Computer Science 145, Homework 4

Michael Wu
UID: 404751542

February 22nd, 2019

Problem 1

For purity and normalized mutual information, we can assign the cluster labels to a ground truth label according to the majority of the ground truth labels in the cluster. This would mean cluster 1 corresponds to ground truth label 2, cluster 2 corresponds to ground truth label 3, cluster 3 corresponds to ground truth label 1, and cluster 4 corresponds to ground truth label 4. We can construct the following table.

Ground \ Cluster	1	2	3	4	Total
2	5	0	0	0	5
3	0	5	0	0	5
1	0	1	4	0	5
4	0	0	1	4	5
Total	5	6	4	4	20

Overall 18 of the data points are matched, so our purity is 0.9. For normalized mutual information, we can use this table to obtain the following expression for information using the base 10 logarithm.

$$I(C,\Omega) = \frac{5}{20}\log\left(\frac{20\times 5}{5\times 5}\right) + \frac{5}{20}\log\left(\frac{20\times 5}{6\times 5}\right) + \frac{1}{20}\log\left(\frac{20\times 1}{6\times 5}\right)$$
$$+ \frac{4}{20}\log\left(\frac{20\times 4}{4\times 5}\right) + \frac{1}{20}\log\left(\frac{20\times 1}{4\times 5}\right) + \frac{4}{20}\log\left(\frac{20\times 4}{4\times 5}\right) \approx 0.513254$$

Because each ground truth label has 5 data points, the entropy of the ground truth labels is given by the following expression.

$$H(\Omega) = -\log\left(\frac{1}{4}\right) \approx 0.60206$$

The entropy of the clusters is given by the following expression.

$$H(C) = -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{10}\log\left(\frac{3}{10}\right) - \frac{2}{5}\log\left(\frac{1}{5}\right) \approx 0.586967$$

Then we have the normalized mean information shown below.

$$NMI(C,\Omega) = \frac{I(C,\Omega)}{\sqrt{H(C)H(\Omega)}} \approx 0.863387$$

In order to calculate the precision, recall, and F-score, we need to know the confusion matrix of this clustering. I obtained this by running the following code.

```
results = [(3,2), (3,2), (1,3), (1,3), (1,3),
  (4,4), (3,2), (3,2), (4,3), (2,1),
  (4,4), (2,1), (1,3), (2,1), (3,2),
  (2,1), (1,2), (2,1), (4,4), (4,4)]

TP=0
FN=0
FP=0
TN=0
for i in range(len(results)):
  for j in range(i+1, len(results)):
    a_ground, a_predicted = results[i];
    b_ground, b_predicted = results[j];
    if a_ground==b_ground:
      if a_predicted==b_predicted:
        TP+=1
      else:
        FN+=1
    elif a_predicted==b_predicted:
      FP+=1
```

```
    else:
      TN+=1

print(TP, " ", FN, " ", FP, " ", TN)
```

This told me that there were 32 true positives, 8 false negatives, 9 false positives, and 141 true negatives. The precision is $\frac{32}{32+9} \approx 0.780488$. The recall is $\frac{32}{32+8} = 0.8$. The F-score is then the following.

$$\frac{2 \times 0.780488 \times 0.8}{0.780488 + 0.8} \approx 0.790124$$

Overall this is a fairly good clustering since all our metrics have high scores.

Problem 2

a)

b)

c)

Problem 3

a)

b)

c)

Problem 4

a)

b)

c)