**Machine learning worksheet**

1.The Answer for question no 1 is: option b

2. The Answer for question no 2 is: option d

3. The Answer for question no 3 is: option d

4. The Answer for question no 4 is: option a

5. The Answer for question no 5 is: option b

6. The Answer for question no 6 is: option d

7. The Answer for question no 7 is: option a

8. The Answer for question no 8 is: option b

9. The Answer for question no 9 is: option a

10. The Answer for question no 10 is: option a

11.The Answer for question no 11 is: option d

12. The Answer for question no 12 is: option a

|  |  |
| --- | --- |
|  | 13) Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data. |
|  | Traditionally, the claculation of cluster analysis consists of three basic steps: |
|  | a) Calculate the distances, |
|  | b) Link the clusters, and |
|  | c) Choose a solution by selecting the right number of clusters. |
|  |  |
|  |  |
|  | 14) |
|  | There are a number of metrics used to quantify the quality of clustering, but one needs to determine which one is more suitable in application. Having said that, the choice of metric also depends on what one considers the purpose of clustering to be. Personally, I think clustering ought to be about identifying different groups of observations that were each generated by a different data generating process. So I would test the quality of a clustering by generating data from known data generating processes and then calculate how often patterns are misclassified by the clustering. Of course this involves making assumptions about the distribution of patterns from each generating process, but one can use datasets designed for supervised classification. |
|  |  |
|  | Others view clustering as attempting to group together points with similar attribute values, in which case measures such as SSE etc are applicable. However, I find this definition of clustering rather unsatisfactory as it only tells you something about the particular sample of data, rather than something generalisable about the underlying distributions. |
|  |  |
|  | 15) |
|  | Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. |
|  | a) Centroid Clustering |
|  | This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you’re a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products. |
|  | The algorithm will start by randomly selecting centroids (cluster centres) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iterations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats. |
|  |  |
|  | b) Density Clustering |
|  | Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified. |
|  |  |
|  | c)Distribution Clustering |
|  | Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data. |
|  |  |
|  | These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty. |
|  |  |
|  | Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to acluster. |
|  |  |
|  | d)Connectivity Clustering |
|  | Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger. |