CIS 419/519: Applied Machine Learning	Fall 2024

Homework 1

Handed Out: September 11 *Due: 7:59 pm October 2*

Name: Alan Wu

PennKey: alanlwu

PennID: 41855518

1 Declaration

- **Person(s) discussed with:** *Your answer*
- **Affiliation to the course: student, TA, prof etc.** *Your answer*
- **Which question(s) in coding / written HW did you discuss?** ***Your answer***
- **Briefly explain what was discussed.** *Your answer*

2 Multiple Choice & Written Questions

1. (a) Increase variance; bias stays the same
   (b) Decrease variance; Increase bias
   (c) Decrease variance; Increase bias
   (d) Variance increases; Bias stays the same
   (e) Variance stays the same; Bias stays the same
   (f) for the following values to decrease test loss
   n: Increase n
   λ: Increase λ
   d: Decrease d
   c: Decrease c
   α: Keep α the same

2. (a) To derive the gradient of the L1 regularization term for the:
   We need to simply take the gradient of the equation with respect to B_j.

$$L_{\ell_1} = \lambda \sum_{j=1}^{p} |B_j| \tag{1}$$

$$\frac{\partial L_{\ell_1}}{\partial B_j} = \lambda \operatorname{sign}(B_j) \tag{2}$$

We ignore the case where $B_j = 0$ and derive that equation from the loss function. Notice that the gradient of the L1 regularized loss function is only dependent on the sign of B_j and the magnitude of λ.

(b) We can analyze the effect of the L1 regularization term on the parameters of the model by looking at the gradient of both the L1 regularization term and the MSE loss function with respect to B_j.

We can derive the full gradient of the loss function as such:

$$L_{\ell_1} = \frac{1}{N}\sum_{i=1}^{N}(y_i - B^T x_i)^2 + \lambda \sum_{j=1}^{p} |B_j|$$

$$\frac{\partial L_{\ell_1}}{\partial B_j} = \frac{-2}{N}\sum_{i=1}^{N}(y_i - B^T x_i)x_{ij} + \lambda \text{ sign}(B_j) \tag{3}$$

Part 1: The MSE Loss Term
Our two cases for the MSE loss that we have to consider is when B_j is predictive of y_i and when it is weakly or not predictive of y_i. We know that the MSE loss is not dependent on the value of λ.

For predictive features, B_j will be larger and thus, the gradient of the MSE loss term will be larger in magnitude (positive or negative) than those of weakly predictive features. And in this way, the gradient of the MSE loss term for predictive features will change more than those that are weakly predictive because they have greater impact on minimizing the loss.

Thus, the magnitude of B_j is dependent on how strongly correlated/predictive the feature x_{ij} is in relation to y_i.

Part 2: The L1 Regularization term
As we derived earlier, the gradient of the L1 regularization term is only dependent on the sign of B_j and the magnitude of λ.

The equation is: $\frac{\partial L_{\ell_1}}{\partial B_j} = \lambda \text{ sign}(B_j)$

This observation gives us two cases where the feature B_j of the L1 loss regularization term will be scaled by the value of λ and the sign of B_j.
And when we scale λ, we will observe the following:
For small λ, the L1 regularization term will have a small impact on the overall gradient. With respect to B_j, this means that many B_j may be non-zero because the MSE loss term dominates the gradient. For both predictive and non-predictive features we may have non-zero values.
For large λ, the L1 regularization term may overtake the MSE loss term in the

gradient. For coeffcients B_j where B_j has small magnitude, the L1 regularization term will dominate and thus this parameter will be pushed down/up based on the sign of the parameter. Once the L1 regularization term dominates we will see that many B_j will shrink to 0, as once the B_j arrives at 0, it will stay there.

(c) ;alskdjf;lak;

3. (a)

   (b)

4. (a)

   (b)

5. (a)

   (b)

   (c)

3 Python Programming Questions