Predicting Haberman's Survival from patients' age and number of positive nodes using k-nearest neighbors

Pronob Barman

Dec 2020

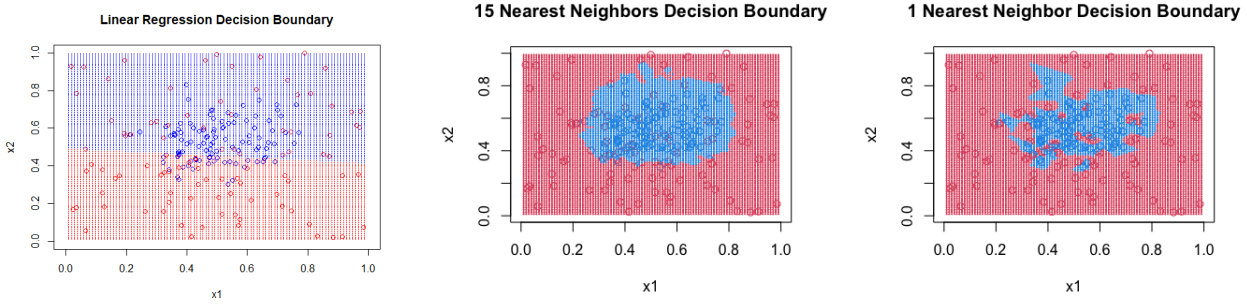
Problem 1. First Task

1. Place the six learners in order of how well they will classify the existing (training) data. Explain your reasoning.

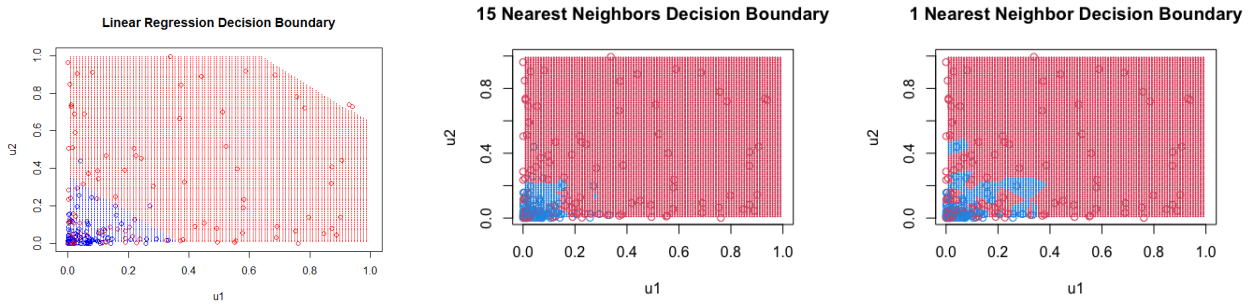
   In order from worst to best, we believe that the order is 1,4, 3, 6, 2, and 5. We believe that the linear regression will produce the worst predictions, given that linear regression is not the correct model for this data set. Specifically, our response variable is binary, and linear regression assumes a continuous response.

   Next, we believe that nearest neighbors with $k = 1$ (learners 3 and 6) will perform better than linear regression, but worse that nearest neighbors with $k = 15$. This is because $k = 1$ implies lower degrees of freedom, and therefore worse predictions than with $k = 15$ (note that we corrected this assumption in part (5), due to a misunderstanding of degrees of freedom in nearest neighbor classification). And lastly, we believe that the transformation (for learners 4, 5, and 6) will produce better predictions, which is reflected in the ordering above.

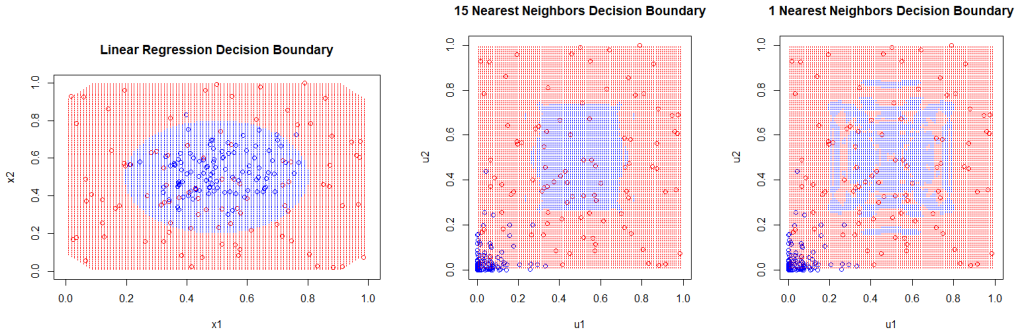
2. Create plots which show the decision boundaries in the x_1x_2 plane resulting from learners 1, 2, and 3 in the introduction.



3. Create plots which show the decision boundaries in the u_1u_2 plane resulting from learners 4, 5, and 6 in the introduction.



4. Create plots which show the decision boundaries in the x_1x_2 plane resulting from learners 4, 5, and 6 in the introduction.



5. For each of the learners, report how many observations from the training data were classified correctly. Compare your findings to what you anticipated in part (1). Were there any surprises, and can you explain them?

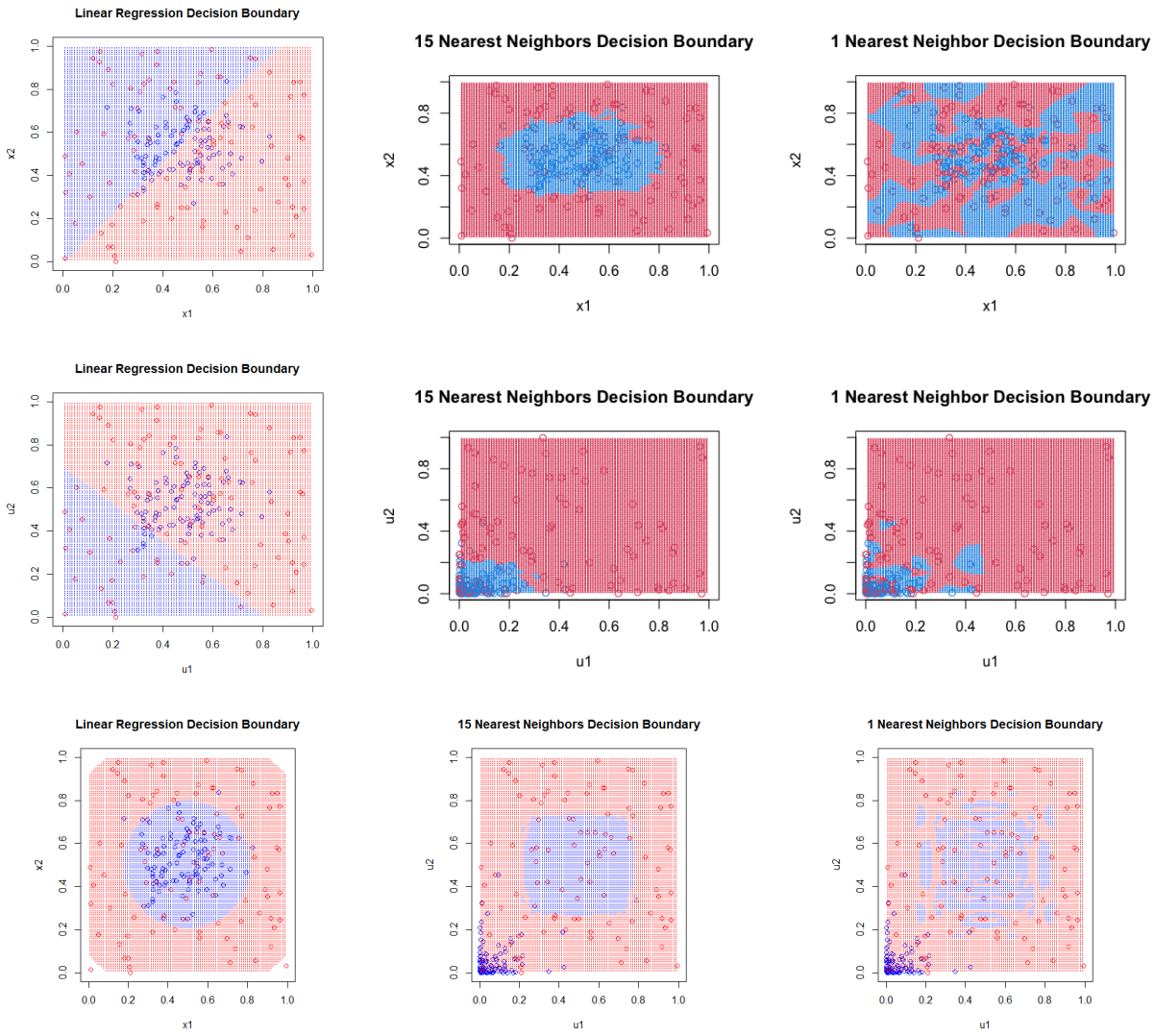
Learner #	Number of $u = 1$ (out of 104) correct	Number of $u = 0$ (out of 96) correct
1	74	52
2	98	64
3	104	96
4	102	59
5	93	73
6	104	96

Based on our results, the order (from worst to best) is 1, 2, 5, 4, 3, and 6 (3 and 6 tied). The first result (learner 1) matched our expectations from part (1). However, we did not expect learners 3 and 6 to perform perfectly. This is because we had misunderstood how many degrees of freedom these learners had. We now understand that for nearest neighbors, a lower k implies higher degrees of freedom, and vice-versa. Therefore, a learner with low k will make very accurate predictions, and a learner with high k will make poor predictions. This is due to the bias/variance trade off for nearest neighbors. Additionally, it is not clear that the transformation made a significant difference in the number of observations classified correctly.

6. Now, place the six learners in order of how well they will classify the new (validation) data.

   Based on our results from part (5), we believe that the order (from worst to best) will be 1, 2, 5, 4, 3, and 6, and that learners 3 and 6 will perform perfectly.

7. Reproduce parts (2), (3), and (4) above, using the validation data.



8. For each of the learners, report how many observations in the new (validation) data were classified correctly. Compare your findings to part (6). Were there any surprises, and can you explain them?

Learner #	Number of $u = 1$ (out of 104) correct	Number of $u = 1$ (out of 96) correct
1	47	58
2	90	73
3	99	101
4	95	69
5	92	75
6	99	101

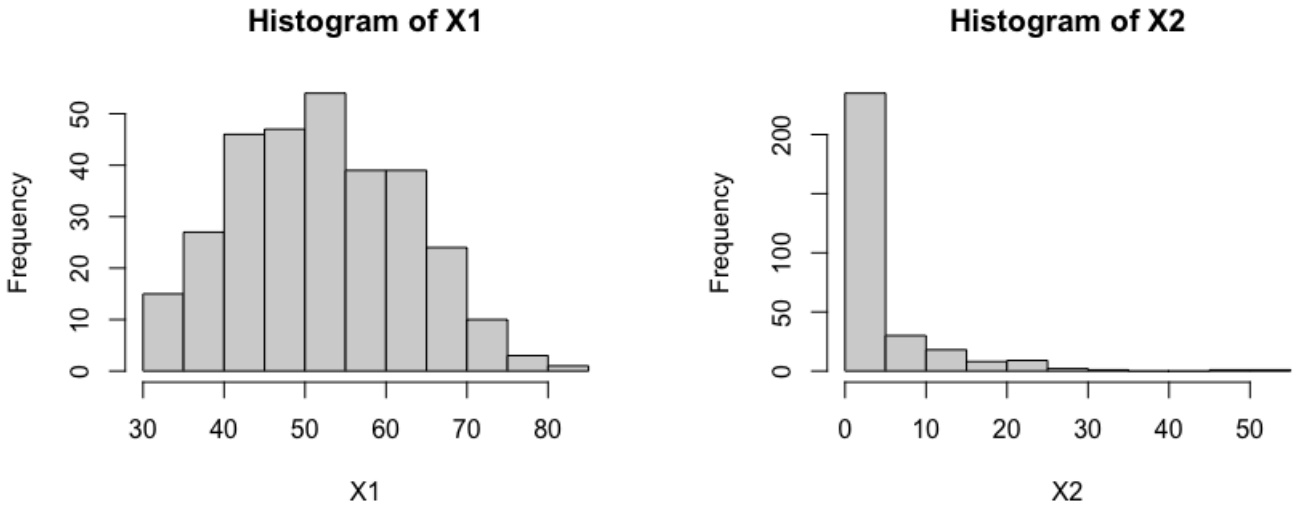
Based on our results, the order (from worst to best) is 1, 2, 5, 4, 3, and 6 (3 and 6 tied). These are the results which we anticipated in part (6).

Problem 2. Second Task

For the second task, involving the survival data, we first inspected the dataset by taking a look at the distribution of each variable, as well as possible relationships between the variables. After modifying the raw data and re-labeling the variables to create a working dataset, we obtained the following summary statistics from R:

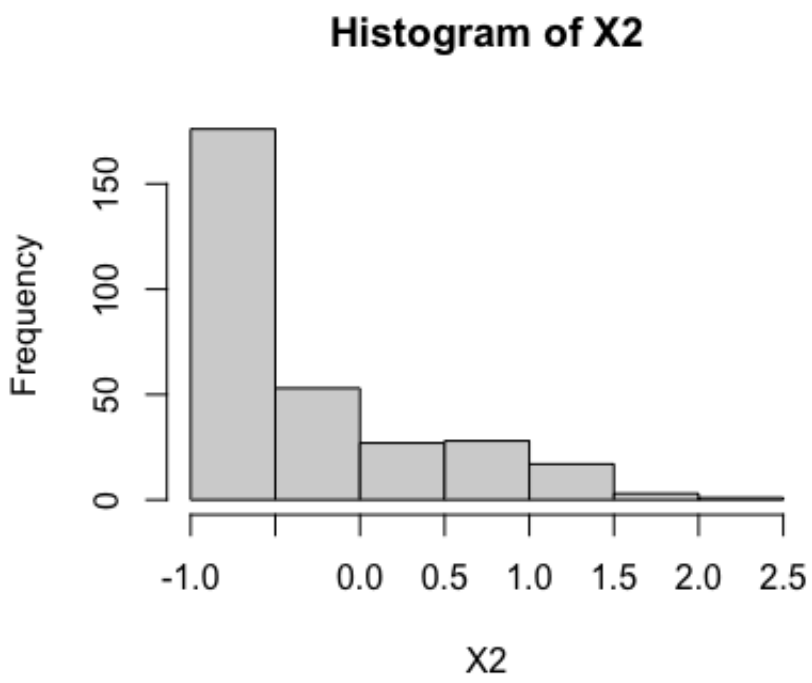
```
> str(surv)
'data.frame': 305 obs. of  3 variables:
 $ X1: int  30 30 31 31 33 33 34 34 34 34 ...
 $ X2: int  3 0 2 4 10 0 0 9 30 1 ...
 $ Y : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 2 2 1 1 ...
> #look at distributions of variables
> summary(X1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   44.00   52.00   52.53   61.00   83.00
> summary(X2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   4.036   4.000  52.000
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.266   2.000   2.000
> cor(X1,X2)
[1] -0.06654809
```

Additionally, we obtained the following histograms for variables $X1$ and $X2$:



As we can see, the distribution of $X1$ (age) is roughly symmetric, whereas the distribution of $X2$ (number of nodes) is right-skewed, with a large number of zeros. Additionally, Y (survival status) is a binary variable, with more patients who survived ($Y = 1$) vs. those who did not ($Y = 2$) as indicated by the mean of Y. Additionally, we may note that there is not a strong correlation between $X1$ and $X2$.

In order to address the issue of skewness with $X2$, we decided add one to each observation, and take the log of that result. The reason for adding one was so that the log of every observation would be defined. After taking this transformation, we obtained the following histogram of our new $X2$:



Additionally, we may note that $X1$ and $X2$ do not have the same units (i.e., years vs. number of nodes). To address this issue, we chose to standardize $X1$ and $X2$, or in other words:

$$X1' = \frac{X1 - \bar{X1}}{sd(X1)} \quad \text{and} \quad X2' = \frac{X2 - \bar{X2}}{sd(X2)}.$$

After inspecting each variable and addressing potential issues with the data, we proceeded to partition our dataset into training and validation data. To help reduce bias, we took a random subset of 80% of our data for training, and the remaining 20% for validation, using the `R` code below:

```
> #partition dataset into training and validation data
> set.seed(677899)
> train.ind<-sample(1:305,244)
> surv.train<-surv[train.ind,]
> surv.valid<-surv[-train.ind,]
```

Next, we investigated the efficacy of several different learners. We did not consider linear regression, due to the structure of the data. In other words, since the response variable (Y) is binary, it would not make sense to use linear regression (which assumes a continuous response) to model the data. Given that linear regression is not appropriate, we tested nearest neighbors at varying levels of k, beginning with the training data.

We began by using the training data to test nearest neighbors with values of k at 1, 5, 15, 25, and 40. After testing each learner, we determined that $k = 1$ and $k = 5$ provided the most accurate classifications without over fitting. The number of correct classifications for each of the learners tested is as follows:

Learner #	Number of $Y = 1$ (out of 179) correct	Number of $Y = 2$ (out of 65) correct
1	174	52
2	168	24
3	177	2
4	179	1
5	179	0

It should be noted that although some of the learners performed similarly in terms of total number of correct classifications, we had to make a decision based on the structure of the data. Specifically, since $Y = 1$ corresponds to a patient who survived and $Y = 2$ corresponds to a patient who did not survive, the consequence of classifying a "2" as a "1" may be considered less severe than the other way around.

Based on the above results, we concluded that nearest neighbors with $k = 1$ and $k = 5$ (learners #1 and #2, respectively) were the best. We then proceeded to investigate the performance of these learners on classifying the validation data. We obtained the following results:

Learner #	Number of $Y = 1$ (out of 45) correct	Number of $Y = 2$ (out of 16) correct
1	45	16
2	42	4

Based on these results, we concluded that nearest neighbors with $k = 1$ was the most appropriate for the data in the second task. Seeing as this learner did not classify the training data with 100% accuracy, we do not have sufficient evidence that this learner overfits the data.