Supervised vs. unsupervised learning again

In *supervised learning*, we had a number of known examples to train our classifier (cf. Lady Gaga classifier)

Unsupervised learning: no training data, we try to discern patterns based on the data itself.

Clustering is an example of unsupervised learning.

Clustering

Two main types of clustering algorithm: *hierarchial* and *partitional* clustering.

- Hierarchial Progressively merge smaller clusters into bigger ones. The lifecycle is recorded in a *dendrogram*.
- Partitional Divide the data into disjoint clusters. k means is a type of partitional clustering. Randomly assign k nodes to k clusters and join the nearest node to the cluster.

For email finding, examined both hierarchial and partitional clustering; knowing the value of k is a key limitation of k means, as well as being non deterministic.

Similarity metric

The purpose of a similarity metric is to find out which nodes are closest together, and thus eligible to be clustered together.

We define a similarity metric S as having the four properties of:

symmetry	$S(x_i, x_j) = S(x_j, x_i)$
positivity	$0 \leq S(x_i, x_j) \leq 1$ for all x_i and x_j
reflexivity	$S(x_i, x_j) = 1$ iff $x_i = x_j$
triangle inequality	$S(x_i, x_j)S(x_j, x_k) \leq [S(x_i, x_j) + S(x_j, x_k)]S(x_i, x_k)$

A similarity metric is a distance metric, inverted.

Hierarchial agglomerative clustering

Demo of clustering in cartesian space.

Hierarchial agglomerative clustering

Demo of clustering in cartesian space.

Method of clustering:

1. Find the closest points and join them (ranking in order)
   Distance between each point : $\sqrt{x_2 - x_1)^2 + (y_2 + y_1)^2}$
2. A cluster's midpoint (*centroid*) is the mean of each constituent point
3. Repeat until everything is joined or distance threshold is exceeded.