Machine Learning – Written Assignment 4

Philip Hartout

November 23, 2016

1. (a) The decision boundaries can be found below.
   (b) It appears that only decision trees, 1-nearest neighbour and logistic regression with quadratic terms classify all data points correctly all data points, but it also appears that they are prone to overfitting[1]; therefore, some form of trade-off is required to make optimal decisions. An idea to highly increase accuracy is to run all algorithms on the same dataset, and, with a given data point belonging to a test set, counting to which class it belongs to according to all 4 algorithms. If, say, all 4 algorithms indicate that the point should be a positive example, then it is highly likely that it is, on the other hand, being aware of the fact that some classifiers do not categorise a given point of a test set as a positive example should require a more careful consideration.

2. First, to do one iteration of the k-means clustering algorithm, we assign each value to the closest cluster centroid, which have means $\mu_{c^{(1)}} = 1$, $\mu_{c^{(2)}} = 3$, $\mu_{c^{(3)}} = 8$, by calculating the squared distance of each point to each cluster centroid. Hence, we end up with the following points assigned to the following cluster centroids:

$$
\begin{aligned}
c^1 = \text{argmin}_j \left\| x^{(1)} - \mu_j \right\|^2 &= 1 \\
c^2 = \ldots = c^7 &= 2 \\
c^8 = \ldots = c^{16} &= 3
\end{aligned}
$$

As a note, the second data point could have taken the value of the first cluster. From there, we can estimate the cost before updating the coordinates the cluster centroids. The formula for the cost function is given as follows.

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

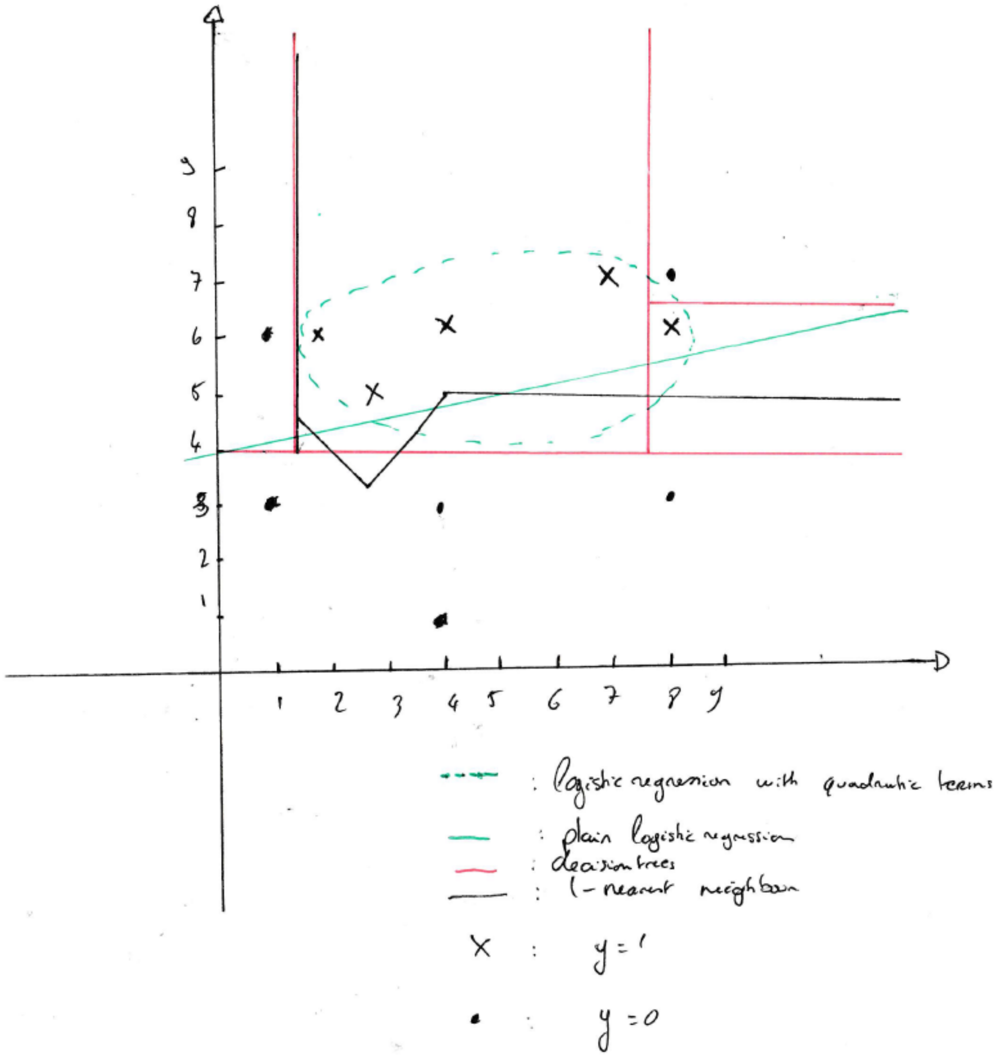
Here, in particular:

$$J(c^{(1)}, \ldots, c^{(16)}, \mu_1, \mu_2, \mu_3) = \frac{1}{16} [\ (1-1)^2 + (2-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 + (5-3)^2$$
$$+ (7-8)^2 + (10-8)^2 + (11-8)^2 + (13-8)^2 + (14-8)^2 + (15-8)^2 + (17-8)^2 + (20-8)^2 + (21-8)^2\ ] = 33$$

Now that we have (1) assigned each data point to a cluster and (2) calculated the cost function, we can now calculate the new position of each cluster centroid, which we achieve by calculating the mean of each cluster, as follows:

$$
\begin{aligned}
\mu_1 &= \frac{1}{1} \cdot 1 = 1 \\
\mu_2 &= \frac{1}{6} \cdot (2+3+3+4+5+5) = \frac{11}{3} = 3.667 \\
\mu_3 &= \frac{1}{9} \cdot (1+2+3+3+4+5+5+7+10+11+13+14+15+17+20+21) = \frac{151}{9} \approx 16.778.
\end{aligned}
$$

[1]Except, perhaps, logistic regression with quadratic terms.



9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
: logistic regression with quadratic terms
: plain logistic regression
: decision trees
: 1-nearest neighbour
X : y = 1
• : y = 0

Which leads us to re-assign the data points to new clusters:

$$
\begin{aligned}
c^1 = c^2 &= 1 \\
c^3 = \ldots = c^9 &= 2 \\
c^{10} = \ldots = c^{16} &= 3
\end{aligned}
$$

Finally, we estimate the cost function:

$$
J(c^{(1)}, \ldots, c^{(16)}, \mu_1, \mu_2, \mu_3) = \frac{1}{16} [\ (1-1)^2 + (2-1)^2 + \left(3 - \frac{11}{3}\right)^2 + \left(3 - \frac{11}{3}\right)^2 + \left(4 - \frac{11}{3}\right)^2 + \left(5 - \frac{11}{3}\right)^2
$$
$$
+ \left(5 - \frac{11}{3}\right)^2 + \left(7 - \frac{11}{3}\right)^2 + \left(10 - \frac{11}{3}\right)^2 + \left(11 - \frac{151}{9}\right)^2 + \left(13 - \frac{151}{9}\right)^2 + \left(14 - \frac{151}{9}\right)^2 + \left(15 - \frac{151}{9}\right)^2
$$
$$
+ \left(17 - \frac{151}{9}\right)^2 + \left(20 - \frac{151}{9}\right)^2 \left(21 - \frac{151}{9}\right)^2 \ ] = \frac{12493}{1296} \approx 9.639
$$