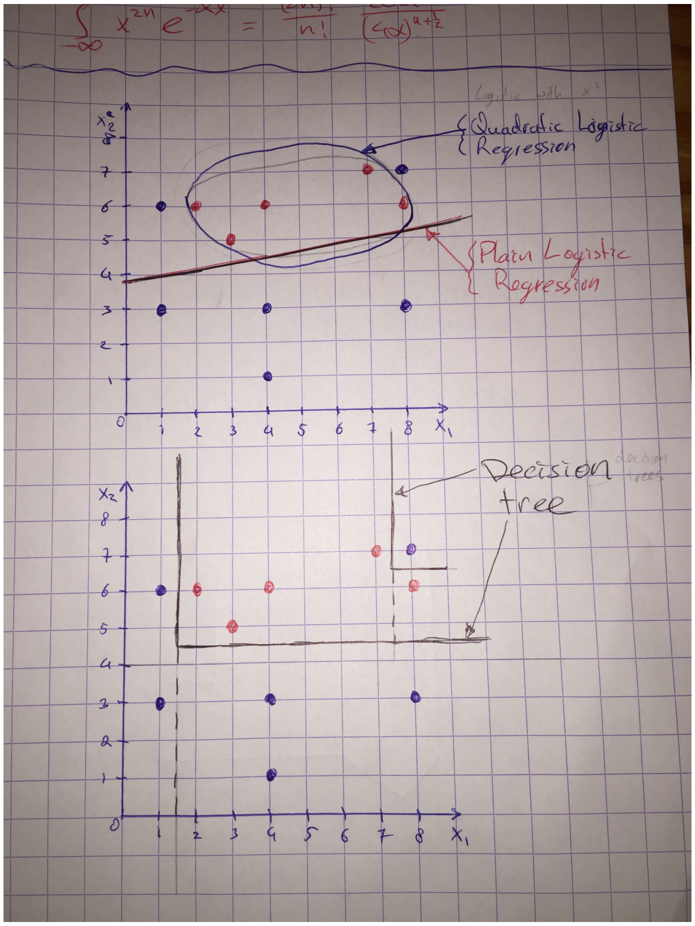
Machine Learning - Written Assignment 4

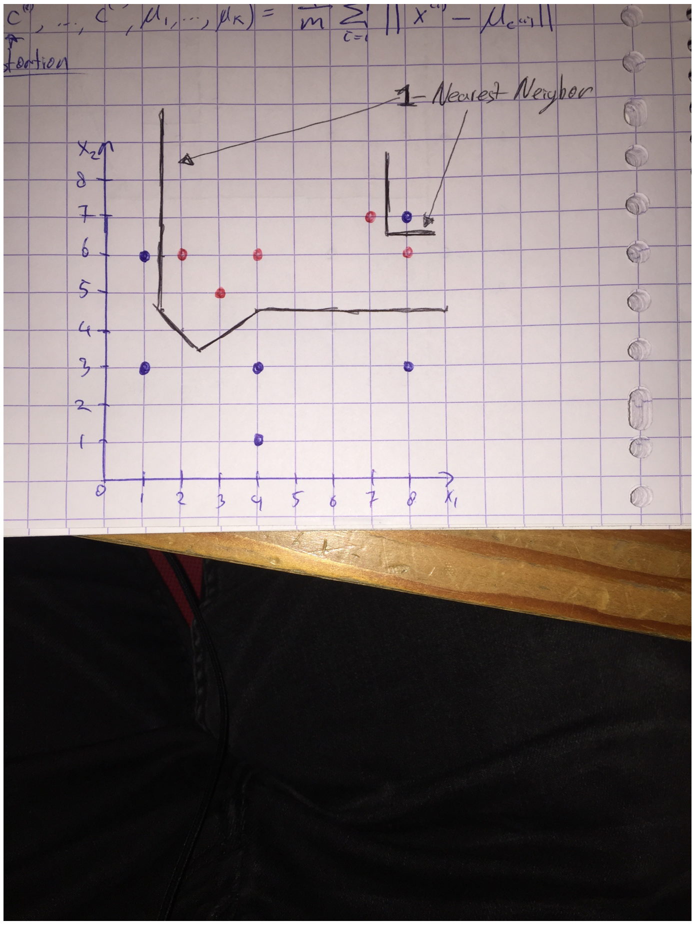
khaled tamimy

November 18, 2016

1 Question 1:

1.1 a:





1.2 b:

Intuitively I would think that the best boundary is that of 1-Nearest Neighbor. Although this algorithm seems very counter intuititive to me; one would not classify something regarding to its nearest neighbor, the boundary seems the best fitting in this case.

One could run all these algorithms and use a best-fit algorithm; some algorithm that looks for a combination of one or multiple algorithms that would best fit the given data set.

Alternatively, to come back to the question: One could use the 1-NN algorithm together with the quadratic logistic regression. The 1-NN algorithm has very edgy boundaries; a combination with quadratic logistic regression could make these boundaries more smooth.

2 Question 2:

We have initialized the means as follows: $\mu_1 = 1$, $\mu_2 = 3$ and $\mu_3 = 8$ And, we have the given dataset: {1,2,3,3,4,5,5,7,10,11,13,14,15,17,20,21}.

We have:

$$c^{(i)} := argmin_j ||x^{(i)} - \mu_j||^2 \tag{1}$$

And, $\mu_{c^{(i)}}$ = value of the cluster centroid to which $x^{(i)}$ has been assigned

Hence, we have:
c$^{(1)} = 1$
$c^{(2)} = 1$
$c^{(3)} = 2$
$c^{(4)} = 2$
$c^{(5)} = 2$
$c^{(6)} = 2$
$c^{(7)} = 2$
$c^{(8)} = 3$
$c^{(9)} = 3$
$c^{(10)} = 3$
$c^{(11)} = 3$
$c^{(12)} = 3$
$c^{(13)} = 3$
$c^{(14)} = 3$
$c^{(15)} = 3$
$c^{(16)} = 3$

And for the value of the respective cluster centroids, we have:
$\mu_{c^{(1)}} = 1$
$\mu_{c^{(2)}} = 1$
$\mu_{c^{(3)}} = 3$
$\mu_{c^{(4)}} = 3$
$\mu_{c^{(5)}} = 3$
$\mu_{c^{(6)}} = 3$
$\mu_{c^{(7)}} = 3$
$\mu_{c^{(8)}} = 8$
$\mu_{c^{(9)}} = 8$
$\mu_{c^{(10)}} = 8$
$\mu_{c^{(11)}} = 8$
$\mu_{c^{(12)}} = 8$
$\mu_{c^{(13)}} = 8$
$\mu_{c^{(14)}} = 8$
$\mu_{c^{(15)}} = 8$
$\mu_{c^{(16)}} = 8$

And, we have:

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

Hence, the distortion (cost), then is:

$$J = \frac{1}{16}(0+1+0+0+1+4+4+1+4+9+25+36+49+81+144+169) \tag{2}$$

Thus:

$$J = 33 \tag{3}$$

Now, we update, or move, the cluster centroid (values):
We have:

$$\mu_k = \frac{1}{n}[x^{(k_1)} + ... + x^{(k_m)}] \tag{4}$$

Thus:

$$\mu_1 = \frac{1}{2}[1+2] = 1.5 \tag{5}$$

$$\mu_2 = \frac{1}{5}[3+3+4+5+5] = 4 \tag{6}$$

$$\mu_3 = \frac{1}{9}[7+10+11+13+14+15+17+20+21] = 14.2 \tag{7}$$

Then, we have:
c$^{(1)} = 1$
$c^{(2)} = 1$
$c^{(3)} = 2$
$c^{(4)} = 2$
$c^{(5)} = 2$
$c^{(6)} = 2$
$c^{(7)} = 2$
$c^{(8)} = 2$
$c^{(9)} = 3$
$c^{(10)} = 3$
$c^{(11)} = 3$
$c^{(12)} = 3$
$c^{(13)} = 3$
$c^{(14)} = 3$
$c^{(15)} = 3$
$c^{(16)} = 3$

And for the value of the respective cluster centroids, we have:
$\mu_{c^{(1)}} = 1.5$
$\mu_{c^{(2)}} = 1.5$
$\mu_{c^{(3)}} = 4$
$\mu_{c^{(4)}} = 4$

$\mu_{c^{(5)}} = 4$
$\mu_{c^{(6)}} = 4$
$\mu_{c^{(7)}} = 4$
$\mu_{c^{(8)}} = 4$
$\mu_{c^{(9)}} = 14.2$
$\mu_{c^{(10)}} = 14.2$
$\mu_{c^{(11)}} = 14.2$
$\mu_{c^{(12)}} = 14.2$
$\mu_{c^{(13)}} = 14.2$
$\mu_{c^{(14)}} = 14.2$
$\mu_{c^{(15)}} = 14.2$
$\mu_{c^{(16)}} = 14.2$

And thus, the corresponding distortion (cost) is:

$$J = \frac{1}{16}[0.25+0.25+1+1+0+1+1+9+(4.2)^2+(3.2)^2+(1.2)^2+0.04+0.64+(2.8)^2+(5.8)^2+(6.8)^2] \tag{8}$$

$$J = 8.2 \tag{9}$$