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.

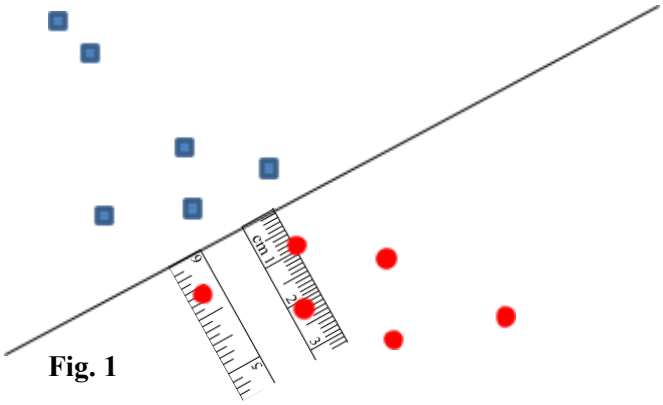

Fig. 1

(a) [1 pt] Draw (approximately) the SVM line separator.

(b) [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?

.5w^2 = 2
w^2 = 4
w = +2, -2
margin = 1/||w||
margin = 1/2 = 0.5

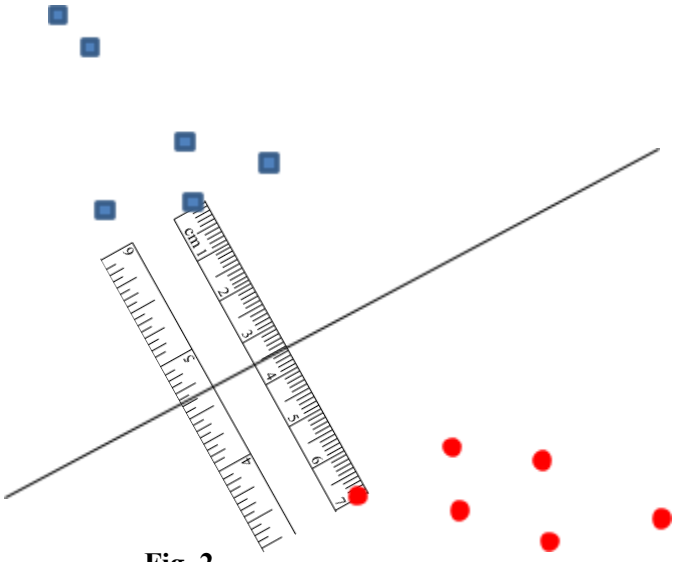

Fig. 2

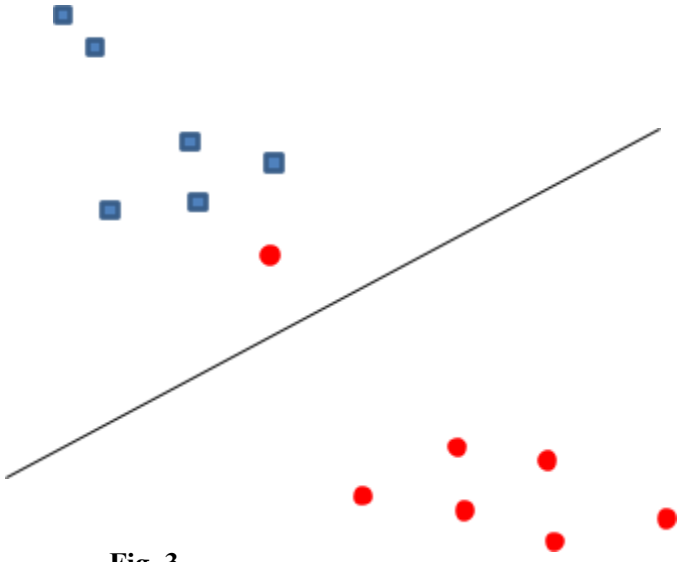

Fig. 3

(c) [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.

It will be greater because the SVM line separator will have a larger margin on either side as the closest support vector points that we use to form the line are further away from each other.

(d) [2 pt] Using a ruler, and the fact that (1/2)*$\mathbf{w}^2$ was 2 previously, find (approximately) the magnitude of the new line coefficient vector, **w**'.
The first margin is about 0.5 cm, the second margin is about 3.5, so 7x larger. This means that the new margin is 1/||w|| = (.5 * 7)
1 / ||w|| = 3.5
||w|| = 1 / 3.5
||w|| = 0.2857

(e) [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).
I drew the line slightly closer to the blue dots than in Fig. 2. The slack variables and C=1 allow us to soften the margin such that the outlier red dot doesn't determine the margin on its own. It is better than a line that perfectly separates all the points because it will likely generalize better to unseen data.

(f) [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?
The optimal line in this case would likely be close to the line that perfectly separates the points but with a slight movement towards the blue points. This will misclassify one of the red circles but otherwise maximizes the margin for the rest of the data set. This is based on the principle that a model which fits the training data too well may not perform well on unseen data. This will hopefully allow the model to generalize better to unseen data compared to the line that perfectly separates all 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.