CMPUT 466 – Machine Learning

Decision Tree

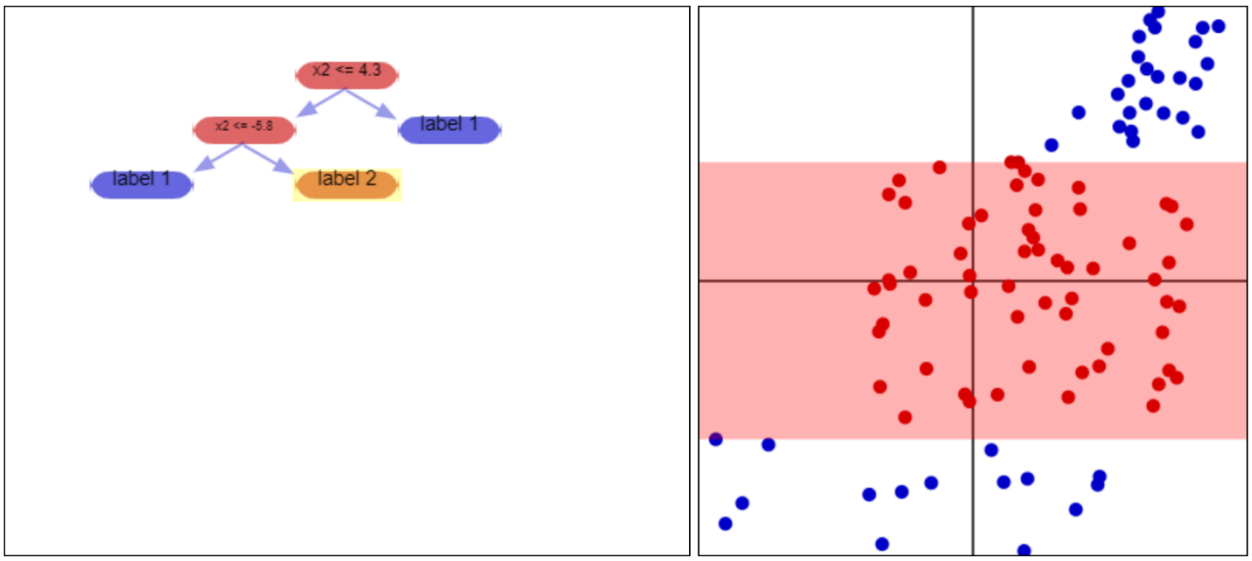
Danh Nguyen

Question 2.

a. When hover over the root node, the whole space is highlighted. When move the cursor to the child nodes, we get the highlights of the 2 partitions of the space. In which 1 of them contains only 1 type of color dot (blue in this case), the other contains a mixture of dots. This corresponds to the leaf node and the sub-tree, respectively. The separation is parallel to 1 of the axes.

   The reason the highlight change is because the root node separates the data into 2 groups, where in each group we have less entropy than before.

b. If we reach the leaf node, it means that we reached the final prediction.



1. Yes, all the vectors in the highlighted region have the same color.
2. The output for this region is “red (label 2)”.
3. For the current dataset, it correctly labeled them all. For future prediction, it will have a good chance to predict the class of the vector if it happens to be in this region. However, there is no guarantee 100% correctness, as we see in the next example.

c. We can say the blue dot happened to appear in the red region is “noise” and if we split this region again into 2 it will lead to “overfitting”. It is true that we should aim for lowest entropy in each leaf; however, decision tree is notorious for easy to be overly complex and causes instability for future predictions (CMPUT 466 Lecture Notes).

d. The cons:

   Stated above, decision tree con grow overly complex very quick (overfitting) and will causes unstable predictions. The blue dot alone is not enough to split further and can be considered noise.

   The pros:

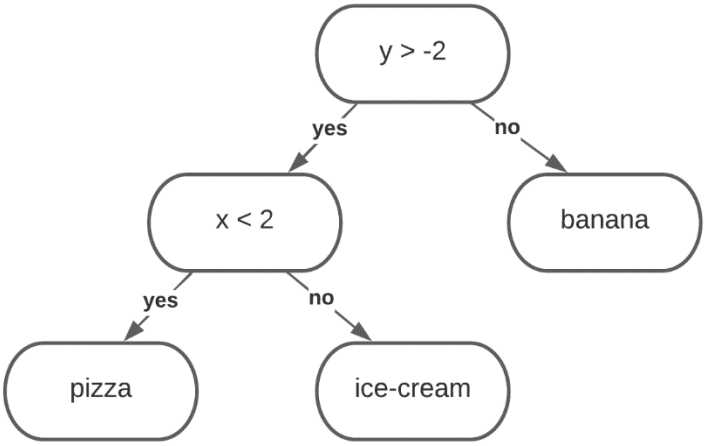
   By observation, this blue dot is far away from the red group (proper notion of metric here is Euclidean distance). Because of that, we can also predict that the blue dot may represent another well-classified group (since they are distinct – their distance is far away). Also, because they are far away, a separator in this region has low chance of over fitting (compare to 2 other regions).

e. Calculations:
   - Region 1 entropy = -(6/8)lg2(6/8) –(2/8)lg(2/8) = 0.811
   - Region 2 entropy = trivially 0 (pure)
   - Region 3 entropy = 0.650

f. AE = 0.811*(8/18) + 0 + 0.650*(6/18) = 0.577

g. Consider the horizontal separation cuts y-axis at about -2, vertical separation at x = 2.

(intentionally left blank, continue next page)



Classifier (DT) for given dataset

h. Pruning:
Performing a Chi-Square test on the "color" node (classification via color):

Results						
	red	green	blue			***Row Totals***
glass	2 (1.92) [0.00]	1 (1.92) [0.44]	3 (2.16) [0.33]			6
plastic	6 (6.08) [0.00]	7 (6.08) [0.14]	6 (6.84) [0.10]			19
Column Totals	8	8	9			**25 (Grand Total)**

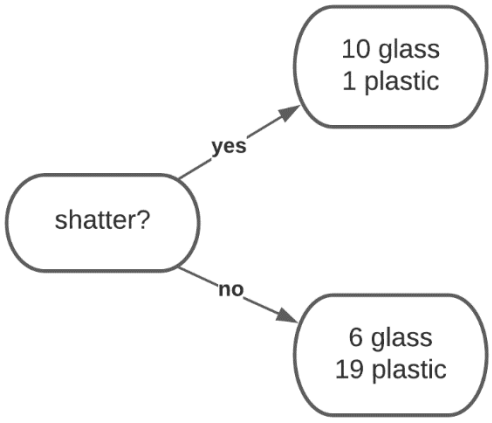
The chi-square statistic is 1.0143. The *p*-value is .602223. The result is *not* significant at $p < .10$.

Calculation and image in courtesy of Chi-Squared calculator at https://www.socscistatistics.com/tests/chisquare2/default2.aspx.

From the result above, chi-value = 1.01, p-value is 0.60. For $p < 0.1$, we conclude that the classes are independent from each other under this classification; and thus, the node should be removed.

(intentionally left blank, continue next page)

Now the tree only has the root node, which is the classification via drop test (shatter or not). Since we pruned the color node, its new look would be:



Now we perform Chi calculation again:

Results						
	shatter	no shatter				***Row Totals***
glass	10 (4.89) [5.34]	6 (11.11) [2.35]				16
plastic	1 (6.11) [4.27]	19 (13.89) [1.88]				20
Column Totals	11	25				**36 (Grand Total)**

The chi-square statistic is 13.8502. The *p*-value is .000198. The result is significant at $p < .10$.

Calculation and image in courtesy of Chi-Squared calculator at https://www.socscistatistics.com/tests/chisquare2/default2.aspx.

From the result above, chi-value = 13.9, p-value is ~0. For $p < 0.1$, we conclude that the classes are dependent from each other under this classification; and thus, the node correct and should NOT be removed.

In conclusion, the tree above is final.

END OF ASSIGNMENT