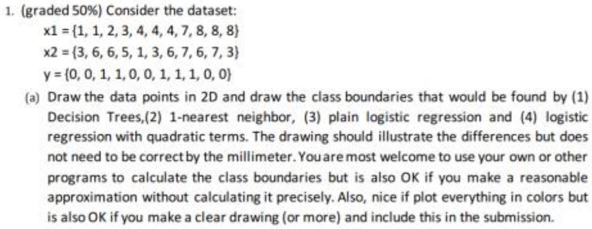
Machine Learning Assignment 4 - Julian Tsang

Thursday, November 17, 2016 2:10 PM



1. (graded 50%) Consider the dataset:

   x1 = {1, 1, 2, 3, 4, 4, 4, 7, 8, 8, 8}

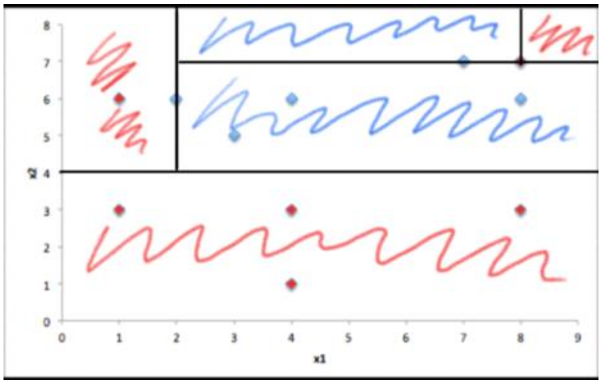
   x2 = {3, 6, 6, 5, 1, 3, 6, 7, 6, 7, 3}

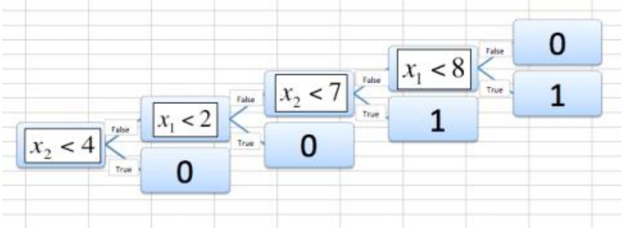
   y = {0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0}

   (a) Draw the data points in 2D and draw the class boundaries that would be found by (1) Decision Trees,(2) 1-nearest neighbor, (3) plain logistic regression and (4) logistic regression with quadratic terms. The drawing should illustrate the differences but does not need to be correct by the millimeter. You are most welcome to use your own or other programs to calculate the class boundaries but is also OK if you make a reasonable approximation without calculating it precisely. Also, nice if plot everything in colors but is also OK if you make a clear drawing (or more) and include this in the submission.

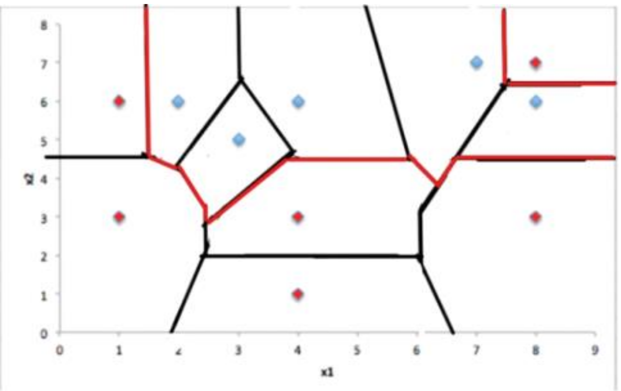
Data points in red are where y=0. Data points in blue are where y=1.

Decision Tree

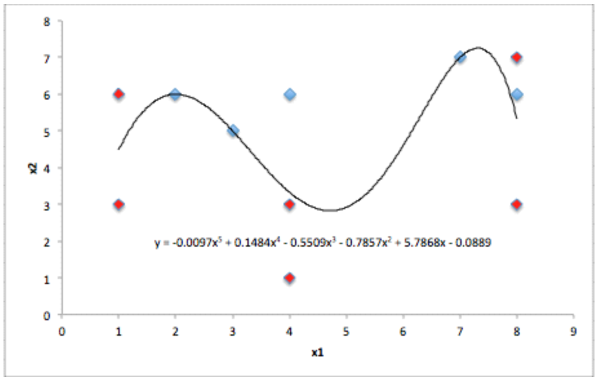




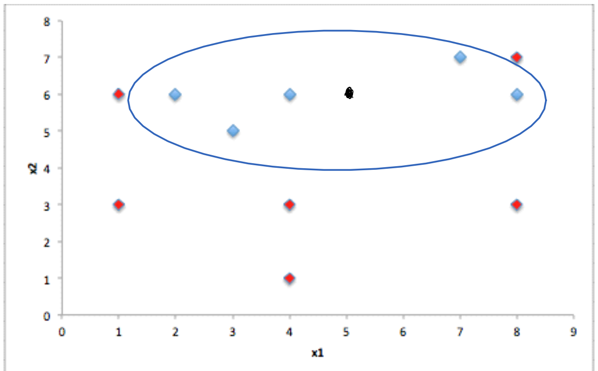
Nearest Neighbor



Logistic Regression

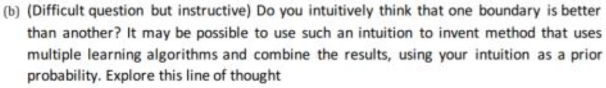


Logistic Regression with Quadratic Terms



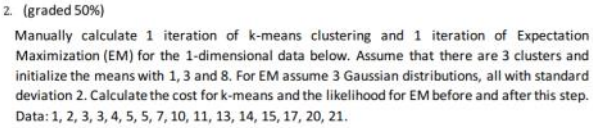
Approximation of Decision Boundary:

$$\frac{(x_1-5)^2}{3.5^2}+\frac{(x_2-6)^2}{2^2}=1$$



(b) (Difficult question but instructive) Do you intuitively think that one boundary is better than another? It may be possible to use such an intuition to invent method that uses multiple learning algorithms and combine the results, using your intuition as a prior probability. Explore this line of thought

The data points in this problem are not linearly separable so it is important to choose a good decision boundary. The use of decision trees and nearest neighbor are not optimal because of the presence of outliers in the data. If interpreted as a polynomial to the second order, logistic regression with quadratic terms has an underfitting problem since the boundary would not fit the overall data. However, if the boundary is interpreted as an ellipse, then logistic regression may give an appropriate boundary. In general, as illustrated in the third part of the question above, logistic regression may have an overfitting problem, but that is to be determined if there is more data to be added to the dataset. If we take into consideration only the data as we have it now, then logistic regression gives a reasonable decision boundary as well.



2. (graded 50%)

Manually calculate 1 iteration of k-means clustering and 1 iteration of Expectation Maximization (EM) for the 1-dimensional data below. Assume that there are 3 clusters and initialize the means with 1, 3 and 8. For EM assume 3 Gaussian distributions, all with standard deviation 2. Calculate the cost for k-means and the likelihood for EM before and after this step.
Data: 1, 2, 3, 3, 4, 5, 5, 7, 10, 11, 13, 14, 15, 17, 20, 21.

Index	$c^{(1)}$	$c^{(2)}$	$c^{(3)}$	$c^{(4)}$	$c^{(5)}$	$c^{(6)}$	$c^{(7)}$	$c^{(8)}$	$c^{(9)}$	$c^{(10)}$	$c^{(11)}$	$c^{(12)}$	$c^{(13)}$	$c^{(14)}$	$c^{(15)}$	$c^{(16)}$
Data	1	2	3	3	4	5	5	7	10	11	13	14	15	17	20	21
Initial Centroid	1	1	3	3	3	3	3	8	8	8	8	8	8	8	8	8
Mean	1.5	1.5	4	4	4	4	4	14.22	14.22	14.22	14.22	14.22	14.22	14.22	14.22	14.22
New Centroid	1.5	1.5	4	4	4	4	4	4	14.22	14.22	14.22	14.22	14.22	14.22	14.22	14.22

K-Means Clustering

Random Initialization

Initialize centroids at 1,3,8.

Cluster Assignment Step

Assign the data points to the centroid at their respective closest distance.

$c^{(i)} \coloneqq min\left\|x^{(i)} - \mu_k\right\|^2$

Labeled as initial centroid

Move Centroid Step

$For\ k = 1\ to\ K, \mu_k = mean\ of\ all\ points\ assigned\ to\ cluster\ k$

The centroid is updated to take on the mean value of the points assigned to it.

Labeled as new centroid

Cost for K-Means

$$J\left(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

Before implementing cluster assignment and move centroid steps:

$$\frac{1}{16}[(1-1)^2 + (2-1)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 + (5-3)^2 + (7-8)^2 + (10-8)^2 + (11-8)^2 + (13-8)^2 + (14-8)^2 + (15-8)^2 + (17-8)^2 + (20-8)^2 + (21-8)^2] = 33$$

After implementing cluster assignment and move centroid steps:

$$\frac{1}{16}[(1-1.5)^2 + (2-1.5)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (5-4)^2 + (7-4)^2 + (10-14.22)^2 + (11-14.22)^2 + (13-14.22)^2 + (14-14.22)^2 + (15-14.22)^2 + (17-14.22)^2 + (20-14.22)^2 + (21-14.22)^2] = 8.18295$$