INT 6151: Statistical Machine Learning **Second Semester 2023-2024**

Homework 1

Submit Date: 31/03

- Feel free to talk to other students in the class when doing the homework. You should, however, write down your solution yourself. You also must indicate on each homework with whom you collaborated and cite any other sources you use including Internet sites.
- You will write your solution in LaTeX and submit the pdf file in zip files, including relevant materials, through courses.uet.vnu.edu.vn
- Dont be late.

1 Gradient Descent - 10pts

Let $f : \mathbb{R}^d \to \mathbb{R}$, start at some points $\mathrm{x}^{(0)} = \{x_1^{(0)}, \ldots, x_d^{(0)}\}$ with $k = 0$.

At the k step we have updated rule:

$$x_i^{(k+1)} \leftarrow x_i^{(k)} - \lambda \partial_{x_i} f(\mathrm{x}^k)$$

with $\partial_{x_i} f : \mathbb{R}^d \to \mathbb{R}$ is the derivative of the function with respect to the ith coordinate, $\partial_{x_i} f(\mathrm{x}^k)$ is the value at x^k and λ is the learning rate.

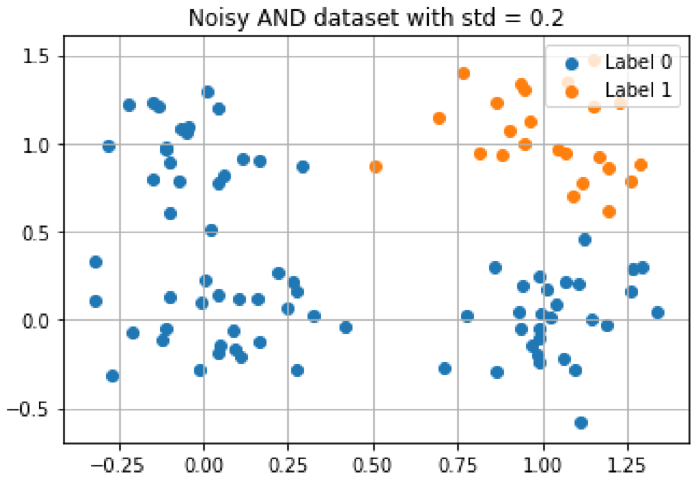
1. Let $f(x) = x^2 - x + 6$ and $\lambda = 0.05$. First, calculate $\partial_x f$.
   Then starting from $x^{(0)} = 1 (k = 0)$, apply the gradient descent with $k = 1, 2, 3$. (5pts)
2. Let $f(x) = (x_1 - x_2^2 - 3)^2 + (x_1 - x_2 - 1)^2$ and $\lambda = 0.2$. First, calculate $\partial_{x_i} f$. Then starting from $x^{(0)} = (-5, -5) (k = 0)$, apply the gradient descent with $k = 1, 2$. How many local maximize of f? Could you find it? (3pts).
3. Draw the contour plot of $f(x)$ in question 2, given the appropriated plot domain and visualize gradient descent with $\lambda = 0.01$ and $\lambda = 0.05$ from $x^{(0)} = (5, 5)$, after 10 steps. Which λ value is more optimized? (2pts)

2 Coding Homework - 10pts

Given a Noisy Dataset given by and_function.ipynb), which follows a predefined distribution $\mathcal{P}(x, y)$. More specifically, we add a noise vector (ϵ^1, ϵ^2) to get $x = (x^1 + \epsilon^1, x^2 + \epsilon^2)$

Let fix the noise be drawn from $\mathcal{N}(0, \epsilon)$ with $\epsilon = 0.2$ and $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_{10}, y_{10})\}$ is a dataset with $n = 10$ samples.

a. Assume we choose a fixed logistic function h which satisfies the condition $\widehat{\mathrm{err}}_D(h) = 0$. What can we say about $\mathbb{P}\{\mathrm{err}_{\mathcal{P}}(h) \ > 0.1\}$ (1pts)



b. Suppose we train a logistic regression model follow the ERM framework on the dataset to reach $\widehat{\text{err}}_D(h) = 0$. (Hint: Create a table to check the above probability for different running seeds) (5pts)

c. What happened when we increase the number of training samples in $\mathcal{D}$. Validate by your experiments? (2pts)

d. What happened when we draw the train samples from $\mathcal{N}(0, 0.2)$ and the test samples from $\mathcal{N}(0, 0.3)$ (2pts)

You should submit your code to verify your answer for the task b, c, d.