1)

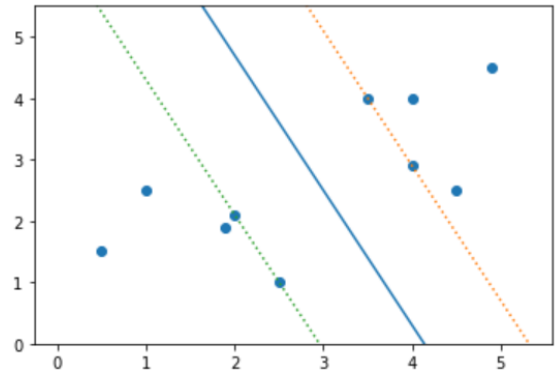
a) h(x) is the hyperplane equation the code used to find and plot this is in the file called SVMs

$h(x) = w^T x + b$

$h(x) = [0.846,\ 0.3852]^T x - 3.50108$

```
w = [0.846, 0.3852]
b = -3.50108
f(x) = x2 = -2.196262x1+9.088993
f(x)- = x2 = -2.196262x1+11.685047
f(x)+ = x2 = -2.196262x1+6.492939

h(x) = -2.196262x1+9.088993-x2
```




```
point x_6 has distance: -1.2498290534100407
point 3,3 has distance: 0.2071071521453788
```

b) it is not within the margin of the classifier and is 1.249829 units from the hyperplane in the direction of the negative class

c) the point 3,3 belongs to the positive class

2) The SVM loss function with slack variables can be viewed as:

$$\min_{\mathbf{w},b} \frac{||\mathbf{w}||^2}{2} + \gamma \sum_{i=1}^{N} \underbrace{\max(0, 1 - y_i f(x_i))}_{\text{Hinge loss}}$$

The hinge loss provides a way of dealing with datasets that are not separable.

a) Argue that $l = \max(0, 1 - yw^\top x)$ is convex as a function of w.

$$\text{if } f''(x) \geq 0 \text{ for all } x \text{ in } I \text{ then } f(x) \text{ is convex.}$$

$$\text{if each part of the max function is convex the whole function is convex}$$

$$l''(w) = \frac{\partial}{\partial w}\frac{\partial}{\partial w} l =$$

$$\frac{\partial}{\partial w}\frac{\partial}{\partial w} \max(0, 1 - yw^\top x) =$$

$$\frac{\partial}{\partial w}\frac{\partial}{\partial w} 0 \leq 0 \rightarrow 0 \leq 0 \rightarrow \text{True}$$

$$\text{and}$$

$$\frac{\partial}{\partial w}\frac{\partial}{\partial w} 1 - yw^\top x \leq 0$$

$$\frac{\partial}{\partial w} 0 - yx =$$

$$0 - 0 = 0 \rightarrow$$

$$0 \geq 0, \text{Therefore } l \text{ is convex as a function of } w$$

b) Suppose that for some w we have a correct prediction of f with x_i, i.e. $f(x_i) = \text{sgn}(w^\top x_i)$. For binary classifications ($y_i = +1/-1$), what range of values can the hinge loss, l, take on this correctly classified example? Points which are classified correctly and which have non-zero hinge loss are referred to as margin mistakes.

$$l = \begin{cases} 1, \text{ if } w^\top x_i = 0 \\ 0, \text{ otherwise} \end{cases}$$

$$Range(\sum_{i=1}^{N} \underbrace{\max(0, 1 - y_i f(x_i))}_{l}) = [0, N]$$

$$\frac{||\mathbf{w}||^2}{2} + \gamma \times [0, N] =$$

$$[\frac{||w||^2}{2}, \lambda N + \frac{||w||^2}{2}]$$

$$\text{final answer} = l = \mathbb{R} \in [0, 1]$$

c) Let $M(w)$ be the number of mistakes made by w on our dataset (in terms of classification loss). Show that:

$$\frac{1}{n} M(w) \leq \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i w^\top x_i)$$

In other words, the average hinge loss on our dataset is an upper bound on the average number of mistakes we make on our dataset.

$$\frac{1}{n} M(w) \leq \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i w^\top x_i) =$$

$$M(w) \leq \sum_{i=1}^{n} \max(0, 1 - y_i w^\top x_i) =$$

$$\sum_{i=1}^{n} |y_i - \text{sgn}(w^\top x_i)| \leq \sum_{i=1}^{n} \max(0, 1 - y_i w^\top x_i) =$$

$$z \leq 2z, \text{let } z = \mathbb{R} \in [0, N]$$

when there is an incorrect sample prediction mistakes will increase by 1 and loss will increase by 2
when there is a correct sample prediction both the loss and number of mistakes increase by 0
the number of (in)corrrect predictions are the same because both use the same set of samples

3)

a) I preprocessed the features by removing ticket name cabin and passengerID as well as removing any data points that had missing values.

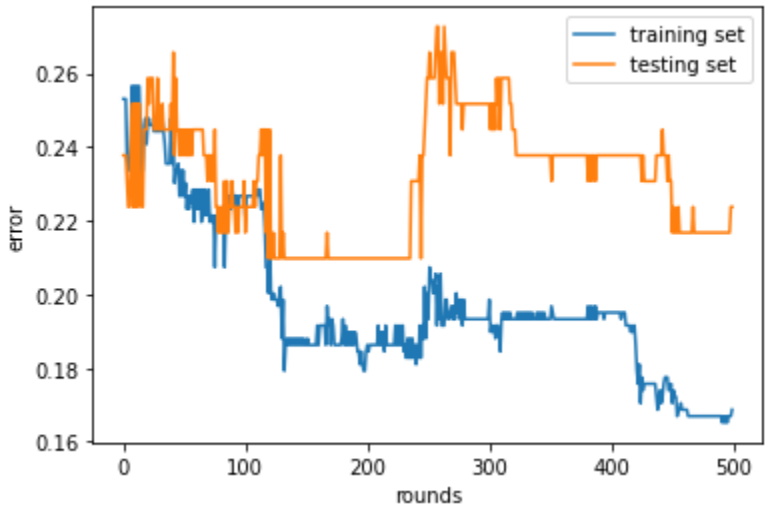
b) The accuracy of Gini-index and information gain both varied greatly based on the max depth and training set but with the max depth set to 2 for both and using the random seeds 0-4 for the test-train-split they mostly performed equally well with the gini index slightly higher.

c) My best observed accuracies were 0.818 and 0.811 for Info-Gain and Gini-Index, respectively at maxdepth=2 random_state=4.

d) My observations were the 3 most important features for the IG were generally the Age Sex and Fare of the passengers. The top 3 features used in the tree for the GI were Pclass Sex and Age. In conclusion, if you were a woman or a child the data suggests you likely would have survived but if you were wealthy that greatly improved your chances.

4)

a) Plot the train and test errors as a function of the number of rounds from 1 through 500.

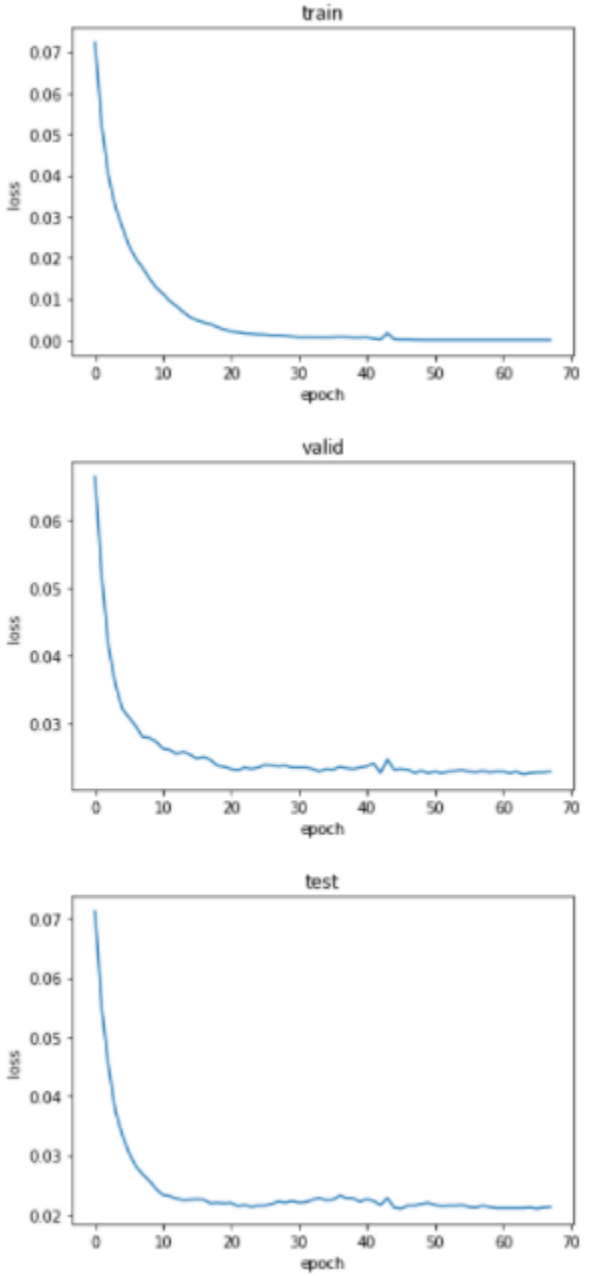


b) The best accuracy on the test data set with random state=1 max tree depth=2 was 0.776224 but at one point during training the model had an accuracy of 0.790210

c) From looking at the graphs we can see the accuracy is very volatile. This suggests high variance most likely from great differences in the lagrange multipliers.
Looking at the testing set we see there exists an overfitting error because we are iterating 500 times and the testing set is very small. We can solve this by setting a threshold for error and when the error is below that threshold stop training.

5)

a) Plot the train, validation, and test errors as a function of the epoches.



b) The best accuracy on the validation and test data sets were 0.9776 and 0.9790

c) From the plots we can see that overfitting is an issue even with an adaptive learning rate. Some possible reasons for this are: that the data early on is already trained to maximum efficiency for this task. Considering the number of neurons in a hidden layer when we have different numbers in these neurons it can vary output between the graphs and using a more complex approach to this small data set and straightforward task is reasonably going to cause the improvement in accuracy to decrease exponentially.