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.

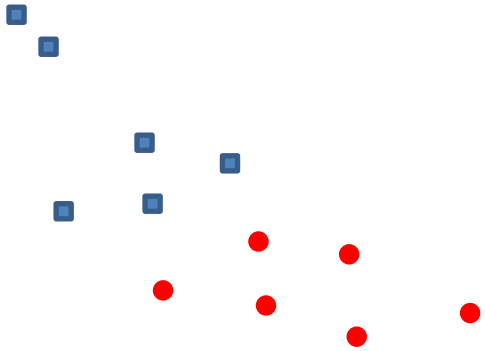
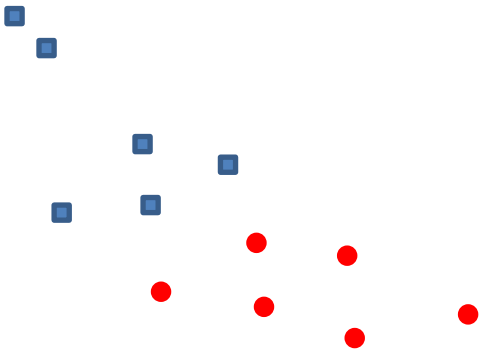


Fig. 1

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



(b) [1 pt] Suppose we find (1/2)*$\mathbf{w}^2$ to be 2 in the SVM optimization. What is the margin, i.e. the distance of closest points to the line?

In SVM, the margin is given by:
Margin = (1/ ‖ w‖) = 1 / [(1/2)w^2]

Given that (1/2)w^2 = 2, we can substitute which will give:
Margin = 1/2 = 0.5

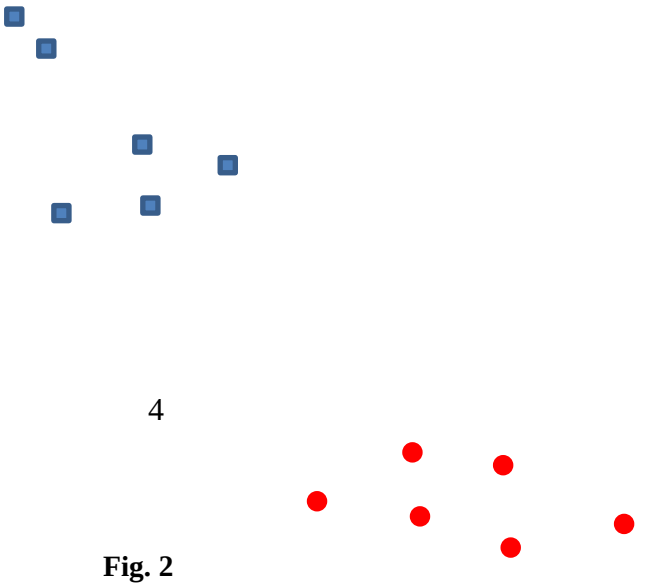


Fig. 2

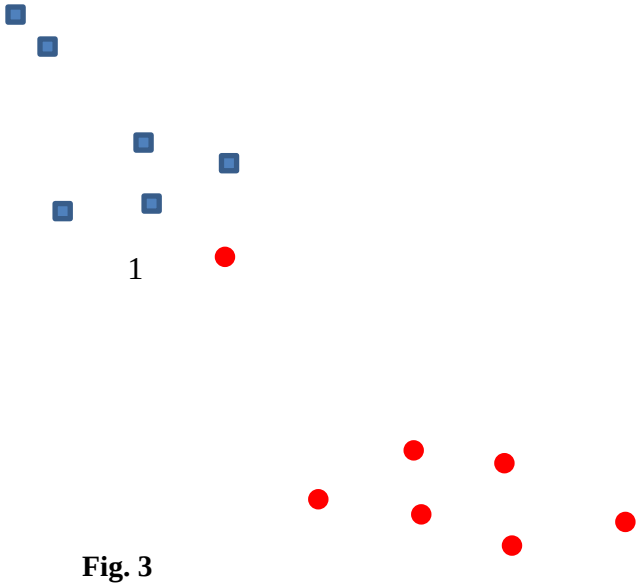


Fig. 3

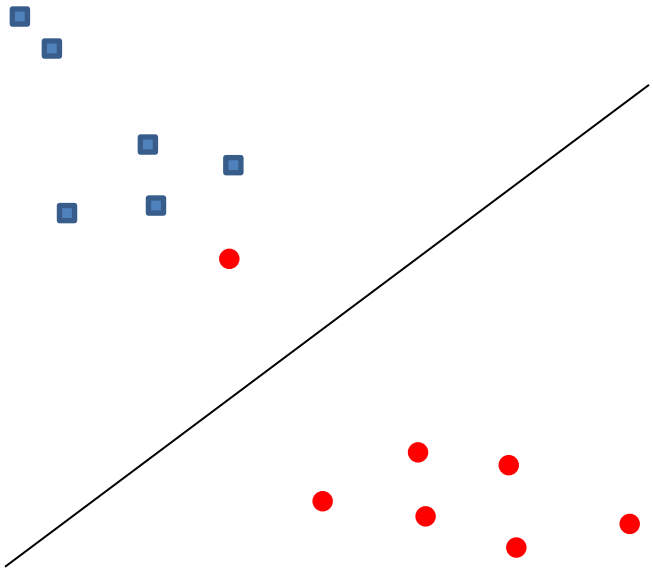
(c) [1 pt] Now consider the dataset in Fig 2 (the red points are shifted below). Will (1/2)*$\mathbf{w}^2$ be smaller or greater than previously? Explain.

If the red points are shifted below, the classes may become easier to separate, or the margin may decrease. Typically, if points are shifted closer to the separating line, the margin will decrease, meaning that the distance between the support vectors and the decision boundary will also decrease. Therefore, (1/2)w^2 is likely to increase, reflecting a smaller margin.

(d) [2 pt] Using the given distance, and the fact that (1/2)*$\mathbf{w}^2$ was 2 previously, find (approximately) the magnitude of the new line coefficient vector, **w**'.

Let's assume that the margin has decreased due to the red points being shifted. If the new margin is m′, we know from SVM theory that:
m′=1/||w′||
If we previously had (1/2)*(w^2)=2, then ||w|| was calculated as 1/m = 2 (since m=0.5m = 0.5m=0.5). If the new margin m′ is now less than the previous margin (for example, suppose it decreases to (0.25), we can determine the new weight vector magnitude:
||w'|| = 1/m′ = 1/(0.25) = 4.

(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). Compute the value of the objective function for each to see mathematically why.



(f) [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?

We would prefer the line in (e) to the line that perfectly separates the points because it achieves a larger margin. A larger margin is desirable in SVM because it can lead to better generalization performance, meaning the model is more likely to correctly classify new, unseen data 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 or Python to compute the counts. However, you should write down in your solution what you computed. Do not submit the spreadsheet or Python program.

Total rows: 2201
Survived: 711
Not survived: 1490
Entropy calculation:
P(Survived) = 711/2201 = 0.3228
P(Not Survived) = 1490/2201 = 0.6772
Entropy = -0.3228log2(0.3228) - 0.6772log2(0.6772) = 0.9183
Best split is on "Sex" reduces to: 0.2065

Male--
Instances: 1667
Survived: 367
Not survived: 1300
Entropy calculation:
P(Survived) = 367/1667 = 0.2202
P(Not Survived) = 1300/1667 = 0.7798
Entropy = -0.2202log2(0.2202) - 0.7798log2(0.7798) = 0.6887
Best split is on "Age" which reduces entropy by 0.0686

Splitting "Male" node on "Age" into:

Instances: 534
Survived: 344
Not survived: 190
Entropy calculation:
P(Survived) = 344/534 = 0.6441
P(Not Survived) = 190/534 = 0.3559
Entropy = -0.6441log2(0.6441) - 0.3559log2(0.3559) = 0.7775
Best split is on "Pclass" reduces by 0.1505

5. (5 pts) Classify using Naïve Bayes method on the titanic dataset the data items:

```
2nd     child   male    ?
2nd     adult   female  ?
```

You can use a spreadsheet or Python to compute the counts. However, you should write down in your solution what you computed. Do not submit the spreadsheet or Python program.

- For P(2nd, child, male|Survived) = 11/711 = 0.0155 approx

Overall for this is P(2nd, child, male) = 11/2201 = 0.005 approx

Applying Bayes P(Survived|2^{nd}, child male) =
P(2nd, child, male|Survived)×P(Survived) / P(2nd, child, male) = 0.0155 x 0.323 / 0.005 = 1.0 approx which means 100% for survival for a 2nd class, child, male.

- For P(2^{nd} , adult, female | Survived) = 80/711 = 0.01125 approx
  For P(2^{nd}, adult, female| Not Survived) = 13/1490 = 0.0087 approx

  P (2^{nd}, adult, female) = 80 + 13 / 2201 = 0.0424 approx

  Applying Bayes wil have P(Survived | 2^{nd} , adult, female) = 0.1125 x 0.323 / 0.0424 = 0.8602 approx

  This means 86% of females have survived