

Lecture 25: Unsupervised learning

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 25 – p. 1/13

Questions?

CSI 5v93: Introduction to machine learning, Lecture 25 – p. 2/13

Unsupervised learning

Unsupervised learning is learning without a training signal.

It can be very different than supervised learning

- no “correct answer”
- goal is different – the concept to learn is not given with the training data
- overfitting and underfitting are still possible, but it’s not as clear when they occur

Data clustering (4.3)

Several views/motivations:

- identify and group items which are similar into several groups
- identify and divide items which are different into several groups
- summarize the data with several prototypes

There are several types of clustering algorithms:

- hierarchical bottom-up
- hierarchical top-down
- iterative “flat” clustering (e.g. k -means and Gaussian EM)
- spectral clustering
- density-seeking (e.g. mean-shift)

Most clustering algorithms are iterative in some fashion.

Hierarchical clustering

k -means finds a fixed number of clusters, based on iterative optimization.

The k -means clusters form a segmentation of the clustered points.

New idea: build a *hierarchy* of clusters.

- form different numbers of clusters
- clusters have a hierarchical relationship

Bottom-up hierarchical clustering

Also called merge-based hierarchical clustering.

Central idea:

- start: every input point is a cluster
- find two closest clusters
- merge these two clusters
- repeat these two steps until only one cluster is left

How to merge two clusters: put the points from each cluster into one new cluster.

Identifying two “closest” clusters

Several typical ways to identify closest clusters:

- furthest pair of points (aka single linkage)
- average distance between all pairs of points (aka average linkage)
- closest pair of points (aka complete linkage)

All three have their biases!

Using distances

For merge-based clustering, we don't need to have the points living in a metric space.

All we need is an $n \times n$ matrix (positive semi-definite) of their distances, for example:

$$D = \begin{bmatrix} 2 & 6 & 2 & 6 & 2 \\ 6 & 5 & 5 & 4 & 6 \\ 2 & 5 & 6 & 7 & 8 \\ 6 & 4 & 7 & 6 & 5 \\ 2 & 6 & 8 & 5 & 5 \end{bmatrix}$$

So as long as we can measure the distance between all pairs of points, we can cluster them with a hierarchical merging algorithm.

However, if we have the data in a metric space, finding closest clusters can be implemented more efficiently.

Identifying two “closest” clusters

Here c and d represent two distinct clusters, and i and j represent points in those clusters.

Closest pair:

$$\min_{c,d} \min_{i \in c, j \in d} D(i, j)$$

Average distance:

$$\min_{c,d} \frac{1}{|c||d|} \sum_{i \in c} \sum_{j \in d} D(i, j)$$

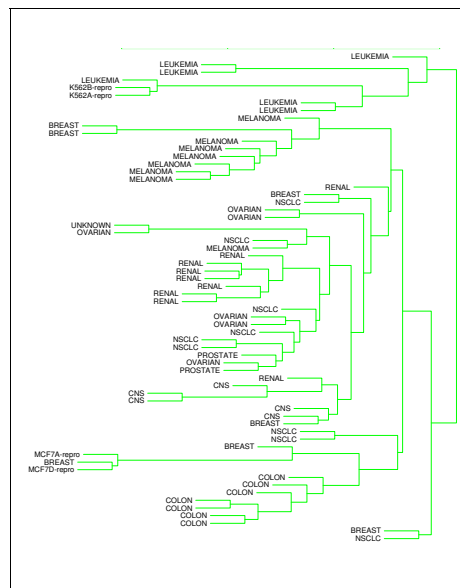
Furthest pair:

$$\min_{c,d} \max_{i \in c, j \in d} D(i, j)$$

Deciding on where to cut the tree

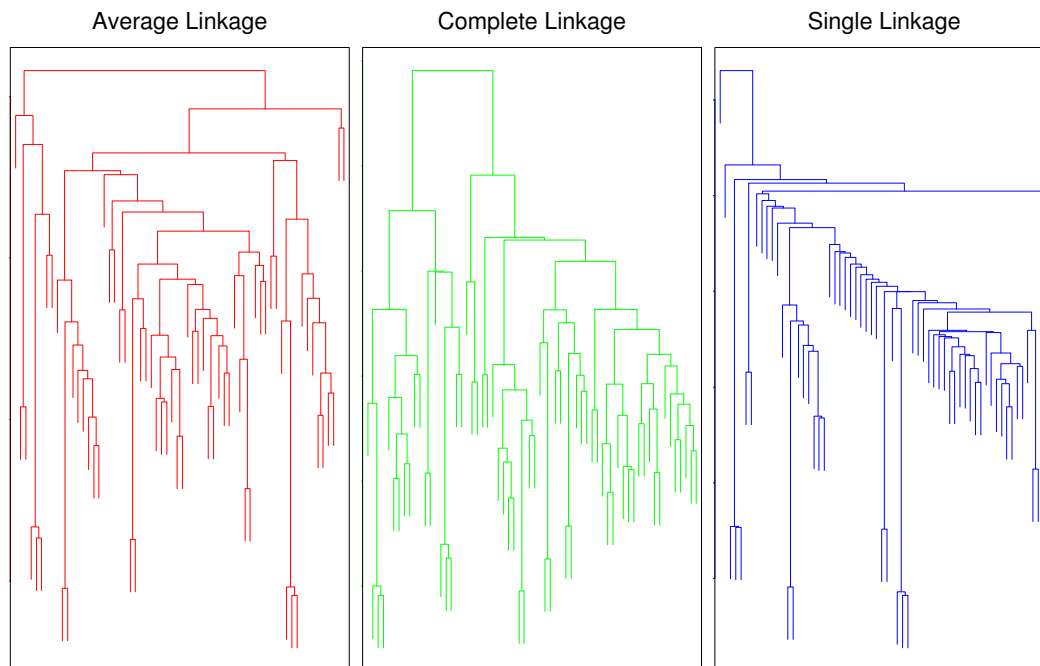
Once the clustering algorithm has run, we have n clusterings of n datapoints.

We can represent these clusterings all at once is with a *dendrogram*:



If you want one clustering, slice the tree at a certain level in the hierarchy.

Choosing the right merging method



CSI 5v93: Introduction to machine learning, Lecture 25 – p. 11/13

Observations about these methods

If the data points are actually well-separated into clusters, all three methods will produce similar results.

Despite the appeal of these common hierarchical methods, they are not theoretically grounded – the clusterings can be arbitrarily bad.

Intermediate decisions are not revisited; hierarchy is not flexible.

More advanced, theoretical methods have recently advanced hierarchical clusterings to address these problems.

Runtime: $O(n^2d)$

CSI 5v93: Introduction to machine learning, Lecture 25 – p. 12/13

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification