|  |  |
| --- | --- |
| STATISTICS WORKSHEET-2 | |
| 1 | B(mean) |
| 2 | C(12) |
| 3 | D |
| 4 | C |
| 5 | B |
| 6 | B |
| 7 | A |
| 8 | B |
| 9 | D |
| 10 | A |
| 11 | C |
| 12 | D |
| 13 | D |
| 14 | A |
| 15 | D |

|  |  |
| --- | --- |
| WORKSHEET 2 SQL | |
| 1 | D |
| 2 | C |
| 3 | A |
| 4 | D |
| 5 | B |
| 6 | A |
| 7 | A |
| 8 | C |
| 9 | C |
| 10 | C |
| 11 | A |
| 12 | C |
| 13 | A |
| 14 | C |
| 15 | C |

|  |  |
| --- | --- |
| MACHINE LEARNING | |
| 1 | A |
| 2 | D |
| 3 | A |
| 4 | A |
| 5 | B |
| 6 | B |
| 7 | A |
| 8 | D |
| 9 | D |
| 10 | A |
| 11 | D |
| 12 | D |
| 13 | The k-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier.   An example, is the average of the salaries of the following people:   $50k, $20k, $35k, $65k and $1 Million |
| 14 | K-means is like the Exchange Sort algorithm. Easy to understand, helps one get into the topic, but should never be used for anything real, ever. In the case of Exchange Sort, even Bubble Sort is better because it can stop early if the array is partially sorted. In the case of K-means, the EM algorithm is the same algorithm but assumes Gaussian distributions for clusters instead of the uniform distribution assumption of K-means. K-means is an edge case of E-M when all clusters have diagonal covariance matrices. The Gaussian structure means that the clusters shrink-wrap themselves to the data in a very nice way. This gets around the serious objections you correctly raise in the question. And E-M is not much more expensive than K-means, really. (I can implement both in an Excel spreadsheet.) But for serious clustering applications, one should really look at the hierarchical spectrum from single-link to complete-link clustering. |
| 15 | The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method. |