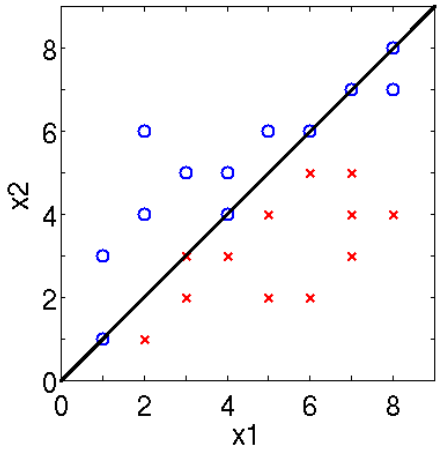
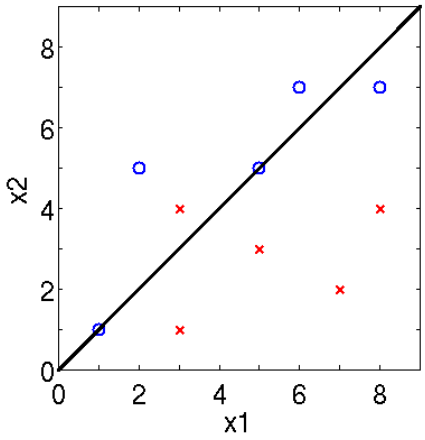
Machine Learning Assignment 2

1. ANSWER to Question 1

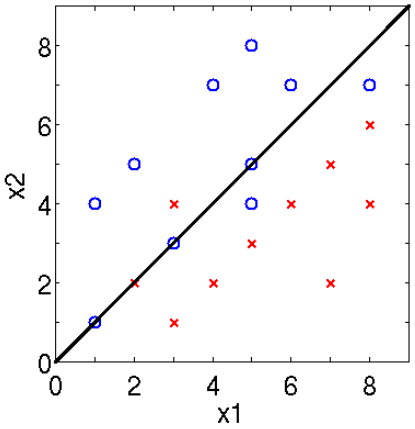
   a. Here are the graphs for S_{train}, S_{test1}, and S_{test2} (in that order) with the line $x_2=x_1$ added. All points on and above the line should be positive, according to this model.



S_{train}: As you can see here, two points, (3,3) and (8,7) out of the 24 points in the set are classified incorrectly according to our model. This gives us an accuracy of 22/24 or about 91.67%.



S_{test1}: As you can see here, two points, (3,4) and (8,7), out of the 10 points in this set are classified incorrectly according to our model. This gives us an accuracy of 8/10 or 80%.



S_{test2}: As you can see here, four points, (2,2), (3,4), (5,4) and (8,7), out of the 20 points in this set are classified incorrectly according to our model. This gives us an accuracy of 16/20 or 80%.

   b. Note: See Appendix A for tables for each of these.

For S_{train}, the decision tree returns the correct values for all of the elements in the set, giving us an accuracy of 100%.

For S_{test1}, the decision tree returns the correct values for all of the 10 elements in the set, except for the two points (3,4) and (5,5). This gives us an accuracy of 8/10 or 80%.

For S_{test2}, the decision tree returns the correct values for all of the 20 elements in the set, except for the five points (3,3), (3,4), (5,4), (5,5), and (8,6), giving us an accuracy of 15/20 or 75%.

c. McNemar test for S_{test1} (using values from the previous two parts) (table directly comparing the parts can be found in Appendix B):

	X_2>=X_1			
Decision Tree		**Predicts Correctly**	**Predicts Error**	**Total**
	Predicts Correctly	7	1	8
	Predicts Error	1	1	2
	Total	8	2	10

Plugging this into the formula for the McNemar test (from Wikipedia) we get:

$$\chi^2 = \frac{(b-c)^2}{b+c} = \frac{(1-1)^2}{1+1} = 0$$

This means we have no confidence that one test is any better than the other, which makes sense, since the two tests acted fairly identically, each getting the same 7 correct and 1 wrong, and each getting one of the other two wrong.

~~Using http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf, it looks like alpha approaches 1 for this value. So the probability that this is significant is 1-1 or 0, meaning we can't say that one model generalizes better than the other for this validation set with any confidence.~~

d. McNemar Test for S_{test2}

	$X_2>=X_1$			
Decision Tree		**Predicts Correctly**	**Predicts Error**	**Total**
	Predicts Correctly	13	2	15
	Predicts Error	3	2	5
	Total	16	4	20

Plugging this into the formula for the McNemar test (from Wikipedia) we get:

$$\chi^2 = \frac{(b-c)^2}{b+c} = \frac{(2-3)^2}{2+3} = \frac{1}{5}$$

Using http://www.danielsoper.com/statcalc3/calc.aspx?id=11, we get a probability of 0.65472085. I think this means that the probability that one test is better than the other is about 65%.

~~Using the same table from earlier, if I'm reading it right, we get alpha = 0.9. So the probability that we can say with confidence that one model generalizes better than the other is 1-0.9 = 0.1.~~

e. We can display our answers to the previous parts in the following way:

	S_{test1}	**S_{test2}**
Decision Tree Model	80%	75%
Linear Model	80%	80%
McNemar	0%	65%

So, the accuracy of the decision tree model decreased from S_{test1} to S_{test2}, and the McNemar value increased. This makes sense, since for S_{test2}, there is more of a difference between the accuracies, making one of the models (the linear one) preferable to the other one (the decision tree one).

Appendix A

Values Returned by Decision Tree for S_{train}		
(x_1, x_2)	**DT Prediction**	**Correct?**
(1,1)	**+**	Yes
(1,3)	**+**	Yes
(2,1)	**-**	Yes
(2,4)	**+**	Yes
(2,6)	**+**	Yes
(3,2)	**-**	Yes
(3,3)	**-**	Yes
(3,5)	**+**	Yes
(4,3)	**-**	Yes
(4,4)	**+**	Yes
(4,5)	**+**	Yes
(5,2)	**-**	Yes
(5,4)	**-**	Yes
(5,6)	**+**	Yes
(6,2)	**+**	Yes
(6,5)	**-**	Yes
(6,6)	**+**	Yes
(7,3)	**-**	Yes
(7,4)	**-**	Yes
(7,5)	**-**	Yes
(7,7)	**+**	Yes

(8,4)	-	Yes
(8,7)	+	Yes
(8,8)	+	Yes

Values Returned by Decision Tree for S_{test1}		
(x_1, x_2)	**Decision Tree Prediction**	**Correct**
(1,1)	+	Yes
(2,5)	+	Yes
(3,1)	-	Yes
(3,4)	+	No
(5,3)	-	Yes
(5,5)	-	No
(6,7)	+	Yes
(7,2)	-	Yes
(8,4)	-	Yes
(8,7)	+	Yes

Values Returned by Decision Tree for S_{test2}		
(x_1, x_2)	**DT Prediction**	**Correct**
(1,1)	+	Yes
(1,4)	+	Yes
(2,2)	-	Yes
(2,5)	-	Yes
(3,1)	-	Yes

(3,3)	-	No
(3,4)	+	No
(4,2)	-	Yes
(4,7)	+	Yes
(5,3)	-	Yes
(5,4)	-	No
(5,5)	-	No
(5,8)	+	Yes
(6,4)	-	Yes
(6,7)	+	Yes
(7,2)	-	Yes
(7,5)	-	Yes
(8,4)	-	Yes
(8,6)	+	No
(8,7)	+	Yes

Appendix B: Tables for 1c and 1d

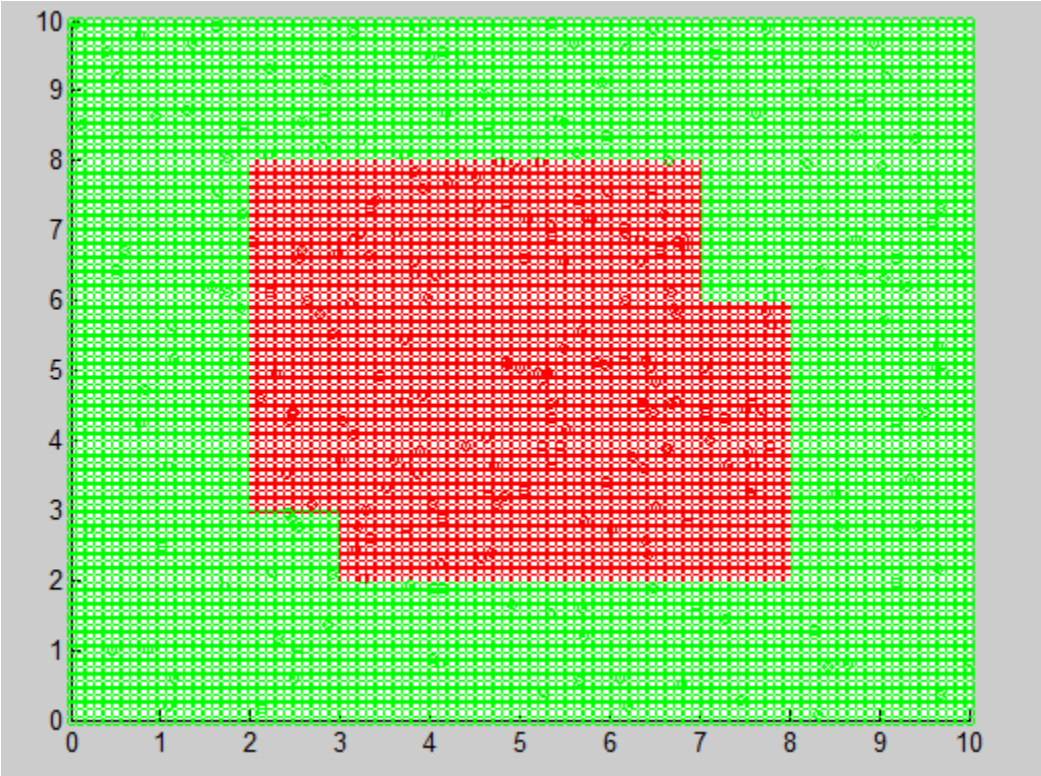
Decision Tree vs $x_2>=x_1$ for S_{test1}		
(x_1, x_2)	**DT Correct**	**$x_2>=x_1$ Correct**
(1,1)	Yes	Yes
(2,5)	Yes	Yes
(3,1)	Yes	Yes
(3,4)	No	No
(5,3)	Yes	Yes
(5,5)	No	Yes
(6,7)	Yes	Yes
(7,2)	Yes	Yes
(8,4)	Yes	Yes
(8,7)	Yes	No

Decision Tree vs $x_2>=x_1$ for S_{test2}		
(x_1, x_2)	**DT Correct**	**$x_2>=x_1$ Correct**
(1,1)	Yes	Yes
(1,4)	Yes	Yes
(2,2)	Yes	No
(2,5)	Yes	Yes
(3,1)	Yes	Yes
(3,3)	No	Yes
(3,4)	No	No
(4,2)	Yes	Yes

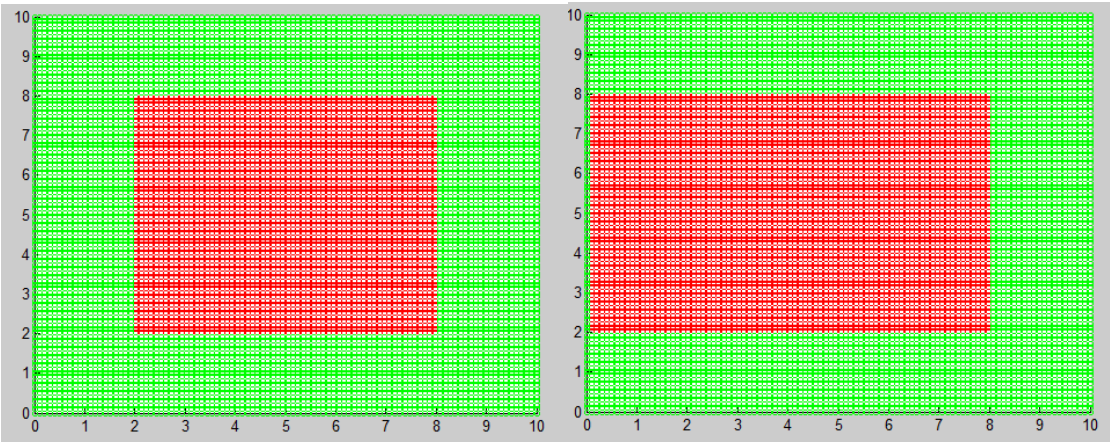
(4,7)	Yes	Yes
(5,3)	Yes	Yes
(5,4)	No	No
(5,5)	No	Yes
(5,8)	Yes	Yes
(6,4)	Yes	Yes
(6,7)	Yes	Yes
(7,2)	Yes	Yes
(7,5)	Yes	Yes
(8,4)	Yes	Yes
(8,6)	No	Yes
(8,7)	Yes	No

Answer to Question 2

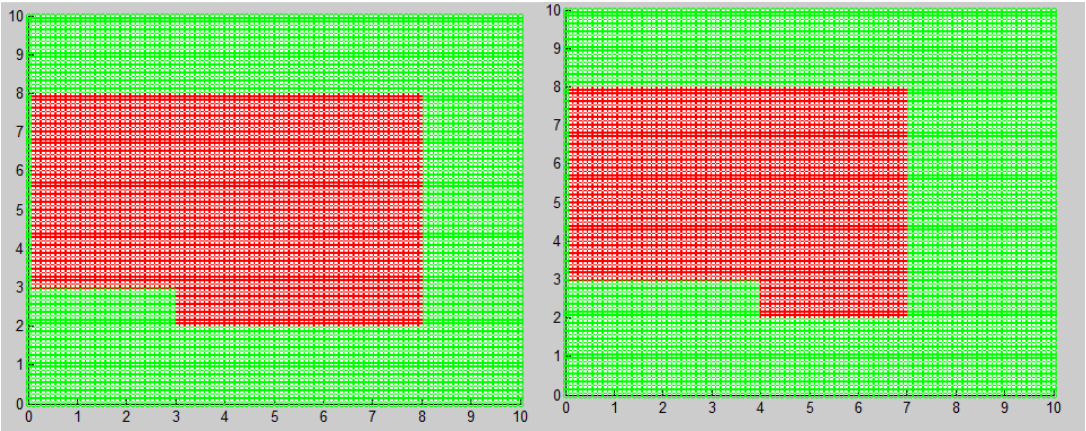
a) The plot of decision tree as tested on the grip points randomly sampled from the square-



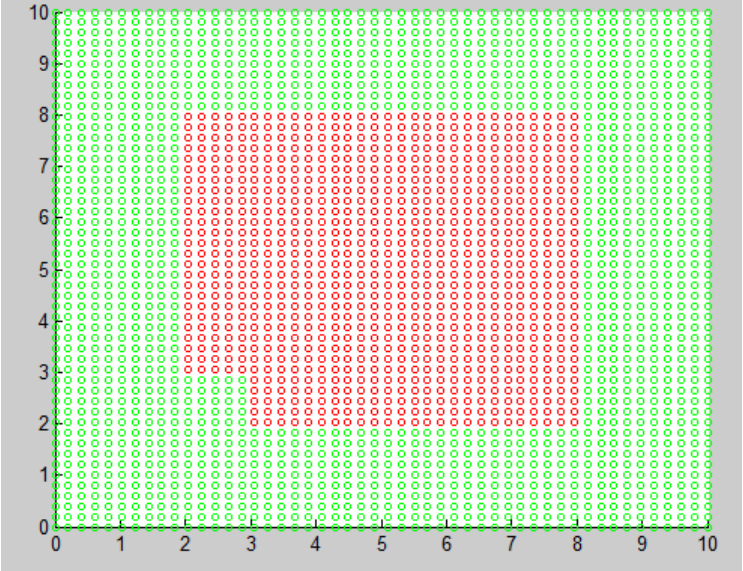
b) Boundary of 2 such classifiers are-



---Two cases of over fitting trees:



---Boundary of the combined classifier on the test data randomly sampled points-



The boundary of the combined classifier is ensemble of the individual classifiers. Individual classifiers vary diversely in boundaries which depend on what part of the test data (1/5th of the overall data) was sampled for training the individual classifier.

The boundary of the combined classifier is closer to the true boundary that was used to generate the data i.e. the circle. It can be attributed to the fact that train data can be sometimes spurious or labels are given wrong. So it makes more sense to part of the data and then aggregate the combined output of all the models to get mean weight. Wisdom of crowd has been proved to be more accurate in the past studies. In this case as well, we find that taking aggregate of multiple predictor labels gives use boundary that is more closer to true function (circle).