Name: EE4211 Homework H3, Fall 2020

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-HW1.zip:

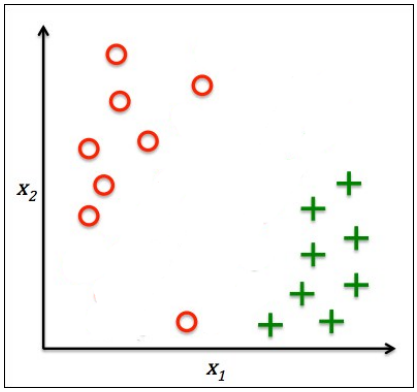
(i) Completed Word file named as StudentID-Name-HW3.docx (with all results)
(ii) Print preview of ipynb file named as StudentID-Name-HW3.pdf (with results)
(iii) Working ipynb file named as StudentID-Name-HW3.ipynb

1. 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) }

a. Plot the training points. Use '+' for positive class and 'o' for the negative class.
b. By inspection, draw a linear classifier that separates the data with maximum margin.
c. The linear SVM is parameterized by h(x) = (**w**^t)(x) + b. What are the parameters **w** and b for this problem?
d. 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) }
What is the linear SVM (in terms of **w** and b) now?

2. Consider the dataset on the right. Consider using the SVM with soft margin classifier with parameter C.

a. Draw the linear classifier when C is large.
b. Draw the linear classifier when C is small.
c. Which value of C yields the classifier most closely resembling the hard margin SVM solution?
d. Using your two examples, explain how the C parameter helps with overfitting in SVMs.



3. 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 or 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

	Accuracy		Precision	Recall
	Train	Test		
SVM1				
SVM2				
SVM3 C =				