Rohit Ravishankar
rr9105@rit.edu

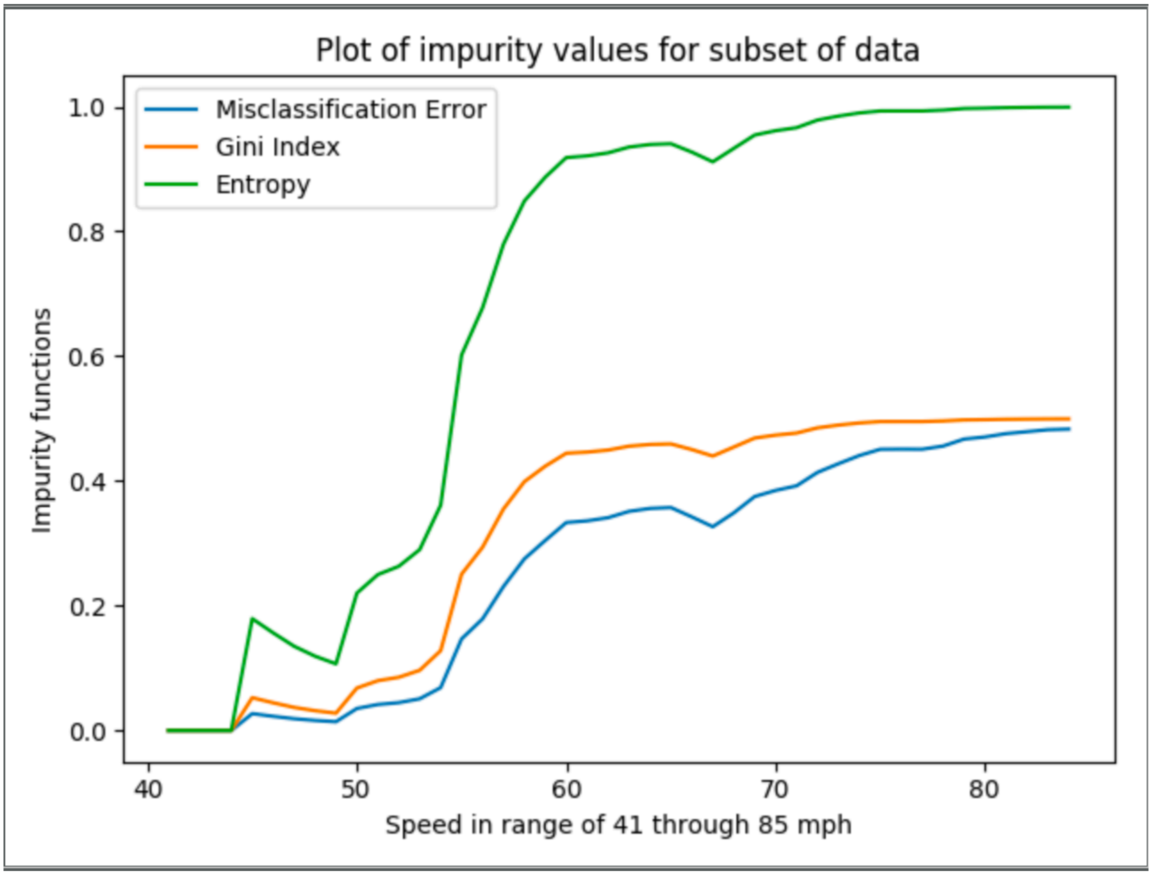
Home Work 3

CSCI - 720 Big Data Analytics

Collaborators: None

3.

d. Plot all three functions on the same graph. Again: only for the sub-set of data under the threshold. Include this plot in the PDF of your write-up.



Misclassification rate is from 0.0 through 0.48310139165009935

Gini index is from 0.0 through 0.49942887407167325

Entropy is from 0.0 through 0.9991758825220223

4.

Write up what you learned here.

Did you manage to write functions that compute the Entropy, Gini, and Entropy? Was there anything particularly challenging? Did anything go wrong?

Ans. We learned how to calculate the misclassification rate, gini index and entropy values for a given subset of data. There was nothing I found particularly challenging. Nothing went wrong during the implementation.

Misclassification error for each threshold was calculated as:

$1 - max(P(\text{wanting to speed}),\ P(\text{not wanting to speed}))$

Gini index for each threshold was calculated as:

$1 - sum(P(\text{wanting to speed})^2,\ P(\text{not wanting to speed})^2)$

Entropy for each threshold was calculated as:

Case 1:

Entropy $= -(P(\text{wanting to speed}) * log_2(P(\text{wanting to speed}) +$

$P(\text{not wanting to speed}) * log_2(P(\text{not wanting to speed}))$

Case 2:

if P(wanting to speed) == 0 **then**

Entropy $= -(P(\text{not wanting to speed}) * log_2(P(\text{not wanting to speed}))$

Case 3:

if P(not wanting to speed) == 0 **then**

Entropy $= -(P(\text{wanting to speed}) * log_2(P(\text{wanting to speed}))$

Case 4:

if P(not wanting to speed) == 0 and P(wanting to speed) == 0 **then**

Entropy = 0

5.

In addition to the original graph, graph all three mixed functions on a single plot.

a. Mixed misclassification error

b. The mixed GINI index

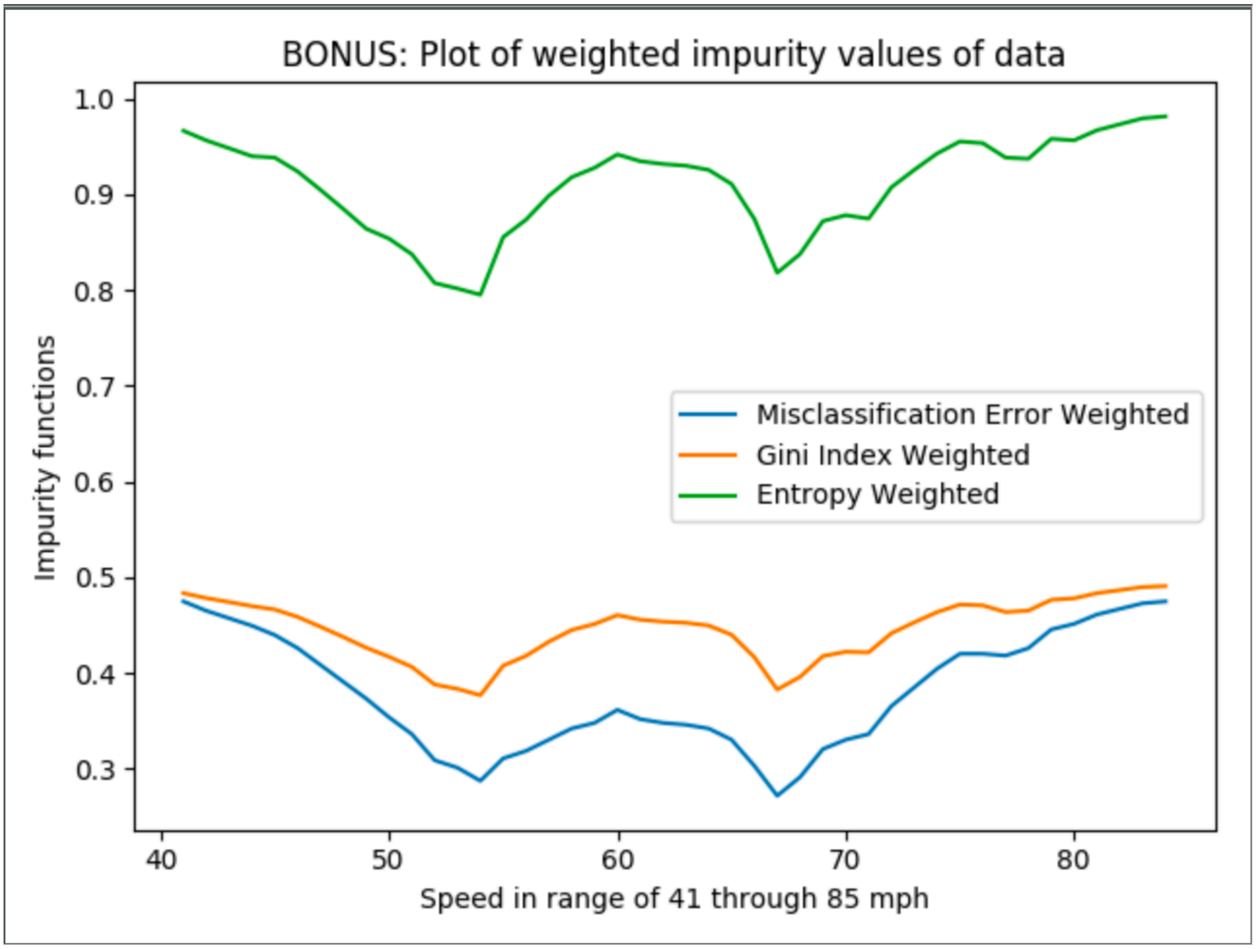
c. The mixed entropy

This requires doing three or four times as much work because you have to compute a Gini Index for both the data below the threshold, and the data ≥ the threshold, and then average them together, for every single possible threshold. Then you have to average all the values together

There is much more room for making a mistake here.

If you do the bonus, explicitly label the word BONUS before your graph, so the grader can identify.

Ans.


BONUS: Plot of weighted impurity values of data
Impurity functions
Speed in range of 41 through 85 mph
Misclassification Error Weighted
Gini Index Weighted
Entropy Weighted