Alexandria University
Faculty of Engineering
Computer and Systems Engineering Dept.

Prof. Dr. Mohamed Abougabal
Prof Dr. Nagwa M. El-Makky
Special Topics in Information Systems

Assignment 3
Data Classification

Objectives

1. Exploring different classification models and tune their hyper-parameters
2. Exploring different techniques for evaluating classification models
3. Learning how to analyze observed results and explain observations in a detailed report.

Problem Statement

As in the previous Lab, you are given the Wisconsin Diagnostic Breast Cancer (WDBC) Data Set. The dataset can be found in wdbc.data and the description can be found in wdbc.names.

It is required to split this dataset into training and testing datasets and apply any required preprocessing techniques. Then, you are required to construct classification models using different approaches, such as **K-Nearest Neighbor, Linear SVM, Nonlinear SVM and Logistic Regression** and tune the hyper-parameters of these models. You are also required to compare the performance of the models with each other.

Lab Session

1. Splitting and Preprocessing

As in the previous Lab, use Stratified Splitting to split the dataset such that the training data would form 70% of the dataset and the testing data would form 30% of it. Then, apply any required preprocessing techniques (e.g., feature normalization, feature selection/ feature projection, etc.) on both the training and testing sets.

2. Classification

In this step, it is required to apply the following classification models and to perform hyper-parameter tuning using **cross-validation**.

a. Classification models:

You are required to build models using the following classification techniques:

- **K-Nearest Neighbor:**

Parameters to be tuned: `n_neighbors`
Hint: use **`KNeighborsClassifier`** from **`Sklearn`**

- **Linear SVM:**

Parameters to be tuned: `C`
Hint: use **`LinearSVC`** from **`Sklearn`**

- **Non Linear SVM (with RBF Kernel)**

Parameters to be tuned: `C, gamma`
Hint: use **`SVC`** from **`Sklearn`**

- **Logistic Regression:**

Parameters to be tuned: `C`
Hint: use **`LogisticRegression`** from **`Sklearn`**

b. Hyper-parameter tuning:

Similar to the previous lab session, use the **cross-validation** approach on the pre-processed training dataset (as discussed in class) to get the best parameter values for each classifier. Test the models trained with best obtained parameter values on the separate pre-processed testing set.

Hint: you can use **`GridSearchCV`** from **`Sklean`** to perform parameter tuning.

c. Evaluation:

Report for each model the following performance metrics: **precision, recall, and F-measure** as well as the resultant confusion matrix using the test data.

Report Requirements

Your report should contain the following:

- Comparison and analysis of the performance results obtained in the evaluation part
- Analysis of the effect of hyper-parameter tuning on the performance of different classifiers

Notes

- This lab session needs time. So, try to start working on it early.
- Use Google Colab.
- You should work in groups of two. Each student should answer any question in the lab session.
  - You should deliver a well-documented code as well as a report showing all your work and conclusions.
  - Copied assignments will be penalized; so not delivering the assignment would be much better.
- You should write your code in python.

References

[1] Chapter 9 of the first reference (J. Han, M. Kamber, and J. Pie, “Data Mining: Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012).

[2] Chapter 21 of the third reference (M. Zaki and M. Wagner, “Data Mining and Analysis: Fundamental Concepts and Algorithms”, Cambridge Univ. Press, 2014).

[3] Dua Dheeru and Efi Karra Taniskidou. 2018. UCI Machine Learning Repository. (2018). http://archive.ics.uci.edu/ml

[4] http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf