Clustering

December 14, 2016

Abstract

This document introduces some fundamental notions of Clustering.

0.1 Clustering (Unsupdervised)

The goal of clustering is to automatically group similar samples into sets.

Since a clustering algorithm has no prior knowledge of how the sets should be defined, and furthermore, since the clustering process is unsupervised, the clustering algorithm needs to have a way to tell which samples are the most similar, so it can group them. It does this the same way we humans do: by looking at the various characteristics and features of the sample.

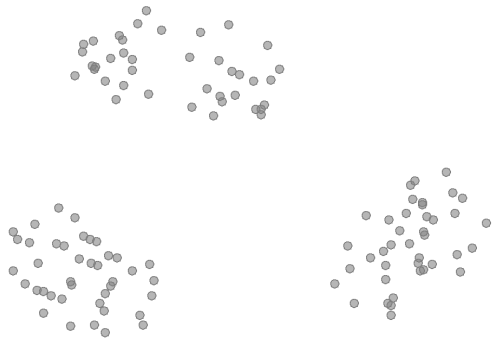

Figure 1: Example of linear regression.

More examples

- Match similar people on a matrimonial site based on their profile question answers.

- Based on search history, recommend houses a prospective home-buyer might be interested in considering.
- Pinpoint the most likely location for a future earthquake using past earthquake seismic data.
- Identify new characteristics shared by different people suffering from the same disease.Calculate an equation to predict the size of a house given its price; or the price of a house given its size.

There are different types of clustering algorithms, some supervised, some unsupervised. There are even semi-supervised clustering methods as well. In this course, you'll only be dealing only with unsupervised clustering. In other words, the clustering algorithm you will use won't need anything except your raw data. No labels hinting at desired clustering outcome will be provided to the algorithm.

source: course.edx.org