Unsupervised SPORF

1. What is Geodesic Learning? Estimating distance between pairs of points in a data corpus, but true distances cannot be achieved in real data applications.
2. Most of the existing algorithms to estimate geodesic distances are for supervised learning.
3. Urerf doesn’t need to compute geodesic distances b/w all pairs of points. It clusters data in sparse linear subspaces and sticks to local structures. Randomer forest help in filtering out noise.
4. Fast BIC -
5. Urerf also introduces geodesic precision which compares NN estimated by geodesic algorithms with set of true NN on manifold. Noise doesn’t interfere much with the Urerf.
6. Non manifold learning approaches – compute geodesic distances in the original manifold in high dimensions and use eucledian distances to compute weights which are not good in high dimensional manifolds.
7. Isomap requires more space and time and doesn’t perform well on noisy data.
8. Distinctions of Urerf –
9. New splitting criterion – Fast-BIC computes approx. Bayesian Information Criterion for Gaussian Mixture model in one dimension
10. Uses random sparse linear combinations of features
11. Generate proximity matrices

Algorithm Explained

1. Overall Algorithm: Input data set of N, build T decision trees from sample size of m (<N), generates d features evaluated on the basis of newly introduced splitiing criterion and then generate proximity matrix
2. Split points are not in original feature space but in random projection framework.
3. Feature generation in nodes: The feature space is transformed into a smaller d dimensional feature space where the d features are a sparse linear combinations of the higher p dimensional features. A is a sparse projection distribution matrix generated by random sampling from (-1,1). A basically transforms the node data X to Xbar where each row of Xbar represents projection of the node data in the 1d space as a sparse linear combination of original feature space.
4. Splitting: each row of the above Xbar is inspected for best split point and split dimension which minimizes the splitting criterion.
5. Splitting criterion: Fast exact univariate splitting
6. Each decision tree operates in one dimension easier and faster soln available for two means
7. Split b/w sequential pairs of points, left one cluster and right one cluster, find mean by MLE, find cutpoint which minimizes the 2 mean objective.
8. Drawback – fails to consider feature vise variance. (Rescaling feature is an issue so not an option)
9. Splitting criterion: GMM Splitting with MClust BIC
10. For each feature data fit into 2 component GMM and Expectation Maximum is used determine parameters and latent variable (denote probability of the sample being in the cluster. Here two clusters
11. Likelihood is found for and each feature is evaluated with BIC for all parameters and latent variables and tree with max BCI is chosen for each node split.
12. Splitting criterion: 2 GMM with Fast BIC
13. Combines two means with BIC. First two means is done to sort and split(Above and below, 2 Gaussian clusters) each feature, then MLE is done to get prior means and variance for both clusters.
14. Fast BIC does hard clustering for two means (0/1 split instead of probability). Faster coz the equations simplify for two means and is tested only for single variance test.
15. Dimension and split point chosen to maximize log likelihood. Gives global MLE unlike previous BIC.
16. Proximity Matrix: Based on similarity matrix. Uses 0-1 loss to check if the pair of points is in the same leaf node or not. For unsupervised trees, proximity matrix is estimated by counting the fraction of times the points occur in the same leaf node in the forest.
17. Geodesic Precision & Recall: Correct neighbors defined by the lower dimensional manifolds. High precision and lower recall better estimates. Both continuous and discrete cases considered. In disconnected case, precision and recall are identical.

Numerical Results

1. Simulations: 4 cases, Linear – Euclidean gives geodesic with no noise; Helix – Similar to swiss jelly roll, latent manifold embedded in 3d space so Euclidean performs poorly but manifold algo perform well; Sphere – True manifold is 2d and could be extended to higher dimension s; GMM – difficult coz lack guarantee for disconnected connected component graphs.