Technische Universität Braunschweig

Institute of System Security

Exercise Sheet 2

Machine Learning for Computer Security

In this exercise, we will review the python unit and implement an estimate of an unknown probability distribution based on observed data to further test our python skills.

Exercise 2 *Numpy and Matplotlib*

Recreate the plots shown in Figure 1 as closely as possible.

(a) (b) (c)

Figure 1: Recreate these plots as closely as possible.

1. Recreate the plot shown in Figure 1a. To do so, you need to superposition two Gaussian distributions, with a standard deviation of 0.5 and 2.0 respectively. The first Gaussian is centered at $(-5.0, -5.0)$ while the second is centered at $(5.0, 0.0)$.
2. Recreate the plot shown in Figure 1b by implementing a function that splits the image into 4×4 squares and inverts every second piece pixelwise. You will find the required image data here: https://www.sec.tu-bs.de/teaching/ss22/mlsec/mlsec-exer02.tar.
3. Recreate the plot shown in Figure 1c. Use as few instructions as possible to draw a circle of different diameters. Make sure the circles are perfectly centered. *Make use of functions offered by Numpy only.*

Exercise 3 *Density estimation*

Density estimation is a classical approach of unsupervised machine learning that is used to construct a simple model for unlabeled data. We assume that a finite set of samples $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ has been drawn from the unknown probability distribution P. Our goal is to find a good approximation to P solely based on X. These are the necessary steps:

1. We choose P to be a Gaussian distribution with a mean value of 0.2 and a variance of 1. Draw 1000 samples from this distribution as your data set X. Verify that the samples are normally distributed by calculating the mean and variance for X.

2. To approximate P on the interval $[-3; 3]$ generate 1000 equally spaced values in this interval and store the resulting set in a variable named S.

3. For a fixed s from S, derive an expression $C_h(s, X)$ to calculate the number of elements from X that are located within a window of size h centered at s.

4. The density estimate at a point s is then obtained by dividing the number of data points in the window by the size of the window and the number of data points, or formally

$$\hat{P}_h(s) = \frac{1}{n}\frac{1}{h} \cdot C_h(s, X)$$

   Calculate and plot three estimates of $\hat{P}_h$ using different values for h to illustrate overfitting and underfitting. Label and explain the plots accordingly.