practice midterm 1

Go Ito

April 20, 2019

Question 1.

Suppose you are given a weighted least square. Somehow, its coefficients consist of three corners of a triangle, so we need to minimize our loss function subject to this "constraint"; that is, three coefficients must sum up to 180. Then, our loss function can be expressed as the following (neglect $\frac{1}{N}$ part here):

$$L = (\mathbf{t} - \mathbf{X}\mathbf{w})^T \mathbf{A}(\mathbf{t} - \mathbf{X}\mathbf{w}) - 2\lambda(\mathbf{1}^T\mathbf{w} - 180)$$

The term $2\lambda(\mathbf{1}^T\mathbf{w} - 180)$ is called a lagrange multiplier, which often appears in optimization problems (and Christou's 100C :P) What is the estimate of our coefficients? i.e. What is $\hat{\mathbf{w}}$?

(For those who want to challenge furthermore, what is $\hat{\mathbf{w}}$ without having λ term? Hint: $\mathbf{1}^T\hat{\mathbf{w}} = 180$ still holds.)

Question 2.

Given a function:

$$f(w_1, w_2) = \sin(w_1) + \ln(w_2) + 4w_1w_2 + 90$$

Obtain a numeric gradient using $\epsilon = 0.01$.

Question 3.

Using the function from Question 2., perform two iterations of gradient descent (get $w_{(2)}$).

Start from $w_{(0)} = [\pi, 1]$, and use $\gamma = 0.05$

Question 4.

Given a dataset

$$\mathbf{X} = \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.7 & 3.2 \\ 4.6 & 3.1 \end{bmatrix}$$

$$\mathbf{t} = [1.4, 1.4, 1.3, 1.5]$$

Perform LOOCV, and obtain the average MSE(Loss) of all four cases.

For each "cases"(row 234, 134, 124, 123), the estimated coefficients are as follows (I purposefully did not add intercept)

$$\hat{\beta}_{234} = [0.18, 0.17]$$
$$\hat{\beta}_{134} = [3.14, -4.19]$$
$$\hat{\beta}_{124} = [0.24, 0.07]$$
$$\hat{\beta}_{123} = [0.37, -0.14]$$

Question 5.

Use the dataset in Question 4., obtain the estimated coefficients using OLS Loss and Ridge Loss($\lambda = 2$).

Question 6.

- When λ is big, what would happen to Lasso estimated coefficients?
- Then, what is the advantage of using Lasso regresison?
- What's the point of gradient descent?
- Why is K-fold CV more preferred than LOOCV? (one reason)
- What is over-fitting? How is it related to bias-variance trade off?
- Mr.Avash was trying to do some gradient descent and obtained the plot below (shame on you Avash :P). How could he improve his algorithm?