Machine Learning

Clustering

Edgar F. Roman-Rangel.
edgar.roman@itam.mx

Digital Systems Department.
Instituto Tecnológico Autónomo de México, ITAM.

May 15$^{\text{th}}$, 2021.

Outline

Supervised learning

So far we have seen supervised learning, where,

- Pairs $\{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$ of input and output data.
- Goal: learn a model $\hat{y}^{(n)} = f(\mathbf{x}^{(n)}; \Omega)$.
- Such that $\mathcal{L}(y^{(n)}, \hat{y}^{(n)}) \approx 0$.

Unsupervised learning

Let's see now a few models for unsupervised learning.

- No labels $y^{(n)}$ for training, i.e., only $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$.
- We do not learn a mapping function.
- Rather, we try to make sense of $\{\mathbf{x}^{(n)}\}$.
- Discover hidden structures on data.
- Examples: clustering, dimensionality reduction.

Outline

Unsupervised learning

k-means clustering

Hierarchical clustering

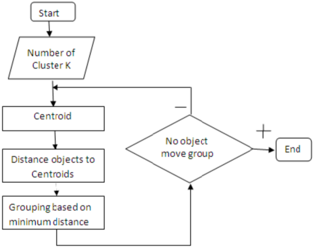
DBSCAN

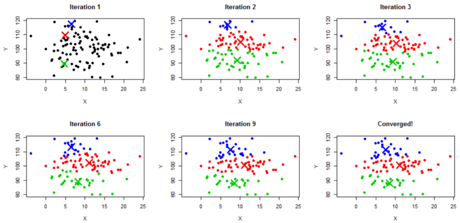
Intuition

- Find k groups (clusters) of similar objects
- Easy to understand and to implement.
- Prototype-based approach (each cluster is represented by a prototype sample, a.k.a., *centroid*).

Method

1. Randomly pick k centroids.
2. Assign each point to its closest centroid.
3. Move the centroids to the center of each cluster.
4. Repeat 2., and 3., until convergence.





Considerations

Distance metric

Euclidean is the most used.

$$d(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{N} (x_n - y_n)^2 .$$

Variants

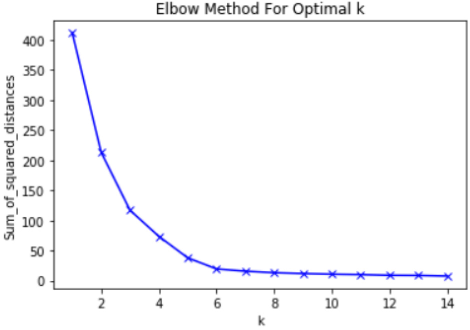
- k-medoids: Manhattan distance, median point as centroid.
- fuzzy C-means: soft assignment (probability distribution).

Validate quality

Given that there is no ground truth label y, it is difficult to tell whether the clustering algorithm is doing well. Use inspection.

Elbow curve

Try different values of k, then pick the one value where the *purity* of the clusters is no longer improved.



Silhouette score

Gives an idea of how tightly grouped the clusters are.

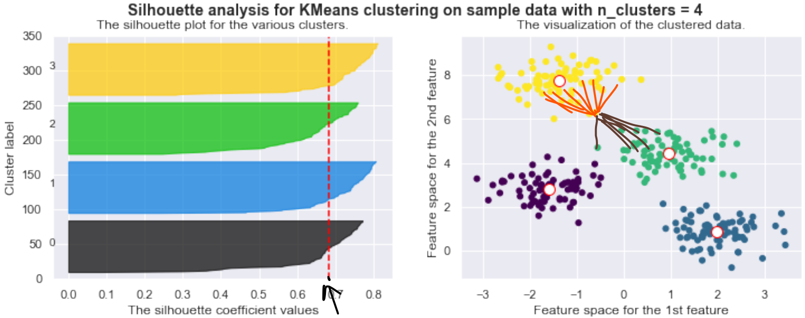
Per each sample $\mathbf{x}^{(i)}$, compute,

1. Cohesion $a^{(i)}$: average distance between the sample $\mathbf{x}^{(i)}$ and all other points in the same cluster.
2. Separation $b^{(i)}$: average distance between the sample $\mathbf{x}^{(i)}$ and all samples in the nearest cluster.
3. Silhouette $s^{(i)}$:
$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)}, a^{(i)}\}}.$$

s can take values between $[-1, 1]$. The higher the better.

Silhouette plot

Silhouette scores can be ploted for comparison.

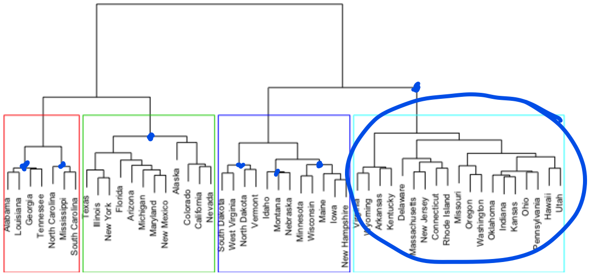


Note: both elbow and silhouette, can be used for most clustering methods.

Outline

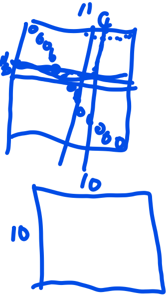
Intuition

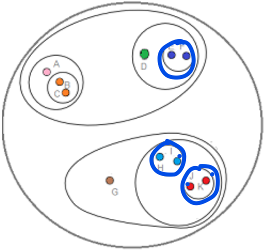
- Dendrogram (tree-like) visualization.
- No need to define the number of clusters a priori.
- We can selected by inspection.
- Can be agglomerative of divisive.

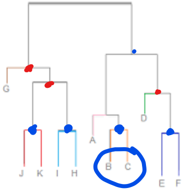


Method

1. Compute distance matrix for all samples.
2. Represent each point as a singleton cluster.
3. Merge the two closest clusters, based on a linkage strategy.
4. Update distance matrix.
5. Repeat 2., 3., 4., until only one single cluster is left.



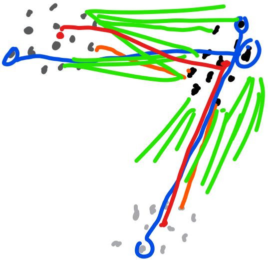




Linkage strategies

Distance between clusters can be defined as,

- **Single**: the distance between their closest members.
- **Complete**: the distance between their farthest members.
- **Average**: the average distance between each pair of members.
- **Ward**: the distance between their centroids.

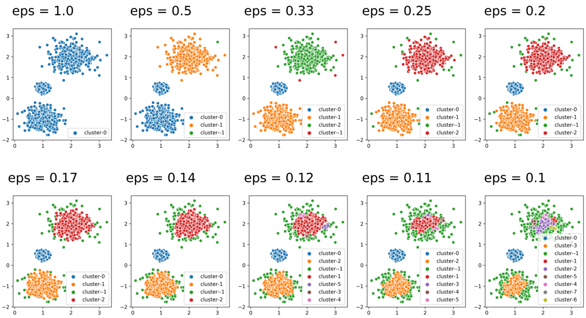


Outline

Intuition

Density-based Spatial Clustering of Applications with Noise (DBSCAN).

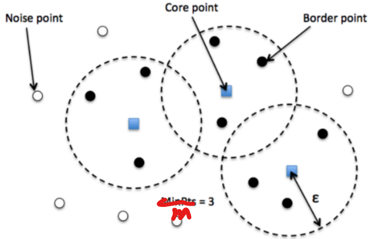
- Define how densely populated clusters must be.
- Density: number of points within a radius ε.



Method, I

Define,

1. **core points**: if at least m neighboring points fall within radius ε.
2. **border points**: if it has fewer neighboring points than m within radius ε, but lies within the radius of a core point.
3. **noise points**: all points that are neither core nor border points.



Method, II

After initial definition, continue with,

1. Form a separate cluster for each core point, or a connected group of core points (core points are connected if they are no farther away than ε).
2. Assign each border point to the cluster of its corresponding core point.
3. All non-assigned points end up marked as outliers.

Other methods

- Affinity propagation. ✔
- Spectral clustering. ✔
- Mean shift. ✔
- Fuzzy C-means. ✔

Q&A

Thank you!

edgar.roman@itam.mx