AS2

Editor: Wang Peng

UID:3035420027

Question 1 (Decision tree)

- a) First split

  split with a

$a=0$	$a=1$
$0:3,\frac{3}{8}$	$1:5,\frac{5}{8}$
$C0:0,0;C1:3,1$	$C0:3,\frac{3}{5};C1:2,\frac{2}{5}$

  $Gini=\frac{5}{8}\times(1-(\frac{3}{5})^2-(\frac{2}{5})^2)=\frac{3}{10}$

  split with b

$b=0$	$b=1$
$0:4,\frac{1}{2}$	$1:4,\frac{1}{2}$
$C0:1,\frac{1}{4};C1:3,\frac{3}{4}$	$C0:2,\frac{1}{2};C1:2,\frac{1}{2}$

  $Gini=\frac{1}{2}\times(1-(\frac{1}{4})^2-(\frac{3}{4})^2)+\frac{1}{2}\times\frac{1}{2}=\frac{7}{16}$

  So,we should first split with a

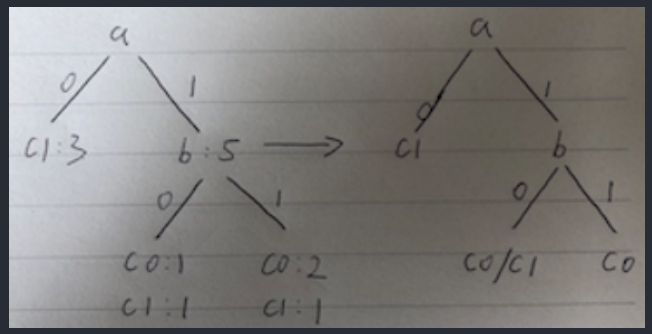
  Second split

  Apparently, we can only split with b

$b=0$	$b=1$
$0:2,\frac{2}{5}$	$1:3,\frac{3}{5}$
$C0:1,\frac{1}{2};C1:1,\frac{1}{2}$	$C0:2,\frac{2}{3};C1:1,\frac{1}{3}$

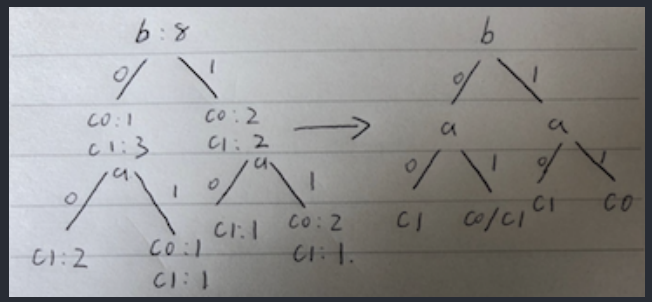
$Gini=\frac{3}{5}\times(1-(\frac{2}{3})^2-(\frac{1}{3})^2)+\frac{2}{5}\times\frac{1}{2}=\frac{7}{15}$

Finally, the tree looks like this:



- b) I suppose not

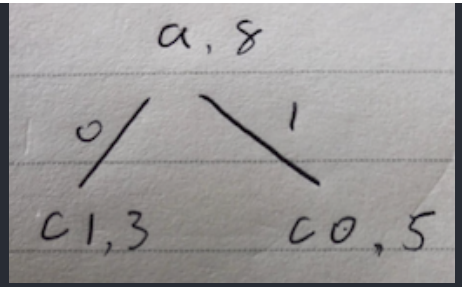
  After calculation, generalization error of first split with a = $\frac{2+0.5\times3}{8}=0.4375$ And the tree structure of first split with b looks like this:



  generalization error of first split with b = $\frac{2+0.5\times4}{8}=0.5$

  We can find that the cost of these 2 trees are really close. So it is too hard to judge without calculation.

- c) Sure, it can be pruned and the new tree looks like this:



We can find that even the leaves of b are pruned, the misclassfied items are still 2.But the number of leaves decreases to 2, which means the generalization error decreases as well.

- d) $a = 1, b = 1 -> class0$

Question 2 (NaiveBayes)

junior	senior
$P(junior) = \frac{7}{11}$	$P(senior) = \frac{4}{11}$
$P(marketing\|junior) = \frac{2}{7}$	$P(marketing\|senior) = \frac{1}{4}$
$P(31-35\|junior) = \frac{3}{7}$	$P(31-35\|senior) = \frac{1}{4}$
$P(46k-50k\|junior) = \frac{3}{7}$	$P(46k-50k\|senior) = \frac{1}{4}$
$Product = \frac{18}{539}$	$Product = \frac{1}{176}$

Apparently, $P(junior) > P(senior), -> junior$

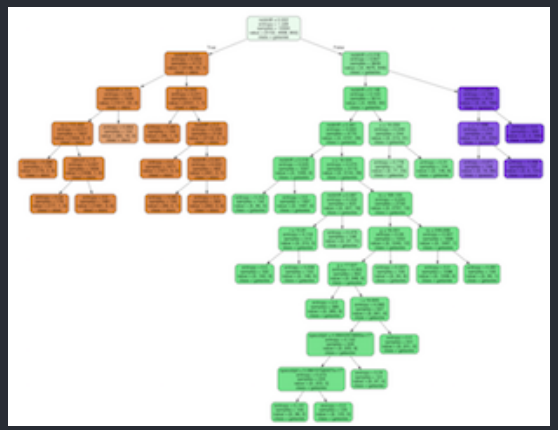
Question 3 (hands-on on decision tree classifier)

All the details can be found in question3.ipynb

1.

2. $GeneralizationError = 0.00515$
3. With min_samples_leaf=100, the GeneralizationError increase to 0.0126 compared with Part 2 as the num of misclassified labels increases even though the num of leaves decreases. To minimize the new GeneralizationError, most parameters do not work in DecisionTreeClassifier because of min_samples_leaf=100. However, I find that when Criterion is changed to 'entropy', the error will decrease slightly to 0.0126 from 0.0127
4. I think it depends: if you want 100% accuracy, the tree built in Part 1 is perfect as its num of misclassified labels is 0; However, if you want to classify faster and can accept a low error rate, the modified tree is OK with error rate lower than 0.0113.

   In this case, my recommendation is the modified tree. In the next part, you can find that the structure of it is really clear with acceptable error rate
5.

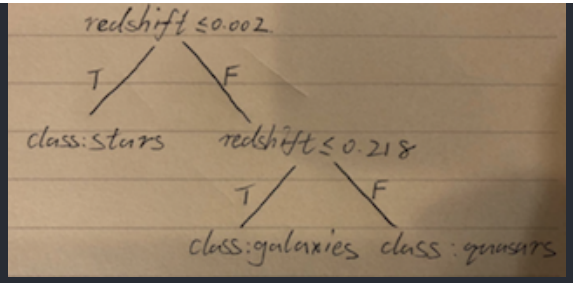


As you can see from the modified tree, the structure of this tree is amazing. All the work is done with only 2 splits:

- split with redshift <= 0.002, then we will get all the class 'Star' on the left side;
- split with redshift <= 0.218, then class 'quasars' will be on the right side while the remaining items are all class 'galaxies'

6. Sure. According to part 5, only the first 2 splits in the modified tree are necessary, which can classify really well with low error rate.

   The pruned tree looks like this:



7.

Question 4 (hands-on naive Bayes classifier)

All details can be found at question4.ipynb

1. average accuracy of MultinomialNB = 0.7375

   average accuracy of GaussianNB = 0.7125
2. As you can see, the accuracy of **random classifier** is around 0.137, which much worse than **MultinomialNB**
3. After test, the accuracy increases to 0.7875 when we do not ignore stopwords. I guess maybe some stopwords contain key information and can boost the accuracy
4. seta=[‘polguns’ , ‘hockey’ , ‘machw’]

   setb=[‘electronics’ , ‘ibmhw’ , ‘machw’]

   After test, the accuracy of seta is higher than setb. I suppose that it is because the seta contains 3 totally different items while the items of setb all belong to technology area.