**Q1- Part 1**



J value is no more decreasing after k=5

* Y axis values might be different. If they showed similar curvature, give them full credit
* If they missed the source code, -5

**Q1-Part 2**



GMM and k-Mean have shown same clustering results with various k for this data set.

Hierarchical Clustering result can be shown

* Give them full credit when they show different clustering across k
* Depending on their programming language, the clustering results might be different. If they implemented correctly and provide any conclusion, give them full credit
* Just showing off the plot alone, -3

 Distance: Euclidean, Linkage: complete

* If they missed to specify the parameters (distance, linkage), -5
* If they missed the source code -5

**Q2-Part 1**

**(1)**



J values was decreasing drastically from k=2

* If their curve has a pick (something wrong), -3
* If they did not use k={1,..,8}, -2

**(2)**



Distance metric: Euclidean distance

Linkage : Complete, single, average (from left to right)

* Dendrogram might be different in terms of their language. Just focusing on the dendrogram should be different based on their metric
* If they did not ignore the last column, -5

**Q2-Part 2**

K mean clustering

|  |  |  |
| --- | --- | --- |
| Cluster | 1 | 2 |
| Benign | 447 | 224 |
| Malignant | 17 | 11 |
| Total | 464 | 235 |

*P* = (447+224)/699 = 0.96

GMM clustering

|  |  |  |
| --- | --- | --- |
| Cluster | 1 | 2 |
| Benign | 367 | 91 |
| Malignant | 0 | 241 |
| Total | 367 | 332 |

*P*=(367+241)/699 = 0.87

* *P* values might be different based on their clustering. Do not grade them strictly
* However, if their solutions are totally different (e.g., P of kmean clustering is 0.6), -3 for each clustering method
* If they provided the P values separately, e.g., providing P of cluster 1 and 2 in separate, give them full credit but leave comments that the actual solution should be collective *P*
* If they did not understand the concept of true positive and true negative, -5

**Q3**



BIC curve for random data set (left) and breast cancer (right).

For random data, BIC has shown the minimum point at k=5. It is obvious because the original data is created by 5 clusters. For real data set, we can find that k=5 and k=9 have shown minimum BIC. We can conclude that they(samples in the data) might be clustered with 5 or 9 components.

* If they presented the curves similar to the solution, give full credit
* If they just tried, give 3
* If they corrected the answer only one, give 7.