COMP 4983: Lab Exercise #10 Mark: /35

[Due: Nov 18, 2022 @2359
Assignment Submission
Folders]

Instructions:

In this lab, you will

- perform K-means clustering on paper for a trivial dataset
- compare the classification performance of the support vector classifier (SVC) and the support vector machine (SVM) on a dataset

Part 1: K-means Clustering (on paper)

In this part of the lab, you will perform K-means clustering for a trivial dataset.

Consider a dataset consisting of the following six (6) samples. Perform K-means clustering with K = 2.

Sample	(X_1, X_2)
x_1	(1, 4)
x_2	(1, 3)
x_3	(0, 4)
x_4	(5, 1)
x_5	(6, 2)
x_6	(4, 0)

a) Plot the samples.
b) Assign samples with an odd-numbered index (i.e., i = {1, 3, 5}) to the first cluster and samples with an even-numbered index (i.e., i = {2, 4, 6}) to the second cluster. State the cluster assignment, $C(i)$, for each sample.
c) Compute the centroid for each cluster.
d) Compute the squared Euclidean distance between each sample and each centroid and assign each sample to the cluster whose centroid is closest. State the cluster assignment $C(i)$ for each sample.
e) Repeat c) and d) until there are no further changes to the cluster assignments. State the final cluster assignment, $C(i)$, for each sample.

Part 2: Support Vector Machine

[35 marks] In this part of the lab, you will compare the classification performance of the support vector classifier (SVC) and the support vector machine (SVM) with the radial basis kernel function on a dataset. In addition, you will determine the best value of the cost parameter, C, using 10-fold cross-validation on the training set and evaluate the error rate (percentage of misclassifications) of SVC and SVM on the test set.

Steps:

1) Download the dataset, *data_lab10.csv*, from BCIT Learning Hub (Content | Laboratory Material | Lab 10) and save it in your working directory. The dataset, *data_lab10.csv*, contains 401 rows (including a header row) and 3 columns. Each row contains two features followed by the class label.
2) Download a Python script, *SVM_lab10.py*, from BCIT Learning Hub (Content | Laboratory Material | Lab 10) and save it as *SVM_lab10.py* in your working directory. This script contains the function **plot_svc_decision_function()**, which plots the decision boundary and the margins of a SVC.
3) Add to your script, *SVM_lab10.py*, to read from *data_lab10.csv*.
4) Split the dataset into training and test sets, with the first 75% of the dataset for training and the remaining 25% for testing.
5) For each $C = [0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 100, 1000]$ (which is referred to as the penalty parameter in **sklearn.svm.SVC**, apply SVC on the training set and evaluate the average cross-validation estimate of prediction error using 10-fold cross-validation. Ensure that the argument **kernel='linear'** is specified when instantiating **sklearn.svm.SVC**. Plot the average cross-validation estimate of prediction error as a function of C. Include in your plot, a terse descriptive title, x-axis label, y-axis label and a legend.
6) Determine the best value of C from Step (5).
7) Using the best value of C, evaluate and output the error rate (percentage of misclassifications) on the test set.
8) Plot the samples from the test set, as well as the decision boundary and the margin of the SVC from Step (7). Include in your plot, a terse descriptive title, x-axis label, y-axis label and a legend.
9) Repeat Steps (5) to (8) for the SVM with the radial basis kernel function. Ensure that the argument **kernel='rbf'** is specified when instantiating **sklearn.svm.SVC**.

Deliverable:

All work submitted is subject to the standards of conduct as specified in BCIT Policy 5104. No late assignments will be accepted.

[Nov 18, 2022 @2359] Ensure that your source code for Part 2 is adequately commented and submit using the filename *SVM_lab10.py* to BCIT Learning Hub (Laboratory Submission | Lab 10).