ASSIGNMENT-2

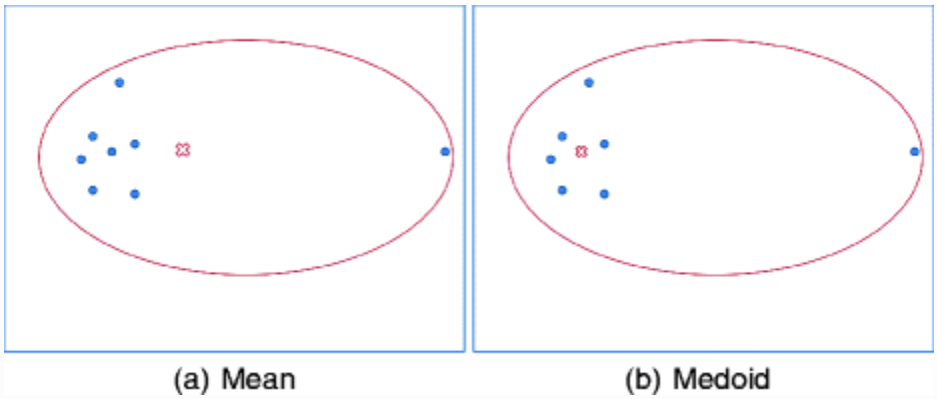
Machine Learning

1. **(b)**
2. (c)
3. (a)
4. (a)
5. (b)
6. (b)
7. (a)
8. (d)
9. (a)
10. (d)
11. (d)

***12 .The K*-means** clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. *K*-Medoids clustering is a variant of *K*-means that is more robust to noises and outliers. Instead of using the mean point as the centre of a cluster, *K*-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points.
Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster centre, while

medoid is robust to the outlier and correctly represents the cluster centre.

K-Medoids Clustering. Figure 1



- Mean vs. medoid in 2-D space. In both figures (a) and (b), the group of points in the right form a cluster and the rightmost point is an outlier. The red point represents the centre found by mean or medoid.

13. k-means is better because it is:

- **Relatively simple to implement.**
- **Scales to large data sets.**
- **Guarantees convergence.**
- **Can warm-start the positions of centroids.**
- **Easily adapts to new examples.**
- **Generalizes to clusters of different shapes and sizes, such as elliptical clusters.**

K-means Generalization

What happens when clusters are of different densities and sizes? Compare the intuitive clusters on the left side with the clusters actually found by k-means on the right side. The comparison shows how k-means can stumble on certain datasets.

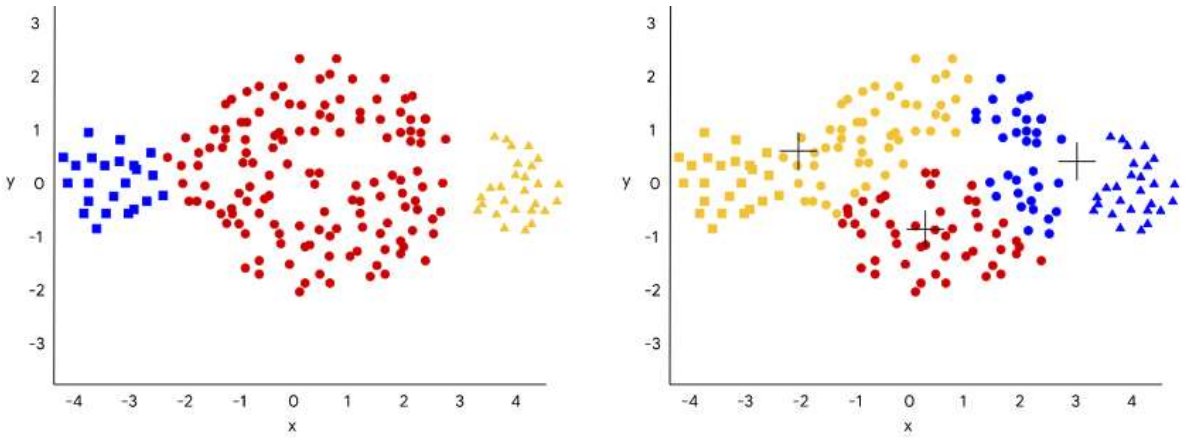
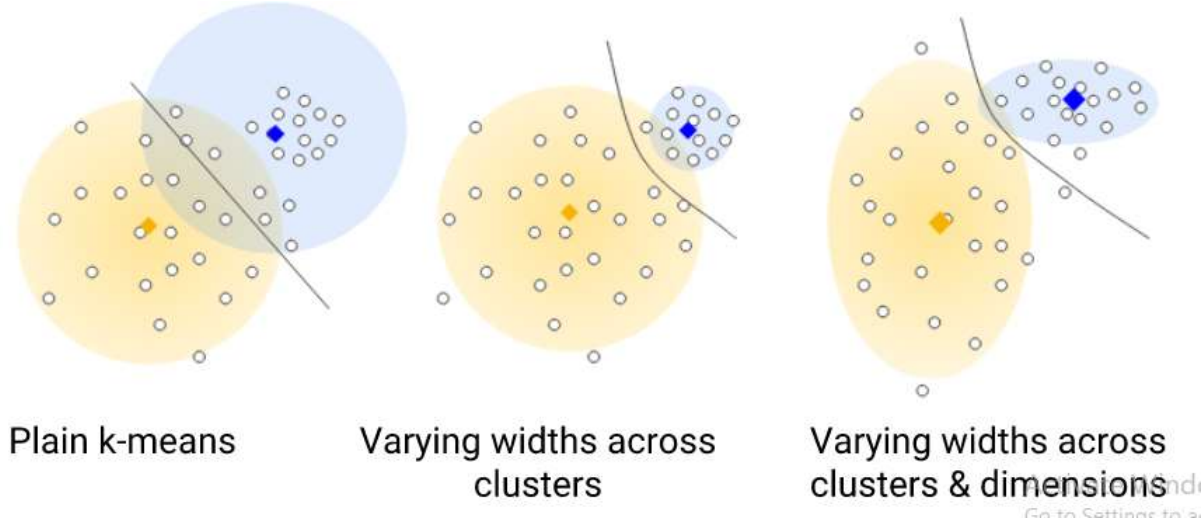


Figure 1: Ungeneralised k-means example.

To cluster naturally imbalanced clusters like the ones shown in Figure 1 , you can adapt (generalize) k-means. In Figure 2, the lines show the cluster boundaries after generalizing k-means as:

- Left plot: No generalization, resulting in a non-intuitive cluster boundary.
- Centre plot: Allow different cluster widths, resulting in more intuitive clusters of different sizes.
- Right plot: Besides different cluster widths, allow different widths per dimension, resulting in elliptical instead of spherical clusters, improving the result.



14. One of the significant drawbacks of K-Means is its **non-deterministic nature**. K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters. Besides, each run of the algorithm for the same dataset may yield a different output