CSCI-737 Pattern Recognition Assignment 2

Prasanth Ganesan
Doctoral Student, Biomedical Engineering,
Rochester Institute of Technology

February 25, 2016

Contents

1 Question 1 - MAP Classifier

Figure 1 illustrates the classification of the two-class data using maximum a posteriori classifier. The yellow and green regions belong to Class 0 and 1 respectively. Similarly, the yellow and green circles are the corresponding data points. The black lines are the region boundaries.

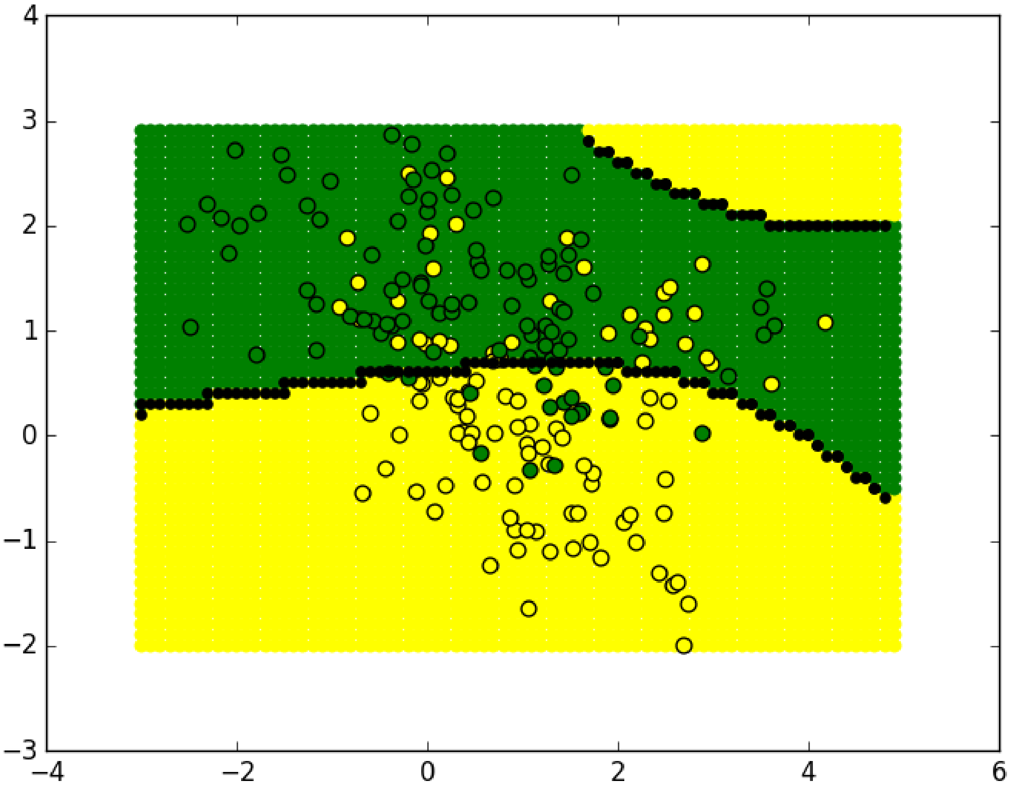

Figure 1: Classification using Maximum A Posteriori – The yellow and green circles represents the input points (yellow = Class 0, green = Class 1), the same colors represent the respective regions and the black line in between the regions is the classification boundary obtained using this method

It can be clearly seen that there are two boundaries splitting the regions into three regions of two classes. The primary reason for this phenomenon is the assumption of a gaussian distribution for deriving the class conditionals. The boundaries are generally those points where the distributions of the two classes intersect each other. Hence, if we imagine the two gaussians intersecting at two points (one distribution on top of the other), the result will be alternating class 0 and 1. This hence explains the result we see in Figure 1.

Figure 2, Figure 3 and Figure 4 illustrates the result of least squares linear classification, k-nearest neighbor for $K = 1$ and $K = 15$ respectively. These classification results are for the same data as the one used in MAP classification above. The least squares method assumes a linear model for the region splitting and the knn works based on the

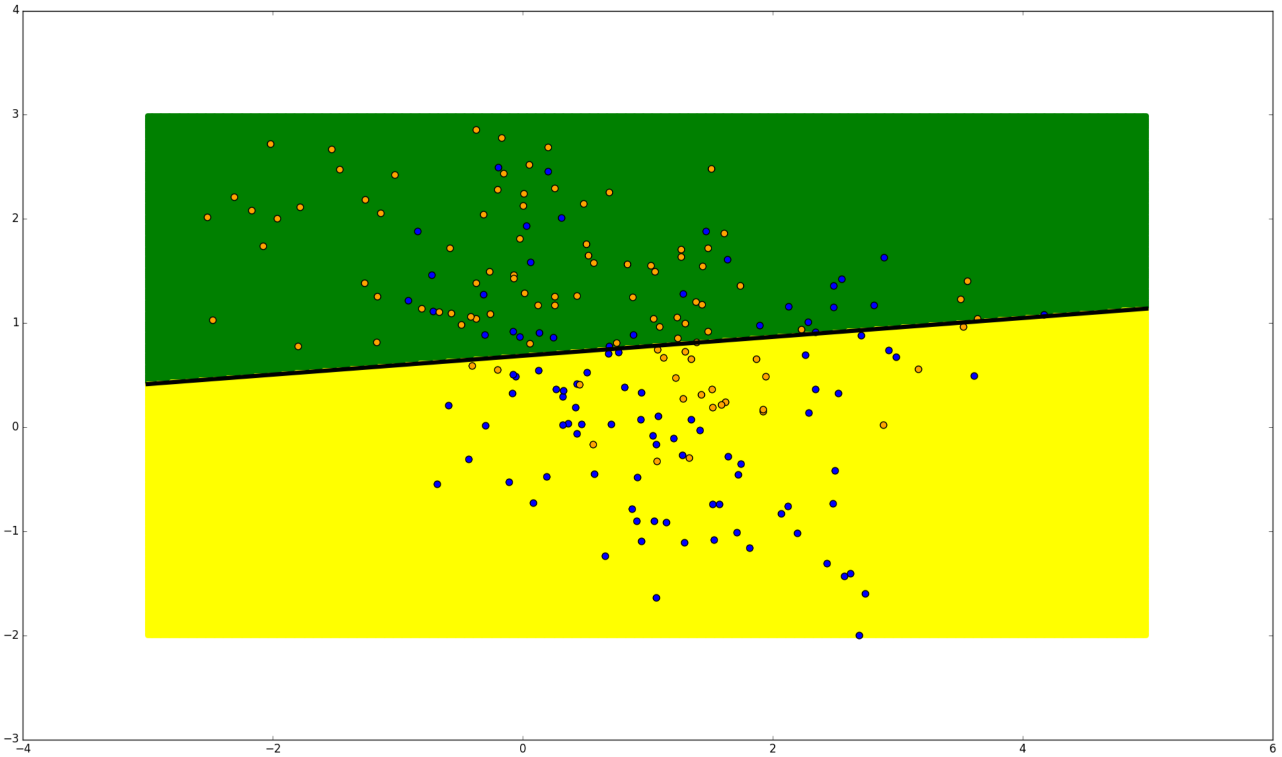

Figure 2: Classification using Least Squares – The orange and blue circles represents the input points (Blue = Class 0, Orange = Class 1), The green and yellow represents the regions (with same respective class labels) and the black line in between the regions represent the classification line obtained using this method

density of the classes present within K datapoints; on the other hand, the MAP classifier designed in our problem assumes the data distribution to be gaussian. If the data points used in training are taken from a gaussian distribution, then the resulting MAP classifier is highly accurate. Even otherwise, the MAP classifier is considered "optimal" due to the assumption that data distribution and the class conditionals are same. Therefore, MAP classifier works better than the other two classifiers, provided the data really follows a gaussian distribution.

Figure 5 consists of the table showing the confusion matrix of MAP classifier. The classification rate was 71.5%.

2 Question 2 - Bayesian Classifier

Bayesian classifier and MAP classifier are very similar to each other. If we include a cost matrix in our classification operation, in order to control the splitting of regions customized to our order of importance, then this is given the name bayesian classifier. This cost matrix used in bayesian classifier could be uniform or non-uniform. In all the figures shown below, Yellow corresponds to Class 1 or bolt, green corresponds to Class 2 or nut, orange corresponds to Class 3 or ring, and red corresponds to Class 4 or scrap.

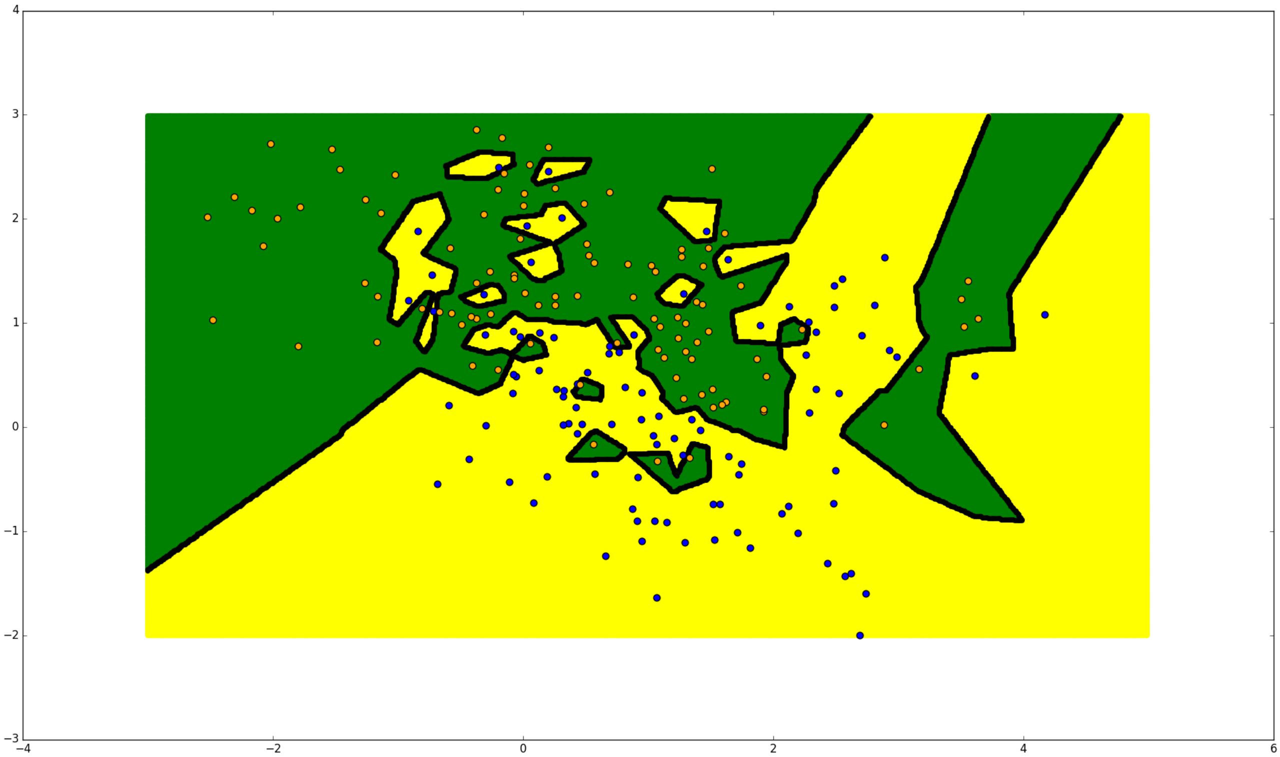

Figure 3: Classification using K-NN, $K = 1$ – The orange and blue circles represents the input points (Blue = Class 0, Orange = Class 1), The green and yellow represents the regions (with same respective class labels) and the black boundary in between the regions represent the classification boundary obtained using this method

(a) Bayes classification using uniform cost

Figure 6 shows the result of classifying a 4 class datapoints using bayesian classifier with uniform cost function.

(b) Bayes classification using non-uniform cost matrix

Here the cost matrix values are decided based on the amount of profit or loss that we would face, in correct classification of the classes. For example, if the profit earned by classifying the scrap accurately is relatively less, then we could assign it a less cost so that the region boundaries are more accurate for specific classes. Figure 7 illustrates the result of bayes classification using the cost matrix given in textbook.

(c) Bayes classification using uniform cost and increased prior for Class 1 data

Figure 8 shows the result of bayes classification using non-uniform cost and by increasing the prior of Class 2 by half the least prior value of all fur classes. The prior of Class 1 is thus increased by the same factor in order to maintain the total probability as 1.

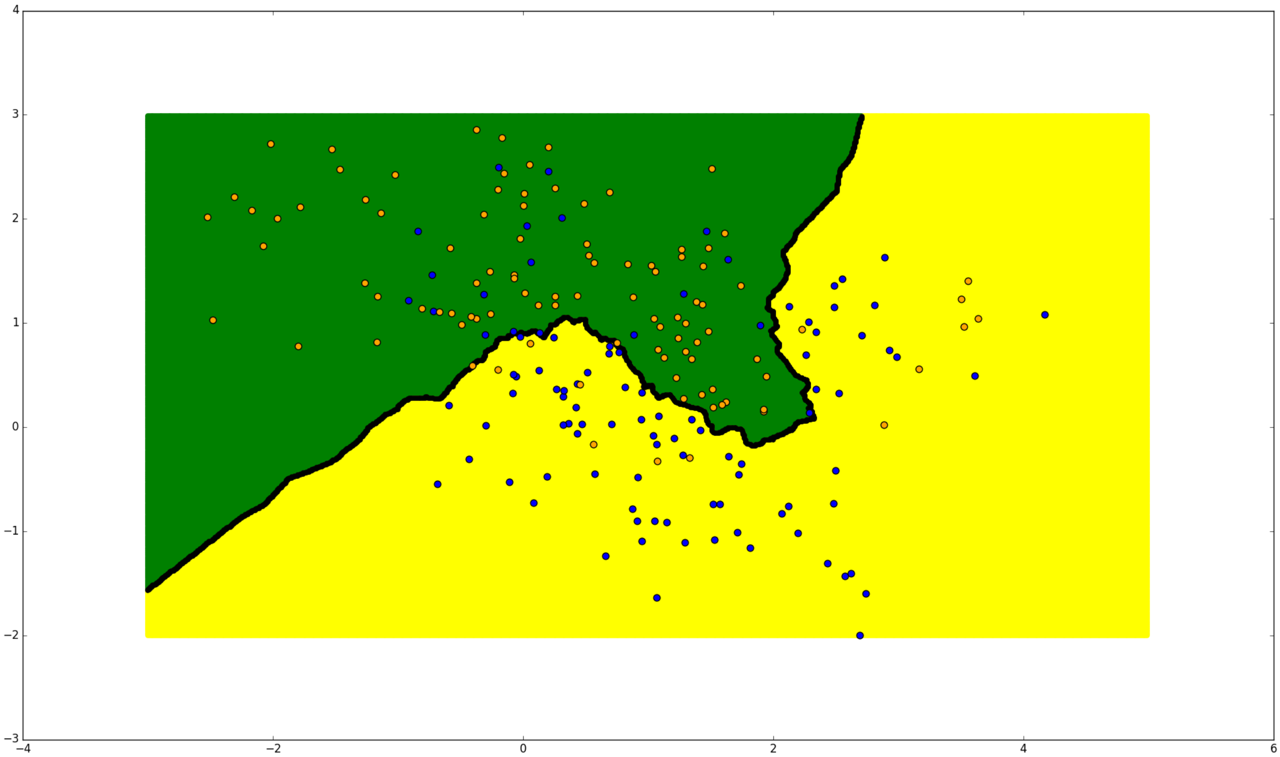

Figure 4: Classification using K-NN, $K = 15$ – The orange and blue circles represents the input points (Blue = Class 0, Orange = Class 1), The green and yellow represents the regions (with same respective class labels) and the black boundary in between the regions represent the classification boundary obtained using this method

On examining the result of three variants of bayes classification, shown in Figures **??**, we could clearly see that the boundaries in non-uniform cost case is pushed away, especially for Class 1 or "bolt". This is mainly because of the fact that this class has a high profit for classifying it correctly. This is denoted by a negative cost. The classifier hence tries to make it more important by increasing the regions for that class. Intuitively, this means that, if a scrap is classified as bolt, then there is no problem, when compared to a bolt classified as scrap which is a great loss. Also, we could notice an additional region of green or Class 2 that occurs in front of Class 3. This is again the effect of cost – the profit for Class 2 (nut) is more than that of Class 3 (which is ring). Now for the case where, the prior of Class 2 is increased, we could see that most of the region is classified as Class 1 and there is no scrap. The reason is that, since the cost matrix already has a high profit for Class 1, an increased prior of Class 2 is affecting the region of that class with least profit cost. Hence the boundaries of Class 2 are pushed away and scrap is totally eliminated and the region is occupied by bolt.

	Class 0	Class 1
Class 0	64	21
Class 1	36	79

Figure 5: Confusion matrix using MAP classification – Table shows the number of points classified as their true class and that are mistakenly classified (False classification).

3 Question 3 - Effect of increasing the scrap

This question is along the lines of what is discussed above. When the prior of a class is increased, it intuitively means that there are more data points for that class and hence the decision boundaries are pushed away from that class in order to give it more importance. Mathematically, the prior is multiplied with te risk function before making the decision. Hence, if the risk function remains the same and the prior is increased, then the risk for that particular class increases. So, in order to get the same classification result, the risk of that class should be decreased; this should specifically be done by decreasing the cost of scrap to scrap classification by the same factor, which is 2. In conclusion, the same decision boundaries can be obtained by dividing the cost of scrap to scrap classification by 2, since the prior is increased by a factor of 2.

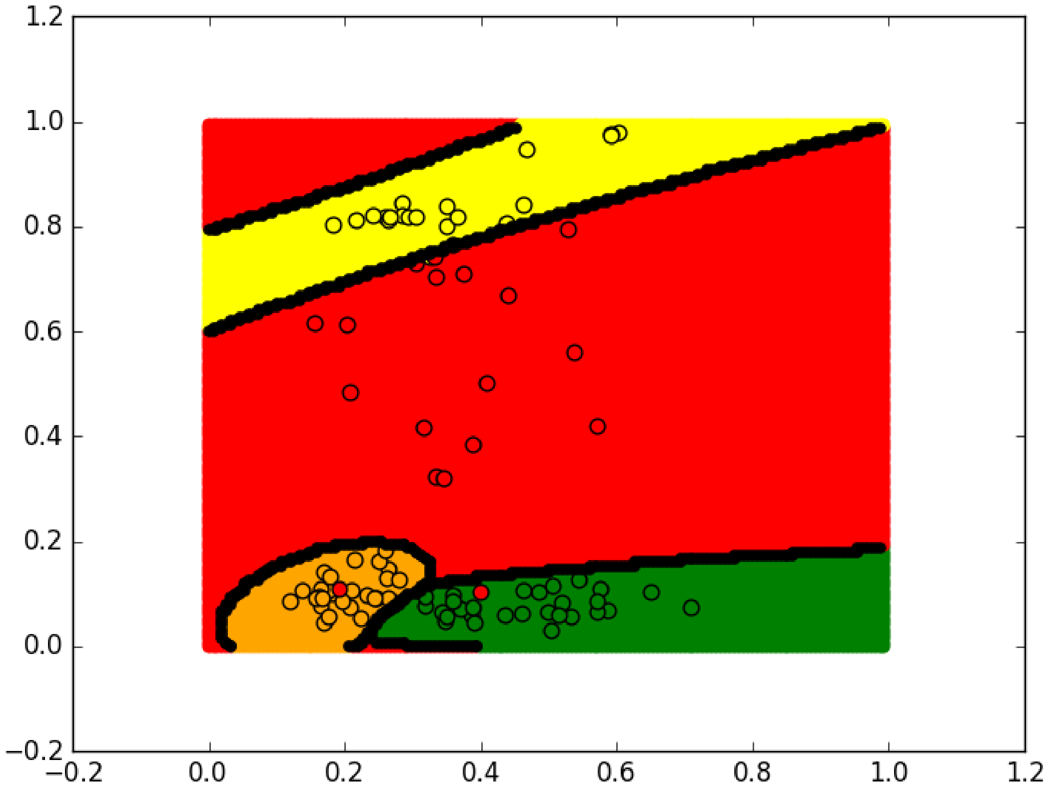

Figure 6: Classification using Bayes uniform cost – Yellow = Class 1 or bolt, green = Class 2 or nut, orange = Class 3 or ring, and red = Class 4 or scrap.

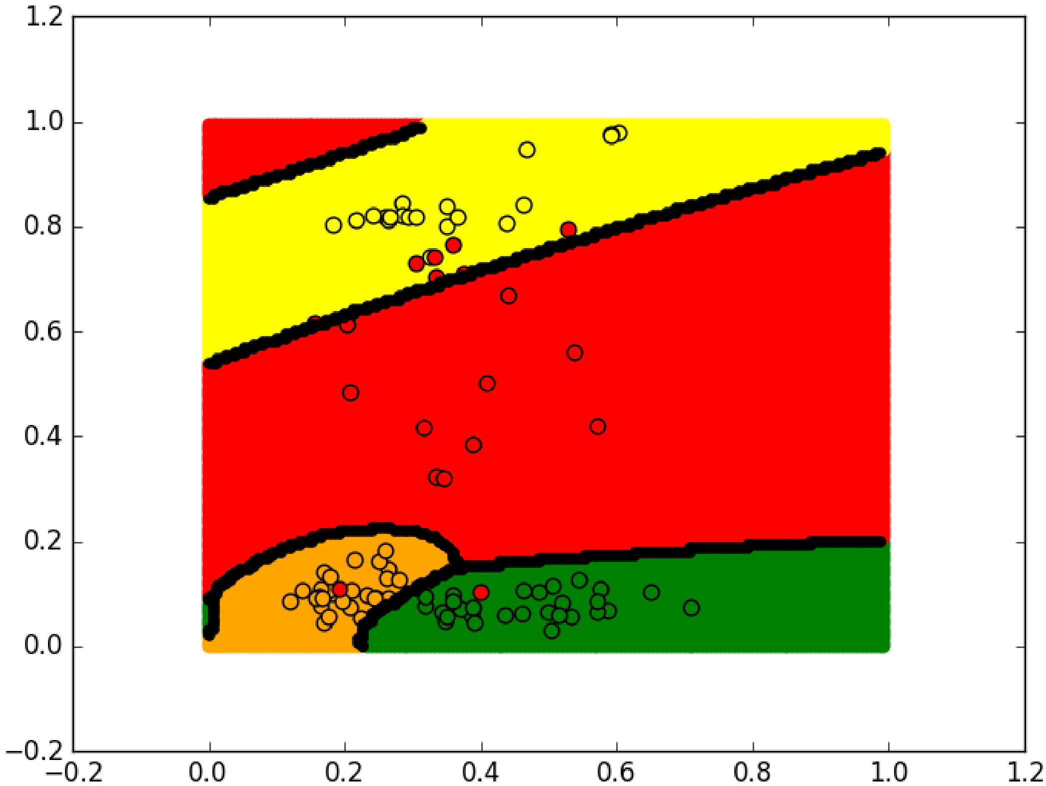

Figure 7: Classification using Bayes non-uniform cost – Yellow = Class 1 or bolt, green = Class 2 or nut, orange = Class 3 or ring, and red = Class 4 or scrap.

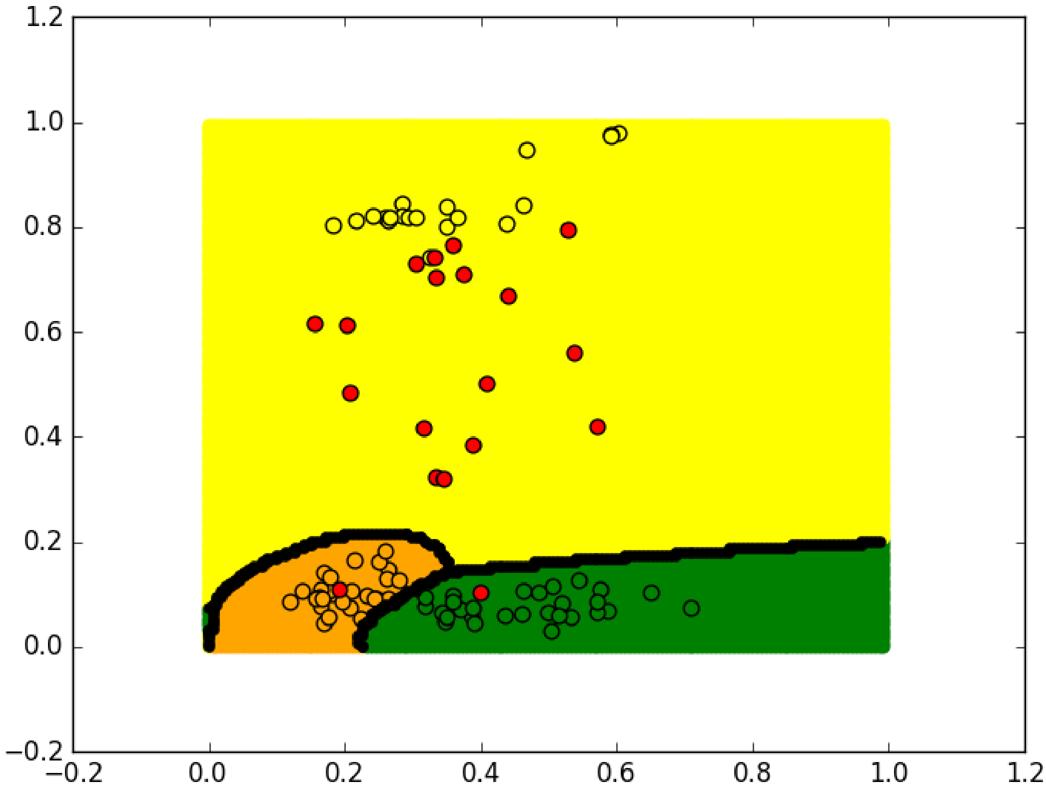

Figure 8: Classification using Bayes non-uniform cost and increased Class-2 prior – Yellow = Class 1 or bolt, green = Class 2 or nut, orange = Class 3 or ring, and red = Class 4 or scrap.

List of Figures