Clustering Techniques: A comparison

Deepthi Hegde, Mathieu Guillaume-Bert, Kyle Miller, Artur Dubrawski

Introduction

- Clustering is the process of automatically finding groups of similar objects in data.
- Marginal clustering is discovering clusters by detecting suitable planes that separate groups of similar data called clusters.

We aim to compare and evaluate various marginal, distance based and density based clustering methods.

Illustration of peaks

Peaks: Points of local maxima and minima
Mountain: Maximum peak
Valley: Minimum peak
Cut: a plane separating clusters

A deep valley between 2 tall peaks indicates a region of low density suitable for a cut.

Split Clustering

Split clustering aims at uncovering clusters in high dimensional data with structure in lower dimensions. Hyperplanes that separate the data into groups are found recursively. Valleys serve as points of cut in the data.

OptiGrid

OptiGrid is a grid partitioning clustering technique that recursively finds optimal cutting planes to find clusters effectively. The algorithm cuts the dataset into grids and recursively finds more cuts (if possible) within each of the dense grids. The cuts are identified based on the density curve of each dimension of the dataset independently. The top q cuts across all dimensions are chosen based on the density score.

Split vs OptiGrid: Noise comparison

Split clustering is robust to noise. The structure of detected clusters is not hampered by the presence of noise points in the data.

Versatility of Split Clustering

Split clustering can handle a wide range of datasets with different cluster shapes. While the 1D and 2D cuts can detect convex clusters, the kernel trick used in the algorithm works well with non convex clusters.

Conclusion

- Simple and easy to implement and interpret
- Robust to noise
- Very low computation time
- Few parameters
- Capable of handling randomly shaped clusters

References


Acknowledgements

Thanks to all the members of the Auton lab and RISS team for the great experience. Special thanks to Mrs. Sowmya Kamath and Prof. Ram Mohan Reddy for their support.