



So in this question, we have to evaluate clustering results from two algorithms and tell which algorithm is best as per to metrices. The first one is purity. Purity is simply about how many objects are in the class where they should be.

So, lets look at the cluster one of algorithm 1. Now there are two objects of square class and two of circle. One must wonder that both classes are equally prevalent so how can we decide that which is the dominant class? Well it actually does not matter. Whatever is the dominant class, two objects are at the right place and the remaining two are at the wrong place.

In the second cluster of algorithm 2, circle is dominant class and two circles are at the right place. So a total of 4 objects (2 in cluster 1 and 2 in cluster 2 are at the right place). Therefore, the purity is 4/7 (total number of objects is 7).

Now, for algorithm 2, lets look at cluster 1. Two objects are at right place (squares). In cluster 2, three objects are at the right place (circle). So 5 out of 7 objects are at the right place. So here, the purity is 5/7

**Clearly, 5/7 is greater than 4/7. Therefore, Algorithm 2 is better than Algorithm 1 as per purity metric.**

**Now as per F1 Score,**

To calculate F1 score, we need to calculate three other metrics. They are, Total Positive, False Positive and False Negative.

No notice a 2x2 matrix given in the question that describes what all these metrices are.

Total Positive is the total number of objects that are of same type and same class.

In Cluster 1, there are 2 squares and 2 circles. So, 2 circles have the same class and some cluster and two squares have the same class and same cluster. In cluster 2, there are two circles with same class and same cluster. So, there are three such instances where two points are of same type and are in same circle. So, **TP here is 3.**

False Positive is defined as total number of points that are in same cluster but are of different class.

Now, in cluster 1, there are two circles and two square. We can form four pairs that satisfy the condition of same cluster but different class. They are, (sq1, c1), (sq2, c1), (sq1, c2) and (sq2. C2). Similarly, in cluster 2, There is a pair (sq1, c) and another pair (sq2, c) that fulfils the condition of same cluster but different class. So, **we have 6 pairs that fulfil this** condition.

Now, false negative is same class but different cluster.

Lets look at square class first. There are two squares in cluster 1 and 1 square in cluster 2. So there are two pairs of same class but different cluster. (Cluster-1’s square 1 with only square of cluster-2) and (cluster-1;s square 2 with only square of cluster-2)

Lets look at circle class next. There are two circles is cluster 1 and two in cluster 2. So there are four pairs of points that satisfy, ‘same class and different cluster’. So, **we have a total of 6 FN.**

**Now, F1 = (2 x 3)/(2 x 3 + 6 + 6) = 6/18 == 1/3**

**Now, in cluster 2,**

True positive (same class, same cluster) = 1 for cluster 1 (sq 1 and sq2). And 3 for cluster 2 (c1, c2), (c2, c3) and (c3, c1). So, a total of 4.

False negative (same class, different cluster = 2 for squares and 3 for circles. So, a total of 5

False positive (different class, same cluster) = 2 for cluster 1 (sq1, c1) and (sq2, c1) and 3 for cluster 2 (c1, sq), (c2, sq), (c3, sq). So, a total of 5.

**Now, F1 = (2 x 4) / (2 x 4 + 5 + 5) = 8/18 = 4/9**

Now, 4/9 is more than 1/3, so algorithm 2 is better than algorithm 1 according to F1 score.