Assignment 3

Q1. Discuss the advantages and disadvantages of having to specify the number of clusters that an algorithm needs to find. Describe different methods for finding the number of clusters and discuss the implications of the shape of the clusters.

Clustering is one of the data mining techniques. Clustering makes sure to capture variance in the dataset and therefore is a good technique. One step in this technique is to specify the number of clusters that the algorithm finds.

Advantages of specifying the number of clusters

- ⇨ Finding # of clusters for algorithm is easy to implement
- ⇨ Already specifying # of clusters works well with large datasets
- ⇨ Specifying this # also helps with adapting to new examples

Disadvantages of specifying the number of clusters

- ⇨ Choosing the number of clusters manually takes lot of computational power. It is like a trial-and-error method to find an optimal number of clusters value.
- ⇨ The number decided for clusters maybe dependent on the initial values of the dataset. For low value of this #, usually the algorithm can be run a few times with different results for better results.
- ⇨ For dataset with varying sizes and density, finding # of clusters is difficult. This is why clustering is not good for dataset with varying size and density.

Determining optimal number of clusters is important for algorithms to perform well and provide correct outputs.

Methods of finding number of clusters

1. Elbow Method: After trial and error of a number of values, the values are squared and added and graphed. The graph is then searched for a change in slope – which creates a shape like an elbow – this determines the optimal number of clusters.

2. Gap Statistic: This method for different values of 'k', compares it with the expected values under null reference distribution of data. The optimal # of clusters is then the one that maximize the gap statistic.

3. Silhouette method: the optimal k value is the one that gives the maximum average silhouette for a range of probable values of k.

4. Sum of square method: Minimize how tight a cluster is and maximize how far a cluster is from another cluster. This helps find the optimal number of k.

5. There are also a few methods that uses R packages like NbClust and Clustree to find optimal number of clusters by considering changes in sample groupings with change in number of clusters. It does not give the right optimal number of clusters, but it does tell probable values.

Implications of shape of clusters

We can reach a certain conclusions/implications depending on the shape of clusters. The shape of the clusters contribute to choosing the right number to clusters to split the data in. Usually having a ball-form cluster helps in finding an optimal value for k. However, there are various conditions in which this gets affected.

- ⇨ Clusters should be enclosed in an equal radius balls centred at their respective cluster centers and there should be no gaps between them. This preserves their local minimum, and the algorithm is able to recognize these ball shaped clusters easily.
- ⇨ There should be good separation between the clusters as this ensures that once clusters are hit by one seed each, they never loose their cluster centers.

References:

https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92

Q2. Research semi-supervised methods and describe in your own words how supervised and
unsupervised methods can be used together. Describe and compare at least two techniques (400 to 500 words). Include the references.

Semi supervised Learning (SSL) in Machine Learning was developed to get over the disadvantages of supervised and unsupervised learning of having limited application spectrum. As the name suggests it is both supervised and unsupervised. Semi supervised learning algorithm uses both labelled and unlabelled data. There is more unlabelled data and less labelled data. We need SSL because supervised learning can only produce insufficient data as field sampling is expensive and time consuming.

The method of procedure is gathering similar data using unsupervised learning then using the labelled data to label that remaining unlabelled data. This ensures we do not have too much training data and that the data is learnt and labelled itself. Labelling data is an expensive procedure. The following characteristics are assumed for a data :

- ⇨ Points closer to each other are assumed to be continuous and have same output label
- ⇨ Points in same cluster will likely have same output label
- ⇨ There is an assumption of manifold

Sometimes in a dataset, we do not have all data labelled and thus we need both supervised and unsupervised leaning to be used together. First the known/labelled data is used as a training dataset and then the model built from it is then used to train more data that is unlabelled.

There can be different techniques in semi supervised learning,

1. Graph-based SSL: One techniques of SSL is graph-based learning where the data is represented as a graph. It is done in two steps – Graph construction and then label interference. The graph is created using the entire data provided – both labelled and unlabelled. Once that is done, then the label inference is performed using which the unlabelled data are labelled by incorporating labelled samples from the constructed graph in previous step.
2. Self-training SSL: This technique uses not only provided labelled data but also environment affecting the data into consideration, it is a wrapper method. First supervised learning is used to train labeled data and then classifier is applied on unlabelled data to create more labelled data by using the previous output as a training data, hence self training itself.

There are many applications of the SSL where different techniques of it is used. For example, in speech analysis – it is a tiring process if speech labelling is done manually, hence SSL is used in this case. Similarly text classification also uses SSL where the final goal is to predict the class label of a given piece of text.

References :

https://www.geeksforgeeks.org/ml-semi-supervised-learning/

https://journals.sagepub.com/doi/pdf/10.1177/2472555220919345

https://openaccess.thecvf.com/content_ICCV_2019/papers/Zhai_S4L_Self-Supervised_Semi-Supervised_Learning_ICCV_2019_paper.pdf

Q3 (next page)

Q3. Describe (in your own words, roughly 300 words) one clustering technique that is not covered in the textbook or the videos that does not use a Euclidean distance measure. Include a reference to a publication that describes the method.

There are a lot of clustering techniques that consider a Euclidean space, and many that do not consider a Euclidean space.

Adaptive Affinity Propagation Clustering

Adaptive Affinity Propagation Clustering (AP) is an algorithm that also works with clusters and data points. In cases with large data, when there are large number of clusters, AP is a good algorithm in case of speed, applicability and performance. One characteristic of this algorithm is that it works simultaneously with all data points and consider all data points as exemplars. The way it works is, for every data point, k, that is considered an ideal candidate to be an exemplar for a point, i, evidence is collected. The evidence 'responsibility' R(i, k) is collected from I to see how suited k is for i. Then evidence 'availability' A(i, k) is collected. This availability depicts how good it would be for point k to serve as an exemplar for point i. This way probable exemplars for a cluster are found. The more the value of R(i, k) + A(i, k), the more chance for that data point k to be the exemplar for the cluster. This is an iterative process of selecting m exemplars for m clusters. They start with n exemplars that then change to fewer and fewer till there are m exemplars found. This is the clustering solution of AP.

There are two parameters in AP - Preferences (p) and damping factor (lam). Preference parameter has initial negative value that indicates that the data point i can be chosen as a cluster center. p influences the output clusters and number of clusters – this means it influences how many exemplars will finally be the main cluster centers. There is no direct correlation between p and output clusters and hence the solution is always an optimal solution for AP. The second parameter, damping factor has two functions. First, in every iterative step, there is an update to the value of R and A,
$R_i = (1\text{-}lam)\times R_i + lam\times R_{i\text{-}1}$, $A_i = (1\text{-}lam)\times A_i + lam\times A_{i\text{-}1}$, where damping factor $lam \in [0,1]$ and default lam=0.5. The second function that damping factor follows is improving the convergence if AP fails to converge upon oscillations. Upon increasing the damping factor, oscillations get eliminated. When there are un-convergent cases, lam is increased manually and then gradually the AP algorithm upon re-running a few times converges.

This Adaptive AP adjusts the damping factor to eliminate oscillations by decreasing p when damping fails.

The publication discusses and goes into more depth of adaptive affinity propagation followed by one experimental case.

Reference: https://arxiv.org/ftp/arxiv/papers/0805/0805.1096.pdf
(400 words)