"Machine Learning and Computational Statistics"

4th Homework

Exercise 1:

Consider the **regression problem**

$$y = g(\boldsymbol{x}) + \eta$$

and let $E[y|\boldsymbol{x}]$ denoting the minimum MSE estimate of y given $\boldsymbol{x}$. Consider the estimator $f(\boldsymbol{x};D)$.

(a) Under what conditions (theoretically) the quantity $E_D[(f(\boldsymbol{x};D) - E[y\,|\,\boldsymbol{x}])^2]$ becomes zero?

(b) Why this cannot be achieved in practice?

Exercise 2 (python code + text)**:**

Consider the **regression problem** (1-dep., 1-indep. variables)

$$y = g(x) + \eta$$

where y and x are jointly distributed according to the normal distribution $p(y,x) = N(\boldsymbol{\mu}, \Sigma)$

with $\boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $\qquad \Sigma = \begin{bmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$

(a) Determine theoretically $E[y|x]$ and plot the corresponding curve (recall the relevant theory concerning the normal distribution case).

(b) Generate 100 data sets D_i, i=1,…100, each one consisting of N=50 randomly selected pairs (y_n, x_n), $n=1,\ldots,N$, from $p(y,x)$.

(c) Adopt a linear estimator $f(x;D)$ and determine its instances $f(x;D_1)$,…, $f(x;D_{100})$, utilizing the LS criterion.

(d) Plot in a single figure **(i)** the lines corresponding to the above 100 estimates (blue color), **(ii)** the line that results by averaging over the 100 lines (red color) and **(iii)** the line corresponding to the optimal MSE estimate (green color). (consult the slides for (ii), in order to see how the average of $f(x;D_1)$,…, $f(x;D_{100})$ is taken)

(e) Repeat steps (b)-(d) where now each data set consists of N=5000 points.

(f) Discuss the results.

Exercise 3 (python code + text)**:**

Consider the set up of exercise 2 and recall the E[y|$\mathbf{x}$] determined there.

(a) Generate a single data set D' of 100 pairs (y_n,x_n), n=1,…,N from $p(y,x)$.

(b) Determine the linear estimate $f(x;D')$ that minimizes the MSE criterion, based on D'.

(c) Generate randomly a set of additional 50 points x'_n, n=1,…,50. For each one of these points determine the estimates $y_{n'} = f(x_n;D')$ (50 numbers (estimates) should be finally computed).

(d) For the previous 50 points determine the estimates $\grave{y}$ = E[y|$\mathbf{x}$].

(e) Based on the previous derived estimates for the 50 points from both $f(x_n;D')$ and E[y|$\mathbf{x}$], propose and use a (practical) way for quantifying the performance of $f(x_n;D')$ in terms of that of E[y|$\mathbf{x}$].

NOTE: Please give **brief explanations** in all **exercises**.