**Instructions**: Please complete and submit your work to the appropriate folder in LumiNUS. You may work in study groups, but each student must be responsible for their own submission.

Please submit all the following documents as a single zip file named StudentID-Name-H3.zip:

1. Completed Word file named as StudentID-Name-H3.docx (with all results)
2. Print preview of ipynb file named as StudentID-Name-H3.pdf (with results)
3. Working ipynb file named as StudentID-Name-H3.ipynb
4. Consider building an SVM classifier for the following two-class training data:

Positive class: { (-1, 3) (0, 2) (0, 1) (0, 0) }; Negative class: { (1, 5) (1, 6) (3, 3) }

1. Plot the training points. Use ‘+’ for positive class and ‘o’ for the negative class.
2. By inspection, draw a linear classifier that separates the data with maximum margin.
3. The linear SVM is parameterized by h(x) = (**w**^t)(x) + b. What are the parameters **w** and b for this problem?
4. Suppose you observe an additional set of points, all from the positive class.

Additional data points in positive class: { (−2, 0) (−2, 1) (−2, 3) (−1, 0) (−1, 1) }

A picture containing clock

Description automatically generatedWhat is the linear SVM (in terms of **w** and b) now?

1. Consider the dataset on the right. Consider using the SVM with soft margin classifier with parameter C.
   1. Draw the linear classifier when C is large.
   2. Draw the linear classifier when C is small.
   3. Which value of C yields the classifier most closely resembling the hard margin SVM solution?
   4. Using your two examples, explain how the C parameter helps with overfitting in SVMs.
2. In this problem, we will look at the Breast Cancer Wisconsin (Diagnostic) Data Set available UCI Machine Learning Repository. Please use the wdbc.data dataset from:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

* Compute the performance of the SVM algorithm on this dataset for predicting the whether the cancer is malignant orx benign. Use a random train/test data split of 70%/30%. Repeat this process 20 times and compute the average performance.
* Please evaluate the following algorithms:
* SVM1: SVM with linear kernel
* SVM2: SVM with RBF kernel
* SVM3: Same as SVM2 but with regularization (soft margin), vary C and report your best results.
* Please compute the following metrics and fill in the table below.
* Training Accuracy and Test Accuracy
* Precision and Recall (which are important metrics that complement Accuracy)
* You can read about performance metrics at: <https://en.wikipedia.org/wiki/Confusion_matrix>
* SKLearn contains functions to compute these metrics:

<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>

A graph on a white board

Description automatically generated

1a/b)By inspection, the red line indicates the linear classifier with maximum margin that separates the two different classes into their respective classes.

1c)

A white paper with writing on it

Description automatically generatedA piece of paper with writing on it

Description automatically generated

Q1d) Supposed having the additional data points from the datasets we can plot the diagram out as seen below.

A graph on a piece of paper

Description automatically generated

Based on the above with the additional data points the Support vectors remain the same hence the linear classifiers values of w and b stay consistent throughout even with the additional pointers.

Q2a/b)

A paper with writing on it

Description automatically generated

Q2c) The high value of C closely resembles the hard margin SVM solution as with the higher C values the linear classification will be as precise as possible but may lead to overfitting as a result.

Q2d) The C Parameters help controls or serves as a regularization factor for the boundary of the linear classifier, and by referencing both examples of hard margin and soft margin respectively we can conclude that the C factor when reduced from a high level will prevent overfitting of the data, in this case we would want to avoid using example 2a of having a high C level as it might cause overfitting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | | Precision | Recall |
|  | Train | Test |  |  |
| SVM1 | 1.0 | 0.947368 | 0.911634 | 0.951289 |
| SVM2 | 1.0 | 0.953216 | 0.925468 | 0.950779 |
| SVM3  C = 10 | 0.992462 | 0.970175 | 0.97054 | 0.949447 |

Comment: When varying the C value it was found that when C was at 10 it gave the best testing and precision results.