## SENG 474, CSC 503: Assignment 2

**1. (6 pts)** Complete the **students\_post.ipynb** notebook about Logistic Regression.

**2. (9 pts)** Consider the dataset in Fig 1, with points belonging to two classes, blue squares and red circles.

**Fig. 1**

1. [1 pt] Draw (approximately) the SVM line separator.
2. [1 pt] Suppose we find (1/2)\***w**2 to be 2 in the SVM optimization. What is the margin, i.e. the distance of closest points to the line?

**Fig. 2 Fig. 3**

1. [1 pt] Now consider the dataset in Fig 2 (the red points are shifted below). Will (1/2)\***w**2 be smaller or greater than previously? Explain.
2. [2 pt] Using a ruler, and the fact that (1/2)\***w**2 was 2 previously, find (approximately) the magnitude of the new line coefficient vector, **w**’.
3. [3 pt] Consider the dataset in Fig 3 (with one additional red circle quite close to the blue squares). Assuming optimization using slack variables and C=1, draw a line that does not perfectly separate the points, but which is nonetheless better than the line that perfectly separates the points. (Draw it in the figure, and explain why).
4. [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?

**3. (5 pts)** Adapt the Text\_Classification.ipynb notebook to build a classifier for the following tweet dataset. The dataset contains tweets pertaining to disasters and non-disasters. Print the classification report after splitting into a train and test dataset similarly to the mentioned notebook.

<https://raw.githubusercontent.com/nikjohn7/Disaster-Tweets-Kaggle/main/data/train.csv>

You should submit your notebook and a pdf printout.

**4. (6 pts)** Construct the root and the first level of a decision tree for the titanic dataset. Use entropy to decide splits. Show the details of your construction (entropies calculated for each step). You can use a spreadsheet to compute the counts.