Machine Learning and Computational Statistics

Theodoros Georgiopoulos

Homework 3

Exercises for Unit 3: Estimators and Regularization

Solution for exercise 1

Consider the Lagrangian function of the ridge regression problem:

$$\min L(\theta) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \mathbf{x}_n)^2 + \lambda \|\boldsymbol{\theta}\|^2 . \tag{1}$$

We take the gradient of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, equate to 0 and solve.

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= 0 \\
2\sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \mathbf{x_n})(-\mathbf{x_n}) + 2\lambda\boldsymbol{\theta} &= 0 \\
2\sum_{n=1}^{N} (\mathbf{x_n}\mathbf{x_n}^T\boldsymbol{\theta} - \mathbf{x_n} y_n) + 2\lambda\boldsymbol{\theta} &= 0 \\
2\sum_{n=1}^{N} (\mathbf{x_n}\mathbf{x_n}^T\boldsymbol{\theta}) - 2\sum_{n=1}^{N} (y_n\mathbf{x_n}) + 2\lambda\boldsymbol{\theta} &= 0 \\
\sum_{n=1}^{N} (\mathbf{x_n}\mathbf{x_n}^T + \lambda\mathbf{I})\boldsymbol{\theta} &= \sum_{n=1}^{N} y_n \mathbf{x}_n
\end{aligned} \tag{2}$$

We can find the matrix from of the solution too. Let $\boldsymbol{X} = [\boldsymbol{x_1^T}, \boldsymbol{x_2^T}, \ldots, \boldsymbol{x_N^T}]^T$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T$. We have

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{x_1}\boldsymbol{x_1^T} + \boldsymbol{x_2}\boldsymbol{x_2^T} + \cdots + \boldsymbol{x_n}\boldsymbol{x_n^T} = \sum_{i=1}^{N} \boldsymbol{x}_n\boldsymbol{x}_n^T .$$

Additionally,

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{x_1}y_1 + \boldsymbol{x_2}y_2 + \cdots + \boldsymbol{x_n}y_n = \sum_{i=1}^{N} y_n\boldsymbol{x}_n .$$

So, the ridge regression solution can be written as

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Solution for exercise 2

Consider a 1-dimensional parameter estimation problem, where the true parameter value is θ_o. Let $\hat{\theta}_{MVU}$ be a minimum variance unbiased estimator of θ_o. Consider the parametric set F of all estimators of the form

$$\hat{\theta}_b = (1+a)\hat{\theta}_{MVU}$$

with $a \in \mathbb{R}$.

(a) The $\hat{\theta}_{MVU}$ is an unbiased estimator of θ_o. So, the MSE depends only on the variance of the estimator.

(b) We have $E[\hat{\theta}_{MVU}] = \theta_o$. So, $E[\hat{\theta}_b] = (1+a)\theta_o$ and for $a \neq 0$, all $\hat{\theta}_b$ estimators are biased.

(c) We have $MSE(\hat{\theta}_{MVU}) = E[(\hat{\theta}_{MVU} - E[\hat{\theta}_{MVU}])^2]$. MSE in order to be zero, it must have zero variance. For a finite number N, this is impossible because datasets have to be the same which is not the case.

(d) We have

$$\begin{aligned}
MSE(\hat{\theta}_b) &= E[(\hat{\theta}_b - \theta_o)^2] \\
&= E[((\hat{\theta}_b - E[\hat{\theta}_b]) + (E[\hat{\theta}_b] - \theta_o))^2] \\
&= E[(\hat{\theta}_b - E[\hat{\theta}_b])^2] + E[(\hat{\theta}_b - \theta_o)^2] \\
&= E[((1+a)\hat{\theta}_{MVU} - E[(1+a)\hat{\theta}_{MVU})^2]] + (E[(1+a)\hat{\theta}_{MVU}] - \theta_o)^2 \\
&= (1+a)^2 MSE(\hat{\theta}_{MVU}) + a^2\theta_o^2
\end{aligned}$$

(e) In order to have less MSE for the biased estimator, it has to

$$\begin{aligned}
MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU}) &\iff \\
(1+a)^2 MSE(\hat{\theta}_{MVU}) + a^2\theta_o^2 < MSE(\hat{\theta}_{MVU}) &\iff \\
a^2(MSE(\hat{\theta}_{MVU}) + \theta_o^2) + 2aMSE(\hat{\theta}_{MVU}) < 0 &\iff \\
a\left[a + \frac{2MSE(\hat{\theta}_{MVU})}{\theta_o^2 + MSE(\hat{\theta}_{MVU})}\right] &< 0
\end{aligned}$$

So, in order to get $MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU})$, a must be in the range

$$-\frac{2MSE(\hat{\theta}_{MVU})}{\theta_o^2 + MSE(\hat{\theta}_{MVU})} < a < 0.$$

(f) From (e) we have that $a+1 < 1$. So, $|a+1| < 1$. We multiply each side with $|\hat{\theta}_{MVU}|$ gives $|a+1||\hat{\theta}_{MVU}| < |\hat{\theta}_{MVU}|$. So, $|\hat{\theta}_b| < |\hat{\theta}_{MVU}|$.

(g) The minimum value of

$$MSE(\hat{\theta}_b) = (1+a)^2 MSE(\hat{\theta}_{MVU}) + a^2\theta_o^2$$

with respect to a occurs when the derivative becomes zero, that is when

$$2(1-a)MSE(\hat{\theta}_{MVU}) + 2a\theta_o^2 = 0,$$

or, equivalently, when

$$a_* = -\frac{MSE(\hat{\theta}_{MVU})}{\theta_o^2 + MSE(\hat{\theta}_{MVU})}.$$

(h) In practice, a_* cannot be determined because θ_o is unknown.

Solution for exercise 3

Consider a set N pairs $(y_n, x_n), n = 1, \dots, N$, satisfying the equation

$$y_n = \boldsymbol{\theta_o}^T \boldsymbol{x}_n + \eta_n,\ \ \eta_n \sim \mathcal{N}(0,\, \sigma^2)\,. \tag{3}$$

As we know, the LS estimator satisfies the equation

$$(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T)\boldsymbol{\theta} = \sum_{n=1}^{N} y_n \boldsymbol{x}_n. \tag{4}$$

Consider now the special case where the $\boldsymbol{\theta}$ is a scalar and $\boldsymbol{x}_n = 1$ for all n. In this case, we have

$$y_n = \theta_o + \eta_n. \tag{5}$$

(a) The LS estimator of θ_o for this case where all $\boldsymbol{x}_n$'s are now scalars equal to 1 is

$$N\hat{\theta} = \sum_{n=1}^{N} y_n \iff \hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

(b) We have

$$E[y_n] = E[\theta_o + \eta_n] = \theta_o.$$

So, the y_n is an unbiased estimator of θ_o.

(c) We have

$$E[\overline{y}] = E[\frac{1}{N}\sum_{n=1}^{N} y_n] = \frac{1}{N}\sum_{n=1}^{N} E[\theta_o + \eta_n] = \frac{1}{N}\sum_{n=1}^{N}\theta_o = \theta_o.$$

So, $\overline{y}$ is an unbiased estimator of θ_o.

(d) The $\overline{y}$ is the LS estimator for the 1-dimensional case. It is also the minimum variance unbiased estimator and we denote it as $\hat{\theta}_{MVU}$.

(e) We know that

$$\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T + \lambda\mathbf{I})\boldsymbol{\theta} = \sum_{n=1}^{N} y_n\mathbf{x}_n.$$

So, for our case that $x_n = 1$ for every n, we have

$$(N+\lambda)\hat{\theta}_{ridge} = \sum_{n=1}^{N} y_n \iff \hat{\theta}_{ridge} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}.$$

(f) We have

$$\hat{\theta}_{MVU} = \frac{1}{N}\sum_{n=1}^{N} y_n \tag{6}$$

and

$$\hat{\theta}_{ridge} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}. \tag{7}$$

From (6),(7) we have

$$\hat{\theta}_{ridge} = \frac{N\hat{\theta}_{MVU}}{N+\lambda}.$$

(g) We know that $E[\hat{\theta}_{MVU}] = \theta_o$. So,

$$E[\hat{\theta}_{ridge}] = E[\frac{N\hat{\theta}_{MVU}}{N+\lambda}] = \frac{N}{N+\lambda}\theta_o \neq \theta_o.$$

So, the ridge estimator is biased.

(h) It is

$$|\hat{\theta}_{ridge}| = |\frac{N}{N+\lambda}||\hat{\theta}_{MVU}|.$$

Since $|\frac{N}{N+\lambda}| < 1$, for $\lambda > 0$, we have

$$|\hat{\theta}_{ridge}| < |\hat{\theta}_{MVU}|.$$