Cairo University - Faculty of Engineering
Computer Engineering Department
CMP4040 – Machine Learning
Spring 2024

Assignment #2

Solve the following exercises and report your answers.

Problem 2.16 (Handwritten)

In this problem, we will consider x as a one-dimensional variable. For a hypothesis set

$$H = \left\{ h_c(x) = sign\left(\sum_{i=0}^{D} c_i x^i \right) \right\}$$

Prove that the VC dimension of H is exactly $(D + 1)$ by showing that

(a) There are $(D + 1)$ points which are shattered by H.
(b) There are no $(D + 2)$ points which are shattered by H.

Problem 2.24 (Handwritten + Code)

Consider a simplified learning scenario. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval $[-1, 1]$. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$. Thus, the full data set is $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm returns the line fitting these two points as g (H consists of functions of the form $h(x) = ax + b$). We are interested in the test performance (E_{out}) of our learning system with respect to the squared error measure, the bias and the var.

(a) Give the analytic expression of the average function $\bar{g}(x)$.
(b) Describe an experiment that you could run to determine (numerically) $\bar{g}(x)$, E_{out}, bias and var.
(c) Run your experiment and report the results. Compare E_{out} with bias + var. ~~Provide a plot of your $\bar{g}(x)$ and $f(x)$ (on the same plot).~~
(d) Compute analytically what E_{out}, bias and var should be.
(e) Repeat (b) and (c) for a constant hypothesis function (h(x) = c), where c is the average value of the input data points. Repeat the experiment for a different number of input data points (2, 5, 10 and 20). Compare E_{out} with bias + var and provide a plot of your $\bar{g}(x)$ and $f(x)$ (on the same plot) for each number. Comment on your results.

Programming Problem (Code + Report)

Attached is the “Energy_Efficiency.xlsx” dataset file. Given 8 numerical attributes, we are required to predict the heating and cooling load requirements which is given in the two last columns. We will treat this problem as two tasks:

1- A linear regression task on the heating load (Y1).
2- A logistic regression task to predict on the cooling load (Y2). We assume that the sample load is high if the cooling load is > the load median.

The requirements are as follows:

1- Split the dataset into 70% training samples and 30% test samples. For the sake of reproducibility, don’t change the random seed.
2- Implement linear regression.
3- Implement logistic regression using batch gradient descent. Run the training process for each pair of the following hyper parameters:
   a. Learning rates: 0.5, 0.1, 0.05 and 0.01.
   b. Training epochs: 1, 10, 100, 1000.

For each training experiment, report the training and testing errors. For logistic regression, also report the training and testing accuracy. Write your conclusions based on the results.

Deliver your code as a documented Jupyter notebook written in python. You can use the libraries: pandas, numpy and matplotlib. You cannot use the learning algorithm from a machine learning library like Scikit-learn but you can use them to verify your results.

Rules:

- This is an individual assignment.
- Any evidence of plagiarism will be penalized by all the grades of the assignment.
- Beside the name of each problem, you will find the expected solution format which can be either of the following:
  - **(Handwritten)**: These problems should be solved on paper and scanned. The handwriting and the scanning result should both be clear.
  - **(Code + Report)**: These problems require coding to solve. In that case, you should write the needed code, generate the results, and solve the problem accordingly. You will find attached python notebooks. It’s required to update the notebooks with your solutions and comments (in the allocated areas). You should **only** submit the updated notebooks (with solutions displayed clearly) with your report.

Deliverables: One report of your solutions (PDF) + Three python notebooks. Remember to include your name, section & bench number in the first page of the report.

Deadline: Saturday, April 13th, 2024, 23:59