Homework 2 Simulation Exercises

Exercise 1

The goal of this exercise is to fit a regression model on a set of points.

Part A: Generate the following points whose x and y values are given as follows:

X = np.arange(-10, 10, 0.2)

Y = 2cos(x)/-pi + (2x)/(2pi) + 2cos(3x)/(-3pi)

Now, add a Gaussian noise $N(\mu = 0, \sigma = 1)$, and then a Poisson noise $P(\lambda = 2)$ to the Y values. Affect the noises with coefficient 0.1. Now fit regression models with degrees from 1 to 15. Display the output. Mention the MSE, bias, and variance errors for each fitting.

Exercise 2

Generate the following two sets of points:

1) 200 points belonging to a circular sector with (1.5, 0) as its origin and 4 and 9 as its lower and upper bounds.

2) 200 points belonging to a circular sector with (1.5, 0) as its origin and 0 and 6 as its lower and upper bounds.

Part A: Plot the sets, and separate them through using the Logistic Regression classifier. Obviously, the points are not separated linearly, so they should be moved to a latent space with higher dimensions. Define $\phi(\vec{x})$ in the following way:

$$\vec{X} = [x_1, x_2]^T$$

$$\phi(\vec{X}) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, ..., x_2^7]^T$$

$$f : R^2 \to R^{35}$$

Also, use the L2 Regularization technique. Explain how using the technique will prevent the model from over fitting.

Part B: Display the decision boundary for each classification and mention the accuracy values. Analyze the outputs.

Part C: Repeat the same process for the following dataset:

1) 100 points coming from a guassian distribution with (1, 0) as mean and 1 as standard deviation.

2) 200 points belonging to a circular sector with (1.5, 0) as its origin and 2 and 6 as its lower and upper radius.

Exercise 3

In this question, we want to implement the non-parametric method of the Parzen's Window. Implement the algorithm without using prebuilt modules in libraries.

Part A: the ted-talks dataset is attached in the HW's folder. Extract the duration column and estimate its distribution in the following cases using the Parzen's Window with a Gaussian kernel.

Part B: Set the window size value (V_n) to 10, 20, 50, and 100. Which one estimates the real distribution better?

Part C: Increase the number of samples (n) from 250 to the whole dataset with 250 as timestep. Which output estimates the original distribution better?

Part D: Repeat the previous sections using the SKlearn library.