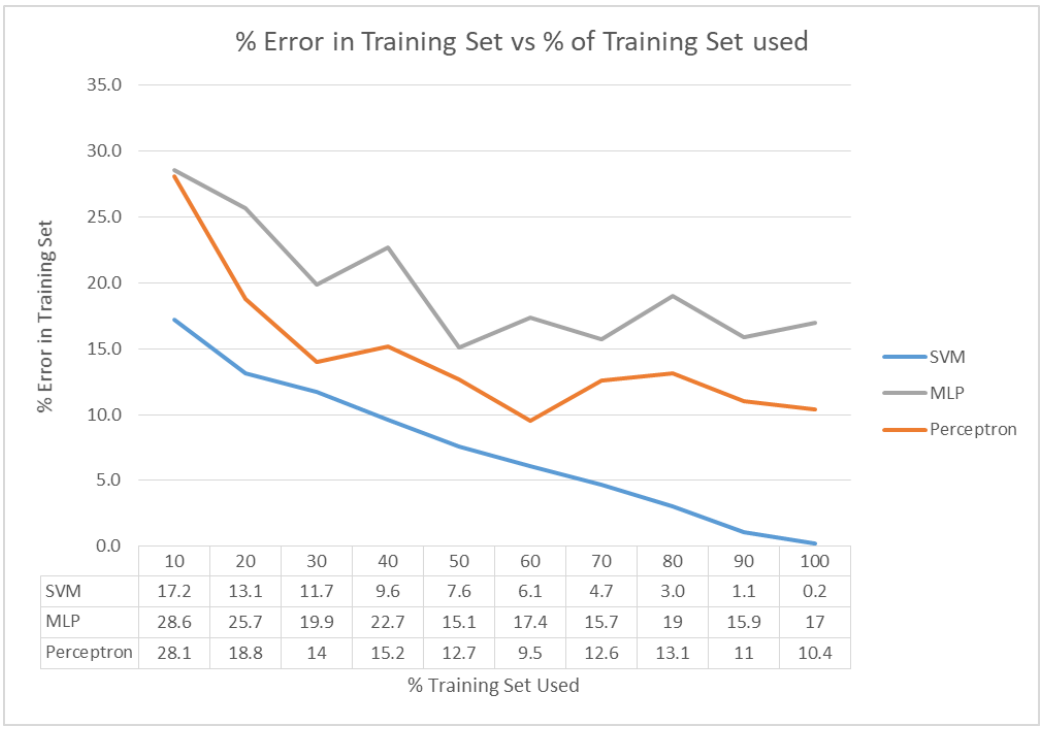
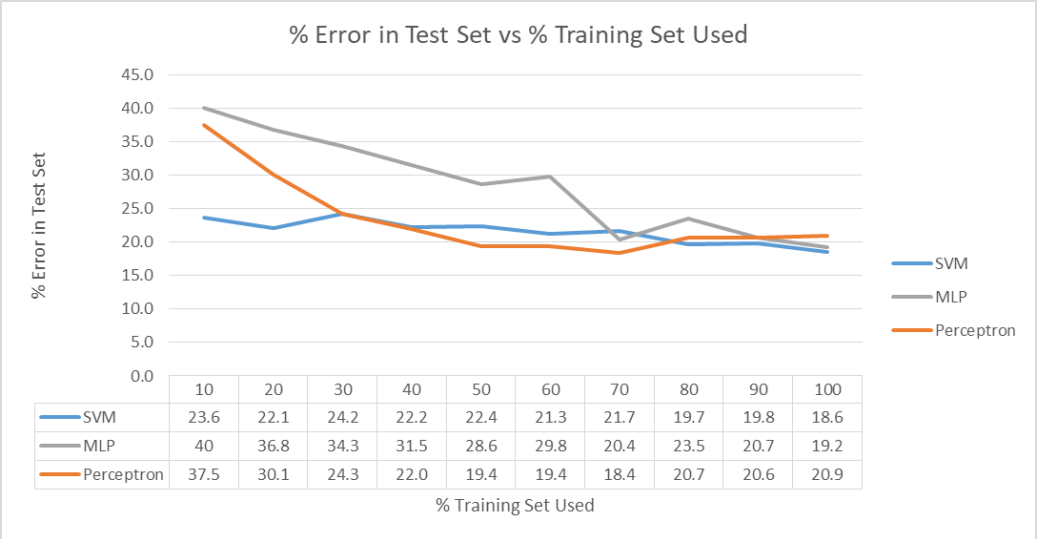
By Kevin Pei (ksp98) and Mengdan Liu (mml176)

# Question 1




| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| SVM | 17.2 | 13.1 | 11.7 | 9.6 | 7.6 | 6.1 | 4.7 | 3.0 | 1.1 | 0.2 |
| MLP | 28.6 | 25.7 | 19.9 | 22.7 | 15.1 | 17.4 | 15.7 | 19 | 15.9 | 17 |
| Perceptron | 28.1 | 18.8 | 14 | 15.2 | 12.7 | 9.5 | 12.6 | 13.1 | 11 | 10.4 |

Figure 1: A graph showing percent correct of the training set as a function of the percentage of the training set used.




| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| SVM | 23.6 | 22.1 | 24.2 | 22.2 | 22.4 | 21.3 | 21.7 | 19.7 | 19.8 | 18.6 |
| MLP | 40 | 36.8 | 34.3 | 31.5 | 28.6 | 29.8 | 20.4 | 23.5 | 20.7 | 19.2 |
| Perceptron | 37.5 | 30.1 | 24.3 | 22.0 | 19.4 | 19.4 | 18.4 | 20.7 | 20.6 | 20.9 |

Figure 2: A graph showing percent correct of the test set as a function of the percentage of the training set used.

Of the three algorithms tested, SVM appears to perform the best on both the training and test set, while MLP performs the worst. For perceptron, we took the dot product of the weights times the training data and selected the layer that had the highest dot product as our hypothesis. If the hypothesis was incorrect, then the training data was subtracted from the weights of the incorrect label and added to the weights of the correct label. For MLP, we used a softmax algorithm. The first layer takes the dot product of the weights times the training data and passes it as input to the softmax layer, which uses the following equation:

$$softmax(label = x) = \frac{e^{w_x t}}{\sum_{n=0}^{9} e^{w_n t}} \tag{1}$$

Where t is the training data and there are 10 labels, 0-9. Then, the highest value of the softmax layer was used as the guess. If the hypothesis was incorrect, then the error was backpropagated to the weights layer by multiplying $y - h_w$ by $s(1 - s)$, where y is the correct label, $h_w$ is the hypothesis, and s is the softmax value for that label. This was added to the weights for the correct and incorrect labels. SVM used the inbuilt SVM functions to maximize the margins between labels and guess that way. Overall, it seems like SVM overfit to the training data but still performed better than per-

ceptron and MLP. MLP was the least well fit to the training data but still performed reasonably well on test data, indicating that it was more generalized than SVM. If SVM was trained on more data, there is a chance that it would have overfit to the data and actually performed worse than perceptron and MLP. MLP and perceptron on the other hand seem to have stagnated when 100% data is used, showing very similar error rates from 50 - 100% of the training data used. This may indicate that further training data may not greatly improve performance for MLP or perceptron.

# Question 2

## Part a

1) 4.0 GPA - P, the data match the decision tree
2) 3.9 GPA - P, the data match the decision tree
3) 3.9 GPA - P, the data match the decision tree
4) 3.8 GPA - yes publications - P, the data match the decision tree
5) 3.6 GPA - no publications - rank 2 university - P, the data match the decision tree
6) 3.6 GPA - yes publications - P, the data match the decision tree
7) 3.4 GPA - no publications - rank 3 university - N, the data match the decision tree
8) GPA 3.4 - No publication - Rank 1 University - N, data match the tree
9) GPA 3.2 - N, data match the tree
10) GPA 3.1 - N, data match the tree
11) GPA 3.1 - N, data match the tree
12) GPA 3.0 - N, data match the tree

## Part b

For the entire set, there are 6 positives and 6 negatives. Hence, the entropy of the set is:

$$E(S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{2}$$

For GPA, the information gained is:

GPA ¿= 3.9: 3 P, 0 N

$$E(GPA >= 3.9) = -1\log_2 1 - 0\log_2 0 = 0 \tag{3}$$

3.2 ¡ GPA ¡ 3.9: 3 P, 2 N

$$E(3.2 < GPA < 3.9) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{4}$$

GPA ¡= 3.2: 0 P, 4 N

$$E(GPA <= 3.2) = -0\log_2 0 - 1\log_2 1 = 0 \tag{5}$$

$$I(GPA) = \frac{1}{4} * 0 + \frac{5}{12} * 0.9710 + \frac{1}{3} * 0 = 0.4046 \tag{6}$$

$$Gain(GPA) = E(S) - I(GPA) = 1 - 0.4046 = 0.5954 \tag{7}$$

For university rank, the information gained is:
University rank = 1: 3 P, 2 N

$$E(rank = 1) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{8}$$

University rank = 2: 2 P, 1 N

$$E(rank = 2) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183 \tag{9}$$

University rank = 3: 1 P, 3 N

$$E(rank = 3) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113 \tag{10}$$

$$I(rank) = \frac{5}{12} * 0.9710 + \frac{1}{4} * 0.9183 + \frac{1}{3} * 0.8113 = 0.9046 \tag{11}$$

$$Gain(rank) = E(S) - I(rank) = 1 - 0.9046 = 0.0954 \tag{12}$$

For whether the student has publications, the information gained is:
Published = Yes : 3 P, 2 N

$$E(Published = Yes) = -\frac{3}{5}\log_2\frac{3}{5}) - \frac{2}{5}\log_2\frac{2}{5} = .9710 \tag{13}$$

Published = No : 3 P, 4 N

$$E(Published = No) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = .9852 \tag{14}$$

$$I(Published) = \frac{5}{12}(.9710) + \frac{7}{12}(.9852) = 0.9792 \tag{15}$$

$$Gain(Published) = E(S) - I(Published) = 1 - 0.9792 = 0.0208 \tag{16}$$

For the quality of the student's recommendations, the information gained is:
Recommendations = Good : 5 P, 3 N

$$E(Recommendations = Good) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = .9544 \tag{17}$$

Recommendations = Normal : 1 P, 3 N

$$E(Recommendations = Normal) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = .8113 \tag{18}$$

$$I(Recommendations) = \frac{8}{12}(.9544) + \frac{4}{12}(.8113) = 0.9067 \tag{19}$$

$$Gain(Recommendations) = E(S) - I(Recommendations) = 1 - 0.9067 = 0.0933 \tag{20}$$

For the root of the decision tree, the best attribute to use is GPA, since it has the highest information gain of all the attributes. Because a GPA of 4.0 always leads to P and a GPA of 3.3 always leads to N, we only need to investigate the next node for when GPA = 3.6.
For the subset, there are 3 positives and 2 negatives. Hence, the entropy of the subset is:

$$E(S) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{21}$$

For university rank, the information gained is:
University rank = 1 : 1 P, 1 N

$$E(rank = 1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{22}$$

University rank = 2 : 1 P, 0 N

$$E(rank = 2) = -1\log_2 1 - 0\log_2 0 = 0 \tag{23}$$

University rank = 3 : 1 P, 1 N

$$E(rank = 3) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{24}$$

$$I(rank) = \frac{2}{5} * 1 + \frac{1}{5} * 0 + \frac{2}{5} * 1 = 0.8 \tag{25}$$

$$Gain(rank) = E(S) - I(rank) = 0.9710 - 0.8 = 0.1710 \tag{26}$$

For whether the student has publications, the information gained is:
Published = Yes : 2 P, 0 N

$$E(Published = Yes) = -1\log_2 1 - 0\log_2 0 = 0 \tag{27}$$

Published = No : 2 P, 1 N

$$E(Published = No) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183 \tag{28}$$

$$I(Published) = \frac{2}{5}(0) + \frac{3}{5}(.9183) = 0.5510 \tag{29}$$

$$Gain(Published) = E(S) - I(Published) = 0.9710 - 0.5510 = 0.42 \tag{30}$$

For the quality of the student's recommendations, the information gained is:
Recommendations = Good : 3 P, 2 N

$$E(Recommendations = Good) = -\frac{3}{5}\log_2\frac{3}{5}) - \frac{2}{5}\log_2\frac{2}{5} = .9710 \tag{31}$$

Recommendations = Normal : 0 P, 0 N

$$E(Recommendations = Normal) = -0\log_2 0 - 0\log_2 0 = 0 \tag{32}$$

$$I(Recommendations) = 1(.9710) + 0(0) = 0.9710 \tag{33}$$

$$Gain(Recommendations) = E(S) - I(Recommendations) = 0.9710 - 0.9710 = 0 \tag{34}$$

The best attribute to use next would be whether the student has publications. Because if the student has publications the answer is always positive, then we only need to examine when the student doesn't have publications to create the next child node in the decision tree.
For the subset, there is 1 positive and 2 negatives. Hence, the entropy of the subset is:

$$E(S) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183 \tag{35}$$

For university rank, the information gained is:
University rank = 1 : 0 P, 1 N

$$E(rank = 1) = -0\log_2 0 - 1\log_2 1 = 0 \tag{36}$$

University rank = 2 : 1 P, 0 N

$$E(rank = 2) = -1\log_2 1 - 0\log_2 0 = 0 \tag{37}$$

University rank = 3 : 0 P, 1 N

$$E(rank = 3) = -0\log_2 0 - 1\log_2 1 = 0 \tag{38}$$

$$I(rank) = \frac{1}{3} * 0 + \frac{1}{3} * 0 + \frac{1}{3} * 0 = 0 \tag{39}$$

$$Gain(rank) = E(S) - I(rank) = 0.9710 - 0 = 0.9710 \tag{40}$$

For the quality of the student's recommendations, the information gained is:
Recommendations = Good : 1 P, 2 N

$$E(Recommendations = Good) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183 \quad (41)$$

Recommendations = Normal : 0 P, 0 N

$$E(Recommendations = Normal) = -0\log_2 0 - 0\log_2 0 = 0 \quad (42)$$

$$I(Recommendations) = 1(.9183) + 0(0) = 0.9183 \quad (43)$$

$$Gain(Recommendations) = E(S) - I(Recommendations) = 0.9183 - 0.9183 = 0 \quad (44)$$

The best attribute to use next would be the university rank. Because there is no further information necessary to categorize positives and negatives, recommendation quality is not needed to determine the final outcome. Thus, the decision tree is complete.

## Part 3

The decision tree we obtained in part b is identical to the decision tree provided in part a. Because all of the data match the decision tree, this is not surprising.

# Question 3

## Part a

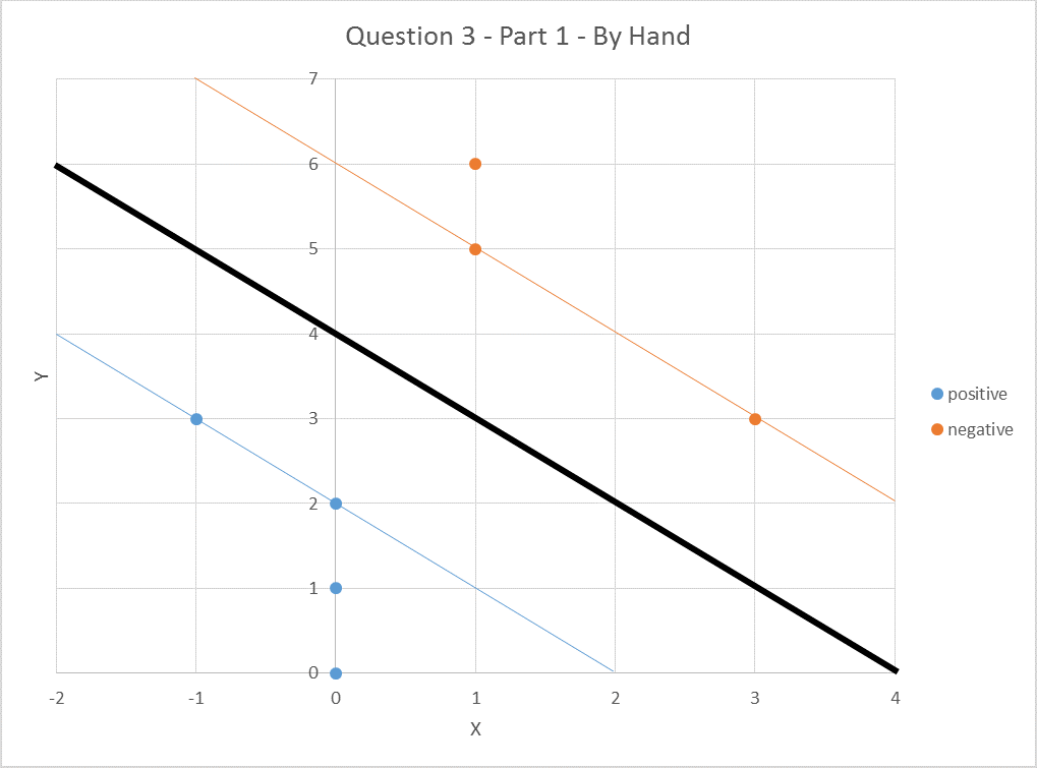



Figure 3: Positive data set is blue. Negative data set is orange. Hand drawn linear classifier in black. Support vectors for corresponding data sets also included

## Part b

The parameters w and b for this line are -1 and 4, respectively. The equation below represents the drawn classifer.

$$h(x) = -x + 4 \tag{45}$$

## Part c

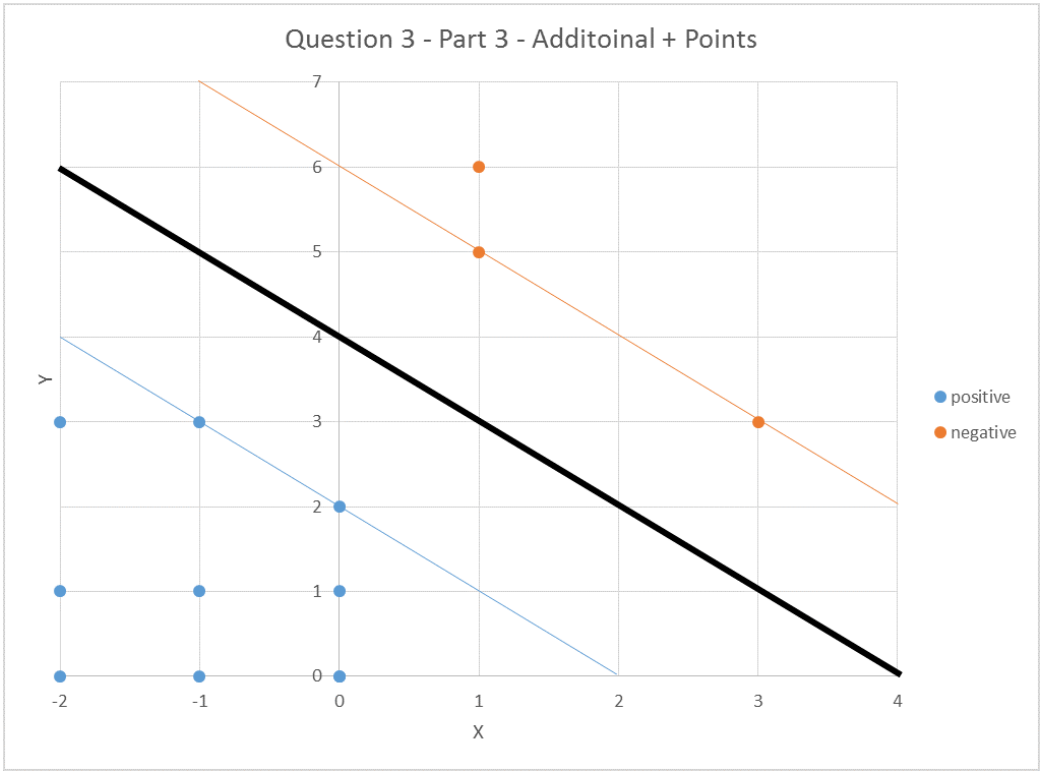



Figure 4: Positive data set is blue. Negative data set is orange. Hand drawn linear classifier in black. Support vectors for corresponding data sets also included

The addition of more data to the positive data set did not change the linear SVM because the values did not change the position of the corresponding support vectors. As a result, the values of w and b are unchanged and the equation below still represents the classifer in the graph.

$$h(x) = -x + 4 \tag{46}$$

# Question 4

## Part a

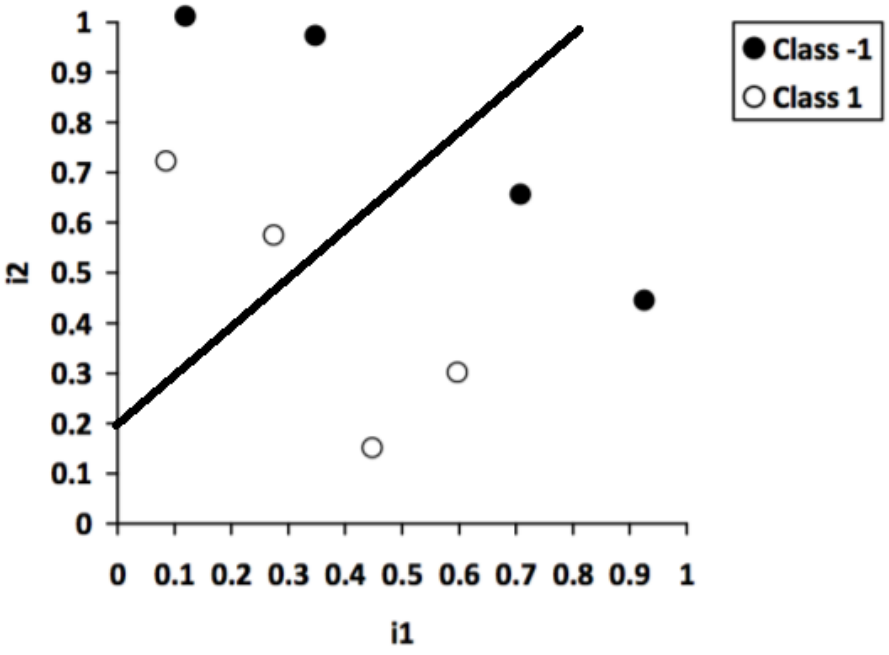



Figure 5: The plot with

The initial line misclassifies exactly 4 samples. Coordinates are approxmiate.

| Sample | Classification | Classified as | Calculation |
|---|---|---|---|
| (.10, .70) | Class 1 | Class -1 | $(1)(.1)+(-1)(.7)+(.2)(1) = -.4$ |
| (.30, .60) | Class 1 | Class -1 | $(1)(.3)+(-1)(.6)+(.2)(1) = -.1$ |
| (.95, .45) | Class -1 | Class 1 | $(1)(.95)+(-1)(.45)+(.2)(1) = .70$ |
| (.70, .65) | Class -1 | Class 1 | $(1)(.7)+(-1)(.65)+(.2)(1) = .25$ |

Selecting an arbitrary misclassified sample : (.3, .6) and running a weight update computation as follows:

(.3, .6) is Class 1 and misclassifed as -1.
$w_1 = 1 + .3 = 1.3$

$w_2 = -1 + .6 = -.4$

$$y = -\frac{1.3}{-.4} - \frac{.2}{-.4} \tag{47}$$

Producing a new linear separator represented by the following equation:

$$y = 3.25x_i + .5 \tag{48}$$

Graphed on the sample space:

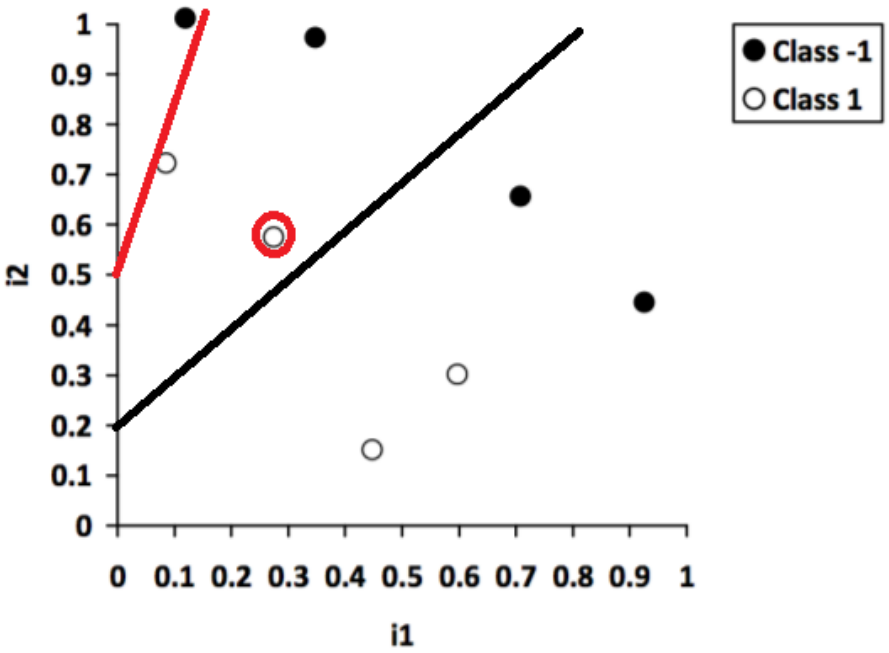



Figure 6: 1st Iteration of Weight Update. New linear classifier and point used to produce it in Red.

This new classifer misclassifes 3 samples and correctly classifies 5. Repeating this algorithm 4 more times:

Iteration 2
(.7, .65) is Class -1 and misclassifed as 1.
$w_1 = 1.3 - .7 = .6$
$w_2 = -.4 - .65 = -1.05$

$$y = -\frac{.60}{-1.05} - \frac{.2}{-1.05} \tag{49}$$

Producing a new linear separator represented by the following equation:

$$y = .57x_i + .19 \tag{50}$$

Graphed on the sample space:

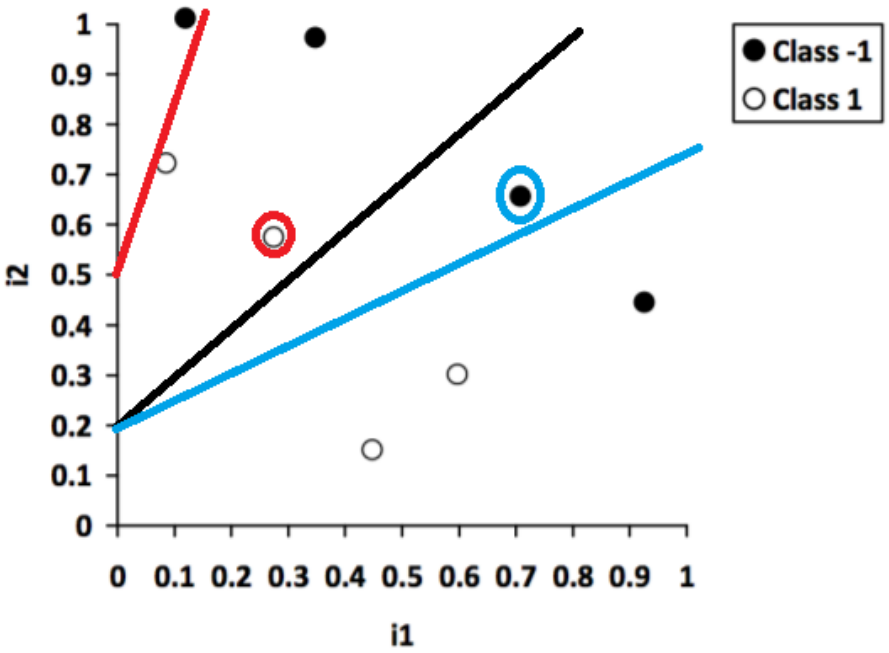



Figure 7: 2nd Iteration of Weight Update. New linear classifier and point used to produce it in Blue.

This new classifer misclassifes 3 samples and correctly classifies 5.
Iteration 3
(.1, .7) is Class 1 and misclassifed as -1.
$w_1 = .60 + .10 = .70$
$w_2 = -1.05 + .70 = -.35$

$$y = -\frac{.70}{-.35} - \frac{.2}{-35} \tag{51}$$

Producing a new linear separator represented by the following equation:

$$y = 2x_i + .7 \tag{52}$$

Graphed on the sample space:

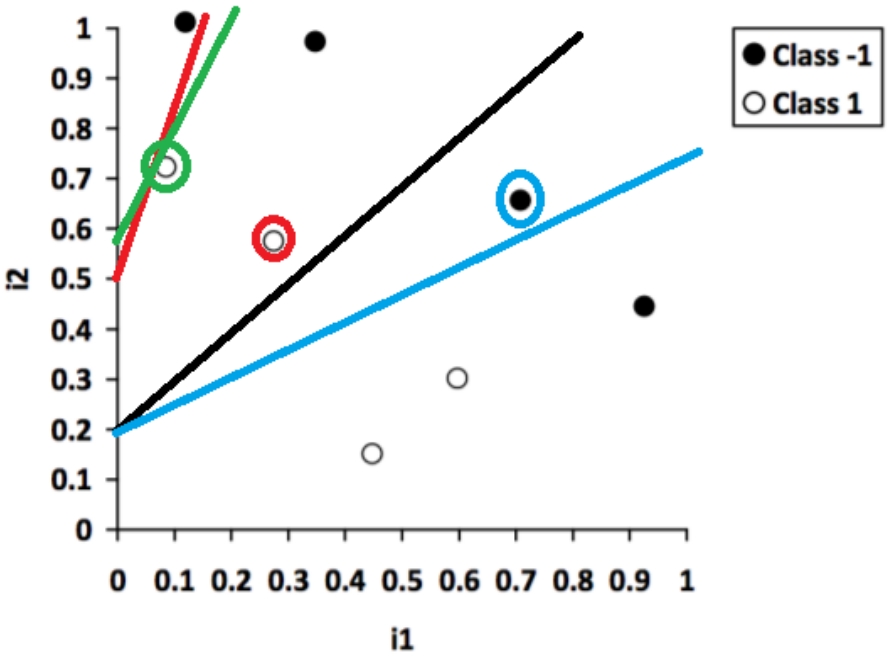



Figure 8: 3rd Iteration of Weight Update. New linear classifier and point used to produce it in Green.

This new classifer misclassifes 3 samples and correctly classifies 5.

Iteration 4
(.35, .95) is Class -1 and misclassifed as 1.
$w_1 = .70 - .35 = .35$
$w_2 = -.35 - .95 = -1.30$

$$y = -\frac{.35}{-1.30} - \frac{.2}{-1.3} \tag{53}$$

Producing a new linear separator represented by the following equation:

$$y = .27x_i + .15 \tag{54}$$

Graphed on the sample space:

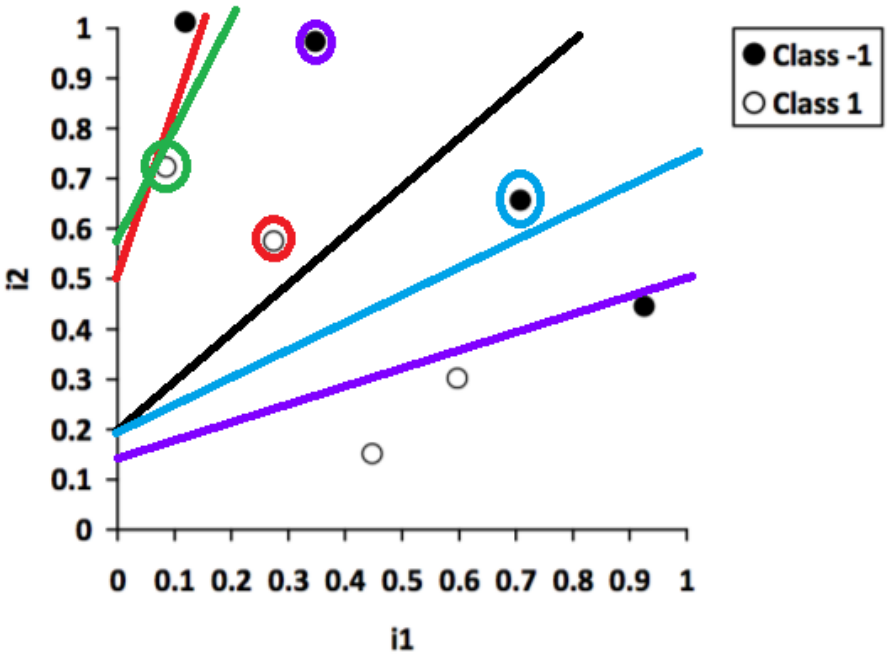



Figure 9: 4th Iteration of Weight Update. New linear classifier and point used to produce it in Purple.

This new classifer misclassifes 3 samples and correctly classifies 5.

Iteration 5
(.95, .45) is Class -1 and misclassifed as 1.
$w_1 = .35 + .95 = .60$
$w_2 = -1.3 - .45 = -1.75$

$$y = -\frac{.60}{-1.75} - \frac{.2}{-1.75} \tag{55}$$

Producing a new linear separator represented by the following equation:

$$y = .34x_i + .11 \tag{56}$$

Graphed on the sample space:

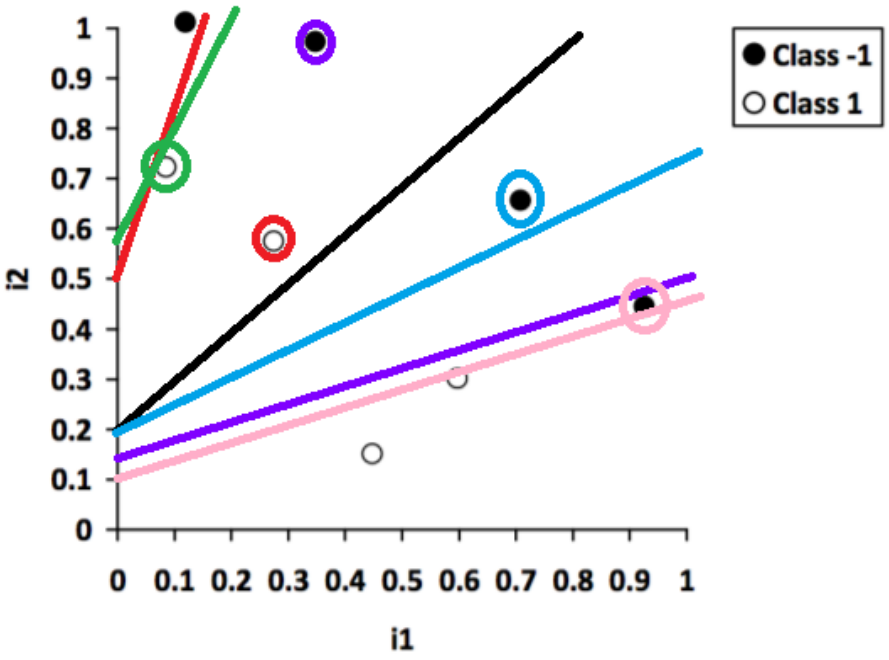



Figure 10: 5th Iteration of Weight Update. New linear classifier and point used to produce it in Pink.

This new classifer misclassifes 2 samples and correctly classifies 6.

We believe that the classifer we calculated in this part of the question failied to arrive at perfect classification because our rate of learning, or alpha value, was too high. We defaulted to an alpha of 1 for simplicity but we believe a smaller alpha is more likely to produce an optimal solution.

## Part b

The classifer in Part a did not reach perfect classification. We have determined that the arbitrary set of weights:
$w_1 = -.154$
$w_2 = -.22$
$w_0 = .2$
With the equation

$$y = -.7x_i + .9 \tag{57}$$

That will produce a perfect linear separator as graphed on the sample space:

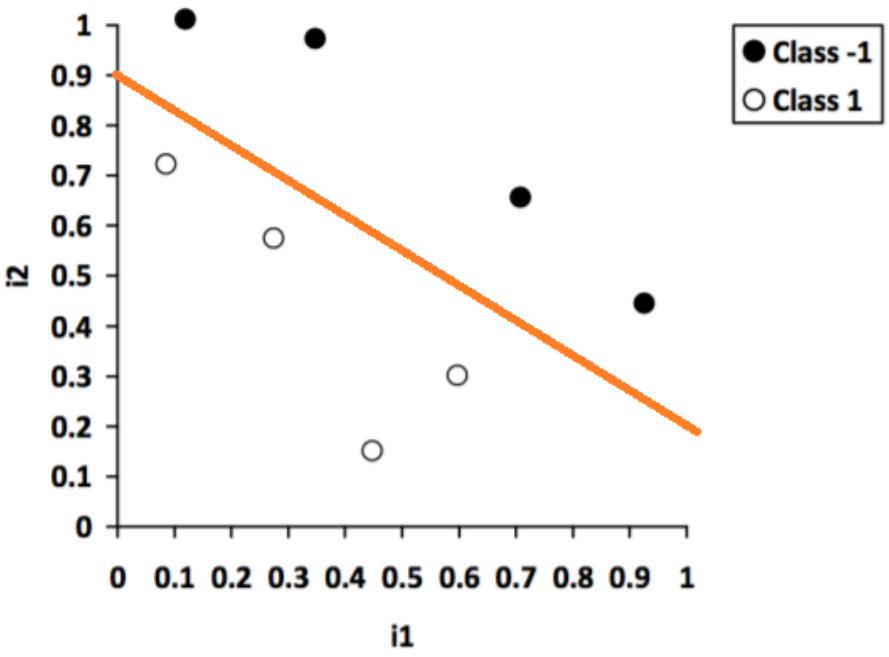



Figure 11: Perfect linear classifier for sample space graphed in Orange.

## Part c

Removing one dimesion of the sample space only permitted the separation of samples by their i1 values and true classification. The removal of i2 also

forces a straight vertical line classifier. We determined the optimal weight for this vertical line classifer to be defined by the equation and weights:
$w_0 = .65$ $w_1 = -1$

$$x = -\frac{w_o}{w_1} = -\frac{.65}{-1} = .65 \tag{58}$$

Graphed on the sample space:

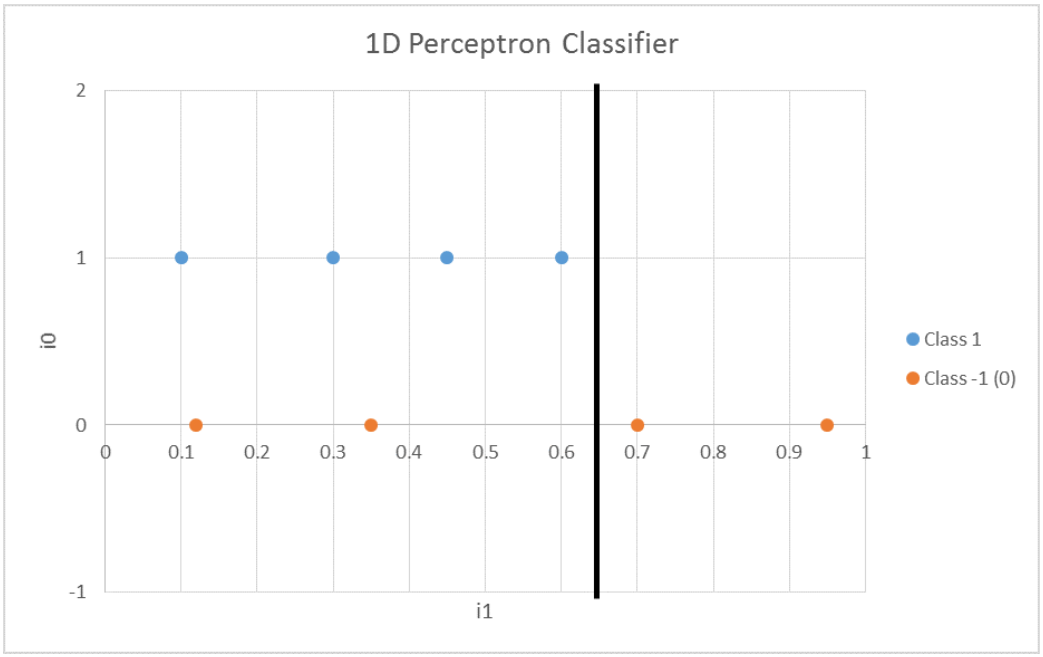



Figure 12: Perfect linear classifier for sample space graphed in black. Left of the classifier is Class 1. Right of the classifier is Class -1.

The classifier separates the sample space with a vertical line at $x = .65$ with an error of 2 misclassified. We believe this is the optimal solution and that other proposed perceptrons will have a higher error for this lower-dimension sample space.

# Question 5

## Part a

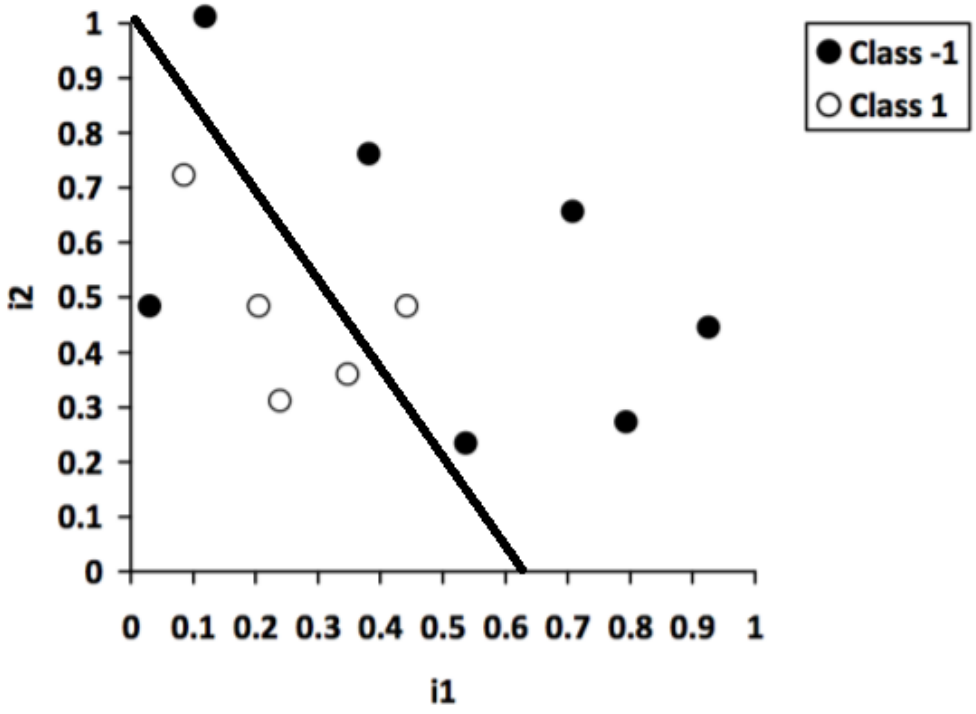



Figure 13: The minimum error possible is 2.

## Part b

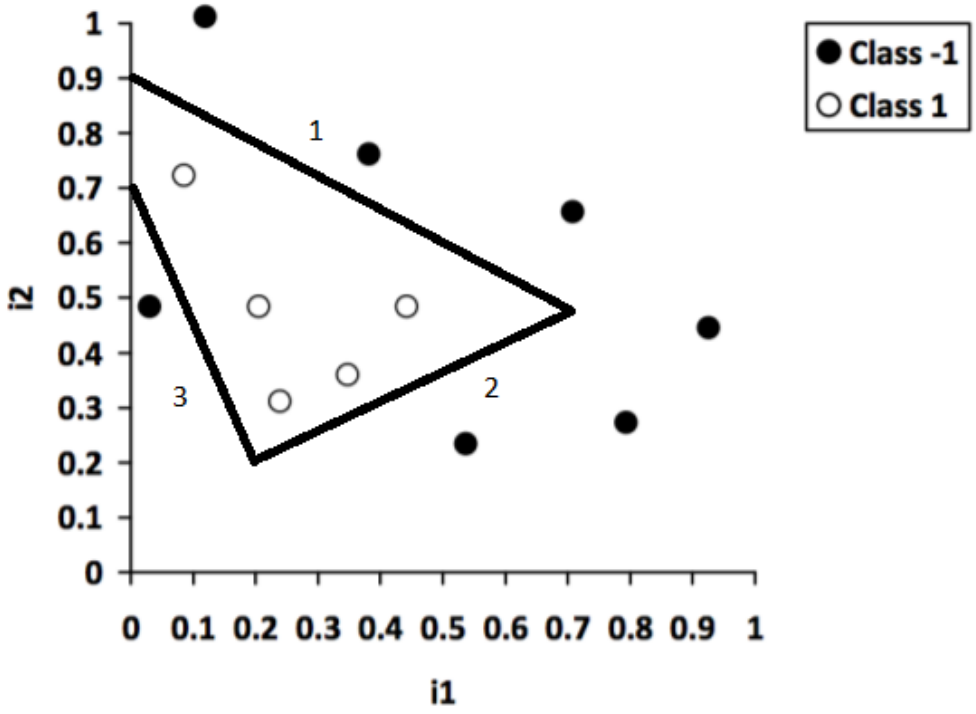



Figure 14: A multilayer perceptron with 3 layers separates the two categories.

The weights for the lines are:

$w_0 = -0.9, w_1 = \frac{4}{7}, w_2 = -1$ for line 1
$w_0 = -0.08, w_1 = -0.6, w_2 = 1$ for line 2
$w_0 = -0.7, w_1 = 2.5, w_2 = 1$ for line 3

Where $w_0$ is the weight for the bias, which we assume to be 1 in this case, $w_1$ is the weight for $i_1$, and $w_2$ is the weight for $i_2$.