Class Test 02

CSE 437 – Pattern Recognition

Department of Computer Science and Engineering

University of Asia Pacific

Total Marks: 20 Time: 40 Minutes

1. Consider the following dataset—

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Tumor Size** | **Stage** | **Smoker** | **Cancer Detection** |
| 01 | Small | B | Yes | Positive |
| 02 | Small | A | No | Negative |
| 03 | Small | B | No | Negative |
| 04 | Medium | B | Yes | Positive |
| 05 | Medium | C | No | Negative |
| 06 | Large | B | Yes | Positive |
| 07 | Large | C | Yes | Positive |

Now, calculate information gain for each attribute. [10]

1. Suppose there are two cluster A and B. Cluster A contains the datapoints— (1, 1), (1, 2), (2, 2), (3,1) and their ***current*** centroid is (3,2). Cluster B contains the datapoints— (4, 4), (4, 5), (5, 6), (6, 7) and their ***current*** centroid is (5, 5). Determine the centroid positions after next iteration. Here one iteration means both ‘color change step’ and ‘centroid update step’ combined. [10]