Machine Learning

Question 1

The goal of this question is to investigate underfitting and overfitting. First, generate the following data using the given code:

```
X = np.arange(-10,10,0.2)
Y = 2*cos(x)/-pi + (2*x)/(2*pi)+2*cos(3*x)/(-3*pi)
```

Now, add white Gaussian noise to the data in the first case, and in the second case, add Poisson noise with $\lambda = 2$. Try to fit a polynomial function of degree 1 to 15 to the data. Determine the best and worst degree. Plot the fitted curves and report the MSE values for degrees 1, 3, 8, and 15. Describe your observations by specifying the bias and variance values.

Question 2

In this question, generate data for the two following cases. In both cases, we have two sets of points with coordinates (X, Y):

- **First case:** The first set contains 200 points within a circle of radius 9 and center $(1.5, 0)$, and the second set contains 200 points within a circle of radius 6 and center $(1.5, 0)$.
- **Second case:** The first set contains 100 random points with a mean of $(1, 0)$ and standard deviation of 1, and the second set contains 200 points within a circle of radius 6 and center $(1.5, 0)$.

Plot the data for both cases. Use logistic regression with L_2 regularization to separate the two classes. Since the data is not linearly separable, increase the feature dimensions as shown below:

$$X = [x_1, x_2], \quad f(X) = [x_1, x_2, x_1^2, x_2^2, x_1x_2, \ldots]$$

Report the classification accuracy and decision boundary for both cases.

Question 3

In this question, we will implement a non-parametric Parzen estimation method. First, without using pre-built machine learning packages, implement the requested algorithm on

the existing dataset. Implement the Parzen window estimation for the `duration` column using a Gaussian kernel. Use a window size of 10. Compare the results using three different window sizes: 20, 50, and 100.

Plot the distribution as the value of n increases and show the convergence process. Finally, compare your results with the library functions for Parzen estimation.