Homework 4 Simulation Exercises

Exercise 1

In this exercise, the Iris dataset is going to be used. The dataset contains three classes, each of which shows one kind of flower.

Part A: Separate the samples using their first two features (sepal.length and sepal.width) and the linear kernel. Display the confusion matrix and mention the output's accuracy, precision, and f1-score.

Part B: Search about non-linear kernels (RBF, polinomial) and explain when using them can be helpful. Employ these non-linear kernels for part A.

Part C: Search about c and γ hyperparameters and explain how they can help the SVM from getting overfitted. Set three different values for each of these hyperparameters and explain how they affect the output.

Part D: Employ the grid search method to find the best classifier. Let the grid search find the best c, γ, degree, and kernel type.

Part E: When carrying out classification for multiple classes, two approaches can be used (one-vs-one, and one-vs-rest). Employ these approaches and display their respective boundaries.

Part F: Repeat the grid search method but this time with regard to the third and forth features (petal.length and petal.width). Explain why separating based on these two features leads to better results.

Exercise 2

Part A: In this exercise, we want to predict the salary attribute of the dataset named Q7-part1.csv using SVR (Support Vector Regression). Employ different kernels (linear, RBF, polinomial) used in the previous section. Display the real and predicted values and explain which kernel has resulted in more accurate predictions.

Part B: We want to Use a SVR model to predict the residence cost of customers in a hotel. The dataset is included in the HW folder, use H1.zip as the training set and H2.zip as the test set. At first, detect the categorical and numerical attributes, then do some preprocessing for each of them. Finally, train the SVR with H1.csv and record its predictions for H2.csv in a output.csv file.