M/L Commando Course 2018

Session 1b - Clustering

Russell Moore

ALTA / Computer Laboratory

July 2018

Introduction - Clustering

Clustering is an *unsupervised learning* technique. We'll look at k-means clustering, which is the absolute classic clustering technique[1].

Clustering is cool because it allows you to find patterns that aren't immediately obvious from vast tables of data. The clustering algo gives you a sort of clairvoyance, allowing you to see through the barrage of numbers.

[1] Other clustering algorithms are available!

K-means clustering

- Say we have a set of samples $S = \{\vec{x_i}\}$
- Each sample has m *features*, e.g. in our iris data each $\vec{x_i}$ has $m = 4$ because we have four different characteristics recorded for each flower. (We can equally think of the sample as a m-dimensional vector.)

K-means clustering: algo overview

Pick a number of clusters, k;
Set 'first estimate' of cluster *centroids*;[2]

Repeat {

    Assign each sample to a cluster based on its 'nearest' centroid;
    For each cluster, reposition the centroid to be the mean of the cluster members (hence 'k means'...);

}
Until the means do not change any more;[3]

[2] There are several ways to optimise this assignment: 'kmeans++' is default in scikit-learn

[3] Or they don't change much, or N iterations have been done, or some other target...

Number of clusters

- There's no 'best way' to choose the number of clusters!
- Sometimes, as with the iris data, we know the number of variants we seek and this guides our choice.
- Otherwise can also do *primary component analysis* to find the most informative dimensions and draw a *scree plot* to see how many we need to consider to explain most variations in the data. We won't do that here though as I'll explain it in a different lecture.
- Lastly, we can just use our judgement to pick a sensible number. This might be based on some exploratory work on the data in question, e.g. visualising the clusters before guessing their number.

Alterative clustering methods

- You may have reason to use other clustering algorithms. Sklearn has lots of options.
- You can set the metric that determines the "distance" between points. It's natural to think of Euclidean distances, but that needn't be the case.
- If you don't know how many clusters you'll have, *affinity propagation* is a modern (2015) clustering algorithm that attempts to estimate the number of clusters. However it has a "temperature" setting, so it doesn't do everything for you...