Northeastern University, Data Mining Techniques - CS6220
Fall 2017

Solutions to Homework 2, Part 3

Nakul Camasamudram

October 25, 2017

3 Evaluation

3.2 Purity

Solution:

Table 8.9 from TSK:

Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

a. Calculation: Let p_i be the purity of the i^{th} cluster, m_i be the number of objects in the i_{th} cluster and m_{ij} be the number of objects of class j that under cluster i.

- $p_1 = \frac{1}{m_1} \cdot max_j(m_{1j}) = \frac{1}{677} \cdot 506 = 0.7474$
- $p_2 = \frac{1}{m_2} \cdot max_j(m_{2j}) = \frac{1}{361} \cdot 280 = 0.7756$
- $p_3 = \frac{1}{m_3} \cdot max_j(m_{3j}) = \frac{1}{685} \cdot 671 = 0.9796$
- $p_4 = \frac{1}{m_4} \cdot max_j(m_{4j}) = \frac{1}{369} \cdot 162 = 0.4390$
- $p_5 = \frac{1}{m_5} \cdot max_j(m_{5j}) = \frac{1}{464} \cdot 331 = 0.7134$
- $p_6 = \frac{1}{m_6} \cdot max_j(m_{6j}) = \frac{1}{648} \cdot 358 = 0.5525$

Overall purity of the clustering $= \sum_{i=1}^{6} \frac{m_i}{m} \cdot p_i = \frac{1}{[354+555+341+943+273+738]} \cdot [506 + 280 + 671 + 162 + 331 + 358] = 0.7203$

b. Purity is a measure of the extent to which each cluster contains only objects from a single class. So, ideally, the overall purity of the clustering should be 1. In the above case, since it is 0.72034, the clustering is good.

By this metric, Cluster 3 is particularly good as it contains objects almost entirely from the "Sports" class.