"Machine Learning and Computational Statistics"

3rd Homework

Exercise 1 (multiple choices question):

In the central square of a town, there are four thermometers A, B, C and D (of slightly different technology from each other) that measure the air temperature. The aim is to estimate the (mean) temperature at noon for the month July. To this end, we record each day the temperature shown by each thermometer at noon. At the end of the month, we average the measurements taken from thermometer A and, separately, we do the same for the thermometers B, C and D. Which of the following is correct?

1. The four averages are all estimates and, in general, have exactly the same value, since they measure the same quantity.

2. The four averages are all estimates of the mean temperature at noon in July.

3. In practice, we can detect which of the four averages/estimates is better, by comparing it with the true mean temperature value.

4. If the true mean temperature at noon in the July is known, there is no need to estimate it.

Exercise 2 (multiple choices question):

In twenty different medical laboratories throughout the world, the scientists aim to estimate the correlation coefficient, r, between the smoking and heart deceases. To this end, each lab uses the relevant data that it has collected during the last five years (for simplicity, assume that each lab has at its disposal a data set $Y_i, i = 1, \dots, 20$, of 10.000 cases) and twenty respective values for r are obtained, namely $r_1, r_2, \dots, r_{20}$. Which of the following are correct?

1. The $r_1, r_2, \dots, r_{20}$ are specific instances of a random variable, r, corresponding to the correlation coefficient.

2. The $r_1, r_2, \dots, r_{20}$ are all estimators of r.

3. The sets $Y_1, Y_2, \dots, Y_{20}$ are specific instances of the "random set" Y, which models all the data sets of the form of Y_i's with 10.000 cases each.

4. The $r_1, r_2, \dots, r_{20}$ are all estimates of the correlation coefficient and they can be considered as instances of the estimator (random variable) r that corresponds to the correlation coefficient.

Exercise 3 (multiple choices question):

Which of the following criteria are suitable to quantify the closeness of an estimator, $\hat{\theta}$, of the true model value θ_o in a single-parameter problem?

1. $\mathrm{E}[|\hat{\theta} - \theta_o|]$

2. $\mathrm{E}[\hat{\theta} - \theta_o]$

3. $\mathrm{E}\left[(\hat{\theta} - \theta_o)^2\right]$

4. $\mathrm{E}\left[(\hat{\theta} - \theta_o)^3\right]$

5. $(\hat{\theta} - \theta_o)^2$

Exercise 4 (multiple choices question):

Why, in practice, the quantity $\mathrm{E}\left[(\hat{\theta} - \theta_o)^2\right]$ cannot be computed explicitly?

1. The estimator $\hat{\theta}$ is unknown.

2. The true value, θ_o, is unknown in practice.

3. While all the involved quantities may be known, the computations are very demanding.

4. The expectation operator prevents the explicit computation.

Exercise 5 (multiple choices question):

In a practical machine learning problem, we have at our disposal only a specific data set Y of size N and, based on this, we compute an estimate $\hat{\theta}$ of the (unknown to us) θ_o parameter of an adopted model. Let $\hat{\theta}$ be the estimator associated with $\hat{\theta}$, for data sets of size N. Assume that it is known that the MSE error, $MSE = \mathrm{E}\left[(\hat{\theta} - \theta_o)^2\right]$ is very small. What implications has this to the obtained $\hat{\theta}$?

1. The estimate $\hat{\theta}$ is very likely to be away from θ_o.

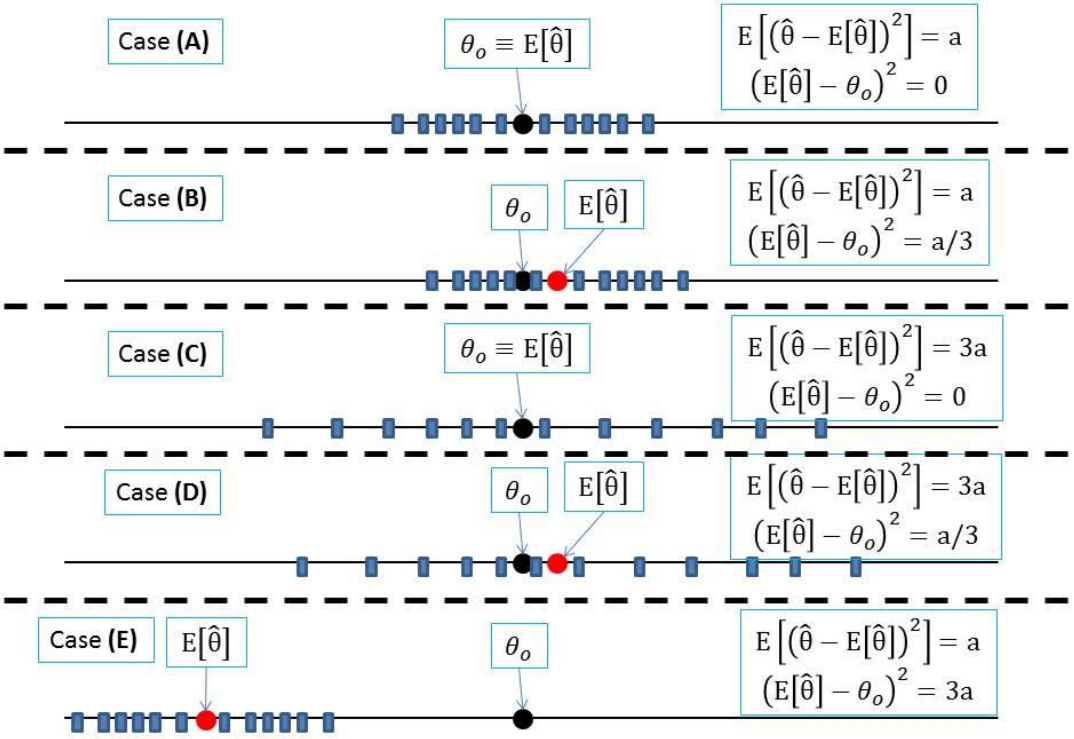
2. Another estimate based on a slightly different data set would also be close to θ_o, with high probability.

3. In general, the specific choice of Y does not play a significant role to the accuracy of the estimate $\hat{\theta}$.

4. The estimate $\hat{\theta}$ of θ_o, based on any data set of size N (for the specific problem) can be adopted with confidence, since, very small MSE implies that estimates are very close to θ_o, with high probability.

Exercise 6:

Consider the following figures, each one showing the locations of the true (unknown) θ_o (black circle) and $\mathrm{E}[\hat{\theta}]$ of an estimator (red circle). The arrangement of the blue boxes is indicative of the variance of each estimator around its mean, which is given for each case (a is a common scalar parameter in all cases).



For each of the above cases (a) determine the associated MSE value and (b) state whether or not estimator is biased or unbiased.

Exercise 7 (multiple choice question):

Consider a single-parameter problem with θ_o being the unknown parameter. Let $\hat{\theta}_1$ be an estimator of θ_o, which minimizes the criterion associated with a specific loss function

and is based on a randomly selected data set comprising N_1 data points. Let $\hat{\theta}_2$ be another estimator of θ_o, which minimizes the criterion associated with the same loss function as $\hat{\theta}_1$, and is based on another randomly selected data set of N_2 ($\neq N_1$) data points. Which of the following statements are true?

1. The two estimators are identical, since the both result from the same loss function.

2. Both estimators coincide with θ_o, since they both minimize a criterion associated with the specific loss function.

3. The two estimators, although they are expressed by the same mathematical formula, they differ since they apply on data sets of different sizes.

4. Any two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ_o associated with the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, coincide with θ_o, since both of them are instances of estimators that minimize the criterion associated with the specific loss function.

Exercise 8 (multiple choice question):

Consider the problem of estimating an unknown parameter θ, based on a set of noisy observations of it, of the form $y_n = \theta + \eta_n$, $n = 1, 2, \dots$, where η_n's are instances of the noise, whose mean value is zero ($\mathrm{E}[\eta] = 0$). Which of the following corresponding estimators of θ are unbiased?

1. $\frac{1}{3}\sum_{n=1}^{3} y_n$

2. $\frac{1}{3}\sum_{n=1}^{3} y_n + 0.2$

3. $\frac{1}{100}\sum_{n=1}^{100} {y_n}^3$

4. $\frac{1}{100}\sum_{n=1}^{100} y_n$

Exercise 9 (multiple choice question):

Consider the problem of estimating an unknown parameter θ, based on a set of noisy observations of it, of the form $y_n = \theta + \eta_n$, $n = 1,2, \dots$, where η_n's are instances of the noise, which is assumed white and Gaussian with zero mean ($\mathrm{E}[\eta] = 0$) and variance equal to ${\sigma_\eta}^2$. In this case, it is known that $\frac{1}{N}\sum_{n=1}^{N} \mathrm{y}_n$ is the minimum variance unbiased

estimator (MVUE). Consider the estimators $\hat{\theta}_4 = \frac{1}{4}\sum_{n=1}^{4} y_n$ and $\hat{\theta}_{200} = \frac{1}{200}\sum_{n=1}^{200} y_n$, which are of the above form. Since they are both unbiased, the MSE for each one of them is $\frac{{\sigma_\eta}^2}{4}$ and $\frac{{\sigma_\eta}^2}{200}$, respectively. Clearly, the second estimator exhibits lower MSE. How this matches with the general conclusion that $\frac{1}{N}\sum_{n=1}^{N} y_n$ is MVUE?

1. The noise variance varies for the two estimators, so that $\frac{{\sigma_\eta}^2}{4}$ and $\frac{{\sigma_\eta}^2}{200}$ to become equal.

2. The general conclusion is valid only for specific N.

3. The general result holds only for the mean of the two estimators.

4. Estimator $\hat{\theta}_{200}$ is biased, so that the addition of the bias term to the variance term, gives the MSE of $\hat{\theta}_4$.

Exercise 10 (multiple choice question):

Consider a multi-parameter regression problem, where $\boldsymbol{\theta} = [\theta_0,\ \theta_1, \dots, \theta_l]^T$ denotes the parameter vector that contains all the parameters of the associated model and let $\boldsymbol{\theta_o} = [\theta_{o0},\ \theta_{o1}, \dots, \theta_{ol}]^T$ be the true value of $\boldsymbol{\theta}$. Also, let $\widehat{\boldsymbol{\theta}} = \left[\hat{\theta}_0,\ \hat{\theta}_1, \dots, \hat{\theta}_l\right]^T$ be an estimator of $\boldsymbol{\theta_o}$. Which of the following statements is correct?

1. If the MSE per parameter $\hat{\theta}_i$ is known, the MSE of $\widehat{\boldsymbol{\theta}}$, is obtained through the multiplication of the MSEs of all the parameters.

2. If $\widehat{\boldsymbol{\theta}}$ is unbiased, its MSE equals to the sum of the variances of $\hat{\theta}_i$'s around their means.

3. In the (utopic) case where the variances of all $\hat{\theta}_i$'s are equal to 0, the MSE of $\widehat{\boldsymbol{\theta}}$ equals to the sum of the bias terms of all $\hat{\theta}_i$'s.

4. Although $\widehat{\boldsymbol{\theta}}$ may be biased, the biased terms corresponding to $\hat{\theta}_i$'s are not taken into account in the computation of the MSE of $\widehat{\boldsymbol{\theta}}$.

Exercise 11 (multiple choice question):

Consider a multi-parameter regression problem, where $\boldsymbol{\theta} = [\theta_0,\ \theta_1, \dots, \theta_l]^T$ denotes the parameter vector that contains all the parameters of the associated model and let

$\boldsymbol{\theta_o} = [\theta_{o0},\ \theta_{o1}, \dots, \theta_{ol}]^T$ be the true value of $\boldsymbol{\theta}$. Also, let $\widehat{\boldsymbol{\theta}} = \left[\hat{\theta}_0,\ \hat{\theta}_1, \dots, \hat{\theta}_l\right]^T$ be the minimum variance unbiased estimator (MVUE) of $\boldsymbol{\theta_o}$. Which of the following statements is correct?

1. One way to get an unbiased estimator with variance less than that of $\widehat{\boldsymbol{\theta}}$, is to magnify the norm of $\widehat{\boldsymbol{\theta}}$.

2. The estimator that results from the shrinkage of the norm of $\widehat{\boldsymbol{\theta}}$ is also an unbiased one.

3. All biased estimators that result from the shrinkage of the norm of $\widehat{\boldsymbol{\theta}}$, have variance less than that of $\widehat{\boldsymbol{\theta}}$.

4. One way to get a biased estimator with variance less than that of $\widehat{\boldsymbol{\theta}}$, is to shrink the norm of $\widehat{\boldsymbol{\theta}}$.

Exercise 12:

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

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

and show that the solution satisfies the equation

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

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

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

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 $X^T X = \sum_{n=1}^{N} x_n x_n^T$ and $X^T \boldsymbol{y} = \sum_{n=1}^{N} y_n x_n$

Exercise 13:

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 $\alpha \in R$. Recall that

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

(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 14:

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

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

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

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

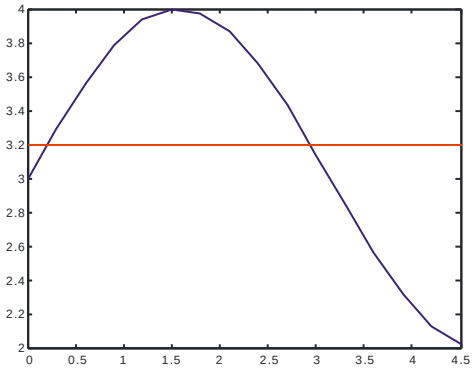
Consider now the special case where the $\boldsymbol{\theta}$ is a scalar and $\boldsymbol{x}_n = 1$ for all n. In this case, eq. (3) becomes

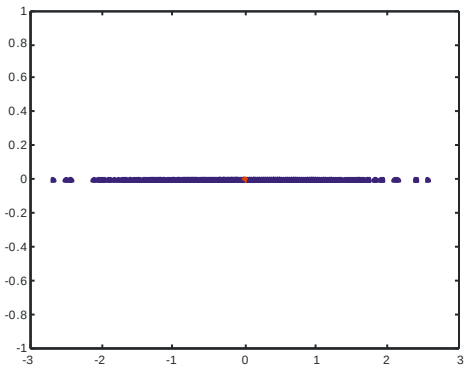
$$y_n = \theta_o + \eta_n. \qquad (3)\,[1]$$

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 $E[y_n] = \theta_o$).
(c) Prove that $\bar{y}$ is an unbiased estimator of θ_o (Show that E$[E[\bar{y}] = \theta_o$).
(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 $\hat{\theta}_{MVU}$.
(e) Prove that the ridge regression estimator for the present 1-dim. case is expressed as $\hat{\theta} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}$ (use the formula (A) proved in exercise 1a).
(f) Denoting as $\hat{\theta}$ the ridge regression estimator, express it in terms of $\hat{\theta}_{MVU}$.
(g) Prove that $\hat{\theta}$ is a biased estimator (Show that E$[\bar{y}] \neq \theta_o$).
(h) Verify that $|\hat{\theta}| < |\hat{\theta}_{MVU}|$.

[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)





(*) Recalling from exercise 3 that $\hat{\theta}_b = (1+\alpha)\hat{\theta}_{MVU}$ 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.

Exercise 15 (python code + text):

Consider a regression problem where both the independent and dependent quantities are scalars and are related via the following linear model

$$y = \theta_o \cdot x + \eta$$

where η follows the zero mean normal distribution with variance σ^2 and $\theta_o = 2$ (thus, the actual model is $y = 2 \cdot x + \eta$).

(a) Generate $d = 50$ data sets as follows:

- Generate a set D_1 of $N = 30$ data pairs (y_i', x_i), where $y' = 2 \cdot x$.
- Add zero mean and $\sigma^2 = 64$ variance Gaussian noise to the y_i' 's, resulting to y_i's.
- The **observed** data pairs are (y_i, x_i), $i = 1, \ldots, 30$, which constitute the data set D_1.

Repeat the above procedure $d = 50$ times in order to generate 50 different data sets.

(b) Compute the LS linear **estimates** of θ_o based on $D_1, D_2, \ldots, D_d$ (thus, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ numbers/estimates will result).

(c) Consider now the random variable $\hat{\theta}$ that models $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ (that is, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ can be viewed as instances of the random variable $\hat{\theta}$)[2] and

(c1) estimate the $MSE = E\left[\left(\hat{\theta} - \theta_o\right)^2\right]$ and

(c2) depict graphically the values $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ and comment on how they are spread around θ_o.

Hint: For (c) approximate MSE as $MSE = \frac{1}{d}\sum_{i=1}^{d}\left(\hat{\theta}_i - \theta_o\right)^2$.

[2] $\hat{\theta}$ is also known as an **estimator** of θ_o.

Exercise 16 (regularization - python code):

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

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 ridge regression estimator and plot the results (data points and the curve resulting from the fit). Output also the estimates of the parameters of the polynomial. Experiment with various values of λ.

(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 $[1, x_i, {x_i}^2, {x_i}^3, {x_i}^4, {x_i}^5, {x_i}^6, {x_i}^7, {x_i}^8]$.