient descent. As a reminder, here's the ℓ_1 regularized linear regression objective.

$$\mathcal{L}_{\ell_1}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

(a) (2 pts) Consider the update rules for weights β in Equation 1 using gradient descent. Write down the gradient of the ℓ_1 regularization term (i.e. second term) with respect to an individual weight β_j. You can ignore the case $\beta_j = 0$ where the gradient may be undefined

(b) (2 pts) Non-important features j tend have coefficients β_j close to zero. ℓ_1 regularization helps push these coefficients to exactly zero, leading to feature selection. Given a sufficiently large regularization parameter λ, explain why this happens from the perspective of gradient descent. Specifically, analyze how the gradient of the linear regression loss term (first term in Equation 1) and the gradient of the ℓ_1 regularization term (second term in Equation 1) contribute to this behavior w.r.t β_j.

(c) (2 pts) Consider ℓ_2 regularization. Does ℓ_2 regularization encourage sparsity (i.e., push some coefficients β_j to exactly zero)? Briefly justify your answer using similar reasoning as in the previous question.

Hint: You do not need to explicitly write out the gradient of ℓ_2 regularization, but a calculation might help.

3. [Linear Regression] (5 pts) We are interested here in a particular 1-dimensional linear regression problem. The dataset corresponding to this problem has n examples $(x_1, y_1), \ldots, (x_n, y_n)$ where x_i and y_i are real numbers for all i. Let $\mathbf{w}^* = [w_0^*, w_1^*]^T$ be the least squares solution we are after. In other words, $\mathbf{w}^*$ minimizes

$$J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2.$$

(a) Find the expression for the derivative of the objective function $J(\mathbf{w})$ with respect to w_0^* and w_1^*.

(b) Show that $\frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(x_i - \bar{x}) = 0$

(c) Is linear regression guaranteed to have a unique solution for any dataset?

where $\bar{x}$ and $\bar{y}$ are the sample means based on the same dataset

4. [Linear Regression] (8 pts)

Suppose we have a weight vector $\boldsymbol{w} = [w_0, w_1, w_2]^T$ with input vectors $\boldsymbol{x_n} \in \mathbb{R}^2$ and $y_n \in \{0, 1\}$. Let us initialize all the weights to be 0. Also, suppose we have $N = 2$ examples in our dataset: $(\boldsymbol{x_1} = [1, -1]^T, y_1 = 0), (\boldsymbol{x_2} = [-1, -1]^T, y_2 = 1)$. Work out on paper the process of training an l_2 regularized $(\lambda = 1)$ linear regression model with batch gradient descent (learning rate = 1) on the above dataset for two epochs (steps).

(a) (1 pts) What is the value of the loss function at the beginning?

(b) (4 pts) What is the final state of the trained weight vector after 2 steps, and the corresponding value of the loss function? (Hint: derive the partial derivative of the loss function with respect to weights, and calculate their values after each step)

$$L = \frac{2}{N}\Sigma_{i=1}^{N}(y_i - \boldsymbol{w}\boldsymbol{x_i})^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

(c) (3 pts) Derive the formula of the closed-form solution for ridge regression.

2 Python Programming Questions

A IPython notebook is linked on the class website. It will tell you everything you need to do, and provide starter code. Remember to include the plots and answer the questions in your written homework submission!