Lecture 6

Clustering 2

Topics

- DBSCAN clustering
- Using DBSCAN
- Hierarchical clustering
- Using Hierarchical clustering
- Evaluating clusterings
- Applications of clustering

K-means: result depends on initialization

Combining clustering and classification

Useful idea when labeling is expensive

- Take a dataset with handwritten digits
- Provide only one label per digit (10 labels for the whole dataset)
- Use 10-means with the ten labeled images as starting points for clustering the whole dataset.
- Then use 1nn for classifying new handwritten digits.

Limitations of K-means clustering

DBSCAN clustering

Steven Bierwagen

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- From 1996

Ingredients for DBSCAN

- A *distance* measure (or metric or similarity measure)
  - often Euclidean distance

scanning radius

- A number defining the meaning of *neighbor*
  - epsilon: the max distance between two points considered neighbors.

- A number defining the meaning of *cluster* (vs outlier or noise)
  - minpts: the minimum number of points in a cluster.

min points inside radius

Two hyperparameters

Labeling step

CORE

BORDER

NOISE

All points are labeled with one of these labels

Neighbors

A *neighbor* of a point p is a point that is within distance ϵ from p.
ϵ
p

Core points

minpts = 4
Core
A *core* point is a point that has at least minpts neighbors

Border points

Noise points

Clustering step

Clusters all core points and border points. Outliers will not be clustered!

Start by picking a new color c and an uncolored core point p.

Clustering step

Put an edge between core points that are neighbors. Color those connected components with c

Also color the border points of those nodes with c

Clustering step

Algorithm

Algorithm 8.4 DBSCAN algorithm.

```
Label all points as core, border, or noise points.
Eliminate noise points.
Put an edge between all core points that are within Eps of each other.
Make each group of connected core points into a separate cluster.
Assign each border point to one of the clusters of its associated core points.
```

Clusterings created by DBSCAN

Ester, Kriegel, Sander, Xu (1996), In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, (KDD), AAAI Press, pp. 226–231

Using DBSCAN

Open dbscan_intro (on Canvas)

K-means vs. DBSCAN

- K-means assigns all points to a cluster, whereas DBSCAN doesn’t necessarily do this. DBSCAN treats outliers as outliers.

- K-means works best when clusters are basically spherical. DBSCAN can find arbitrarily-shaped clusters.

- DBSCAN doesn’t require the number of clusters to be specified by the user.

Hierarchical clustering

Luis Serrano

A new set of points

Suppose we want to cluster these addresses by proximity. No pizza parlors involved this time!

Join the close ones

Join the close ones

Join the close ones

Join the close ones

Join the close ones

Join the close ones

Continue in the same way…

Join the close ones

Continue in the same way…

Join the close ones

Dendrogram

A *dendrogram* shows the entire hierarchichal clustering process (without STOP)

Dendrogram

1
2
3
4
5
6
7
8
Distance
If we want 4 clusters, we can cut here.
If we want 8 clusters, we can cut here.

Dendrogram

If we have a space with billions of points in thousands of dimensions, the dendrogram is still a 2D graph!

For example the tree of life!

Hierarchichal clustering gives more than a clustering: a hierarchy (or taxonomy)

The tree of life

https://www.evogeneao.com/

Hierarchical clustering

- Sometimes called agglomerative clustering, when done bottom-up

- From one extreme case (many clusters, each containing one item) to another (one cluster that contains all items)

Using hierarchical clustering

Open hierarchical_intro (on Canvas)

Clustering is successful, but difficult

- Inherent vagueness in the definition of a cluster
- Can be difficult to define an appropriate similarity measure

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651-666

https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Are the framed cases as desired?

Validating clustering

Stability on subsets

Clustering stable if removing a proportion of random points does not change the clustering fundamentally

Stability on subsets

Note colors change as labeling clusters into first, second, third … changes!

Co-occurrence

For all pairs (i,j) count how frequently i and j are in the same cluster.

Co-occurrence

Stability over repetitions

- Clustering stable if (almost) always same points end up in the same clusters togethers (co-occurrence frequencies) from random initializations

Silhouette coefficient

a: The mean distance between a sample and all other points in the same class.

b: The mean distance between a sample and all other points in the *next nearest cluster*.

$$s = \frac{b - a}{\max(a, b)}$$

https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Clustering clustering algorithms

Jian et al.. (2004) Landscape of Clustering Algorithms, Proc. 17th Int. Conf. on Pattern Recognition (ICPR'04)

Clustering clustering algorithms

Fahad et al. (2014) IIEEE Trans. Emerging Topics in Computing, volume 2, 267-279

Assignment 3

- Using K-means and density-based clustering to cluster the main chain conformations of amino acid residues in proteins.
- To understand the problem domain, look at:
  - http://bioinformatics.org/molvis/phipsi/
  - http://tinyurl.com/RamachandranPrincipleYouTube

Protein main chain

Ramachandran plot

Around 100000 data points shown here

http://bioinformatics.org/molvis/phipsi/

Chellapa and Rose (2012). *Protein Science*, **21**, 1231-1240

"the 11 most populated residue basins in a database of high-resolution protein structures"

Next time

Module 4

By Simon

Bayesian Statistic

Graphical Model

I'll see you again for

Module 8 (Rule-based AI)