COMP 4983: Lab Exercise #10 Mark: /40

[Due: Mar 30, 2020 @1730
Assignment Submission
Folders]

Name:
BCIT ID:
Lab Set: 3M

Instructions:

In this lab, you will

- **construct the maximal margin classifier on paper for a trivial dataset**
- **investigate the effect of the cost parameter, C, for the support vector classifier (SVC) as well as determine the best value of C using 10-fold cross-validation**

Part 1: Maximal Margin Classifier (on paper)

[20 marks] In this part of the lab, you will

- **draw a hyperplane in a 2-dimensional space**
- **construct a maximal margin classifier for a trivial dataset**

1) **[4 marks] Draw the hyperplane $1 + 3X_1 - X_2 = 0$. Indicate the region for which $1 + 3X_1 - X_2 > 0$, as well as the region for which $1 + 3X_1 - X_2 < 0$.**
2) **Consider a training set consisting of the following seven (7) training samples:**

Sample	Input Vector	Output Value
x_1	(3, 4)	Red
x_2	(2, 2)	Red
x_3	(4, 4)	Red
x_4	(1, 4)	Red
x_5	(2, 1)	Blue
x_6	(4, 3)	Blue
x_7	(4, 1)	Blue

   a) **[4 marks] Plot the training samples and draw the maximal margin hyperplane given the training samples.**
   b) **[1 mark] Indicate the margin, M, for the maximal margin hyperplane.**
   c) **[10 marks] Derive the equation for the maximal margin hyperplane in the form of $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ subject to $\sum_{j=1}^{p} \beta_j^2 = 1$.**
   d) **[1 mark] Predict the output value for the test sample (3.5, 2).**

Show all your steps and add comments as necessary to make sure your answers are clear and unambiguous.

Part 2: Support Vector Classifier

[20 marks] In this part of the lab, you will

- **perform classification using the SVC.fit() and SVC.predict() function from sklearn.svm on a dataset**
- **determine the best value of the cost parameter, C, using 10-fold cross-validation on the training set**
- **evaluate the error rate (percentage of misclassifications) of the SVC classifier 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 201 rows (including a header row) and 3 columns. Each row contains two features followed by the class label.**
2) **Create a new Python script using the filename *SVC_lab10_lastname_firstname.py* and save it in your working directory.**
3) **Add to your script, *SVC_lab10_lastname_firstname.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 $C = [0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 100]$ (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.**

Deliverables:

All work submitted is subject to the standards of conduct as specified in BCIT Policy 5104. Late submissions will not be accepted.

[Mar 30, 2020 @1730]

- **Submit your scanned solution to Part 1 of this lab exercise to BCIT Learning Hub (Laboratory Submission | Lab 10). Your submission must include a cover page clearly specifying your name and student number.**
- **Ensure that your source code for Part 2 is adequately commented and submit using the filename *SVC_lab10_lastname_firstname.py* to BCIT Learning Hub (Laboratory Submission | Lab 10).**