"Machine Learning and Computational Statistics"

3rd Homework

Exercise 1:

(a) Use the Lagrangian function of the ridge regression problem

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

and show that the solution satisfies the equation

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda I \right) \hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} y_n \boldsymbol{x}_n \qquad \text{(A)}$$

Hints: Take the gradient of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ equate to **0** and solve.

It is $(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}) \, \boldsymbol{x} = (\boldsymbol{x} \boldsymbol{x}^{\mathrm{T}}) \, \boldsymbol{\theta}$

It is $A\boldsymbol{z} - \lambda \boldsymbol{z} = A\boldsymbol{z} - \lambda I \boldsymbol{z} = (A - \lambda I)\boldsymbol{z}$, where A is a matrix, $\boldsymbol{z}$ is a vector and λ is a scalar and I is the identity matrix.

It is $\frac{\partial ||\boldsymbol{\theta}||^2}{\partial \boldsymbol{\theta}} = \frac{\partial (\boldsymbol{\theta}^T \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\boldsymbol{\theta}$

(b) Prove that the above solution can be expressed in matrix form as

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

where $X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1l} \\ 1 & x_{21} & \cdots & x_{2l} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nl} \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

Hint: Prove that $X^T X = \sum_{n=1}^{N} x_n x_n^T + \lambda I$ and $X^T \boldsymbol{y} = \sum_{n=1}^{N} y_n x_n$

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

$$\hat{\theta}_b = (1 + \alpha)\hat{\theta}_{MVU}, \qquad (1)$$

with $a \in R$. Recall that

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - E(\hat{\theta}))^2 \right] + (\mathrm{E}(\hat{\theta}) - \theta_o)^2 \qquad (2)$$

(a) What can you infer from the fact that $\hat{\theta}_{MVU}$ is an unbiased estimator of θ_o?

(b) Prove that all $\hat{\theta}_b$'s of *F*, for $\alpha \neq 0$, are biased estimators of θ_o.

(c) Find the $MSE(\hat{\theta}_{MVU})$ using eq. (2). Explain why this value cannot be zero for finite *N*.

(d) Express the $MSE(\hat{\theta}_b)$ in terms of $MSE(\hat{\theta}_{MVU})$ (Substitute eq. (1) to eq. (2) and after some algebra, utilize the results of (c) above).

(e) Determine the range of values of the parameter α that result to estimators with MSE lower than $MSE(\hat{\theta}_{MVU})$ (Consider the inequality $MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU})$, substitute $MSE(\hat{\theta}_b)$ from (d) and you will end up with a second order polynomial wrt α, which should be negative. Determine its roots and define the range of values of α where polynomial is negative).

(f) Prove that for any value of α in the range defined in (e), it is $|\hat{\theta}_b| < |\hat{\theta}_{MVU}|$ (Show first that $|1+\alpha|<1$ and then utilize eq. (1)).

(g) Determine the value α^* of the parameter α that corresponds to the estimator giving the lowest MSE (Consider the expression of $MSE(\hat{\theta}_b)$ derived in (d), take the derivative wrt α, equate to 0 and solve).

(h) Explain why in practice α^* cannot be determined.

Exercise 3:

Consider a set *N* pairs $(y_n, \boldsymbol{x}_n)$, n=1,…,N, satisfying the equation

$$y_n=\boldsymbol{\theta}_o^{\mathrm{T}}\boldsymbol{x}_n+\eta_n, \qquad (3)$$

where η_n is normally distributed **zero mean** i.i.d. noise. As it is known, the LS estimator satisfies the equation

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

Consider now the special case where the $\boldsymbol{\theta}$ is a scalar and $\boldsymbol{x}_n$=1 for all n. In this case, eq. (3) becomes

$$y_n = \theta_o + \eta_n. \qquad (3)\ ^{1}$$

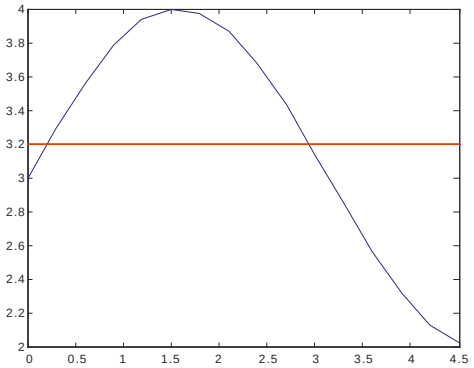
[1] In regression this implies a line perpendicular to the y-axis (left figure), while in classification implies a set of points spread around θ_o (right figure)

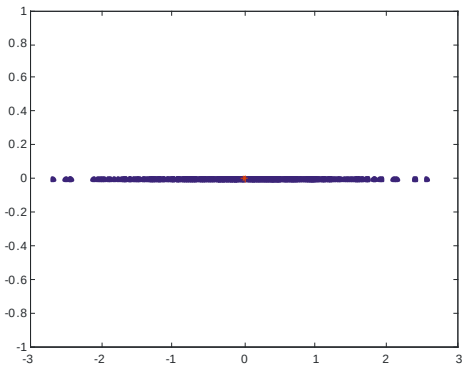
Let y_n denote the rv that models y_n. Let also $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ be the mean of *N* statistically independent rv's that follow the same pdf. Each one of the y_n's models a y_n.

(a) Derive the LS estimator of θ_o from eq. (4) for this case (Note that in eq. (4) all $\boldsymbol{x}_n$'s are now scalars equal to 1).
(b) Prove that y_n is an unbiased estimator of θ_o (Show that E[y_n]= θ_o).
(c) Prove that $\bar{y}$ is an unbiased estimator of θ_o (Show that E[$\bar{y}$]= θ_o).
(d) Since $\bar{y}$ is the LS estimator for the 1-dim. case, it is known that it is minimum variance unbiased estimator. From now on it will be denoted by $\acute{\theta}_{MVU}$.
(e) Prove that the ridge regression estimator for the present 1-dim. case is expressed as $\acute{\theta} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}$ (use the formula (A) proved in exercise 1a).
(f) Denoting as $\acute{\theta}$ the ridge regression estimator, express it in terms of $\acute{\theta}_{MVU}$.
(g) Prove that $\acute{\theta}$ is a biased estimator (Show that E[$\bar{y}$]$\neq \theta_o$).
(h) Verify that $|\acute{\theta}| < |\acute{\theta}_{MVU}|$.
(i) (*) Recalling from exercise 3 that $\acute{\theta}_b = (1+\alpha)\acute{\theta}_{MVU}$ express the quantity α in terms of λ. Then, utilizing exercise 3(e), derive the range of values of λ for which the MSE of the ridge regression estimator is less than that of the least squares estimator.

(*) This is not obligatory.

Exercise 4 (regularization - python code):





Consider the data set given in the attachment file (also the code for reading from python is also given). Specifically, it consists of 10 data pairs of the form (y_i,x_i), i=1,…,10. All y_i's are accumulated in the vector y while all x_i's are accumulated in the vector x.

The aim is to unravel the relation between x_i's and y_i's.

(a) Plot the data.
(b) Fit a 8^{th} degree polynomial on the data using the LS estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial.
(c) Fit a 8^{th} degree polynomial on the data using the regularized LS estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial.
(d) Discuss briefly on the results.

Hint for ***(b), (c)***: The X matrix that needs to be constructed will contain 10 rows, one for each x_i. Each row will have the form [1, x_i, x_i^2, x_i^3, x_i^4, x_i^5, x_i^6, x_i^7, x_i^8].