

EXPLORE || DIGITAL SKILLS

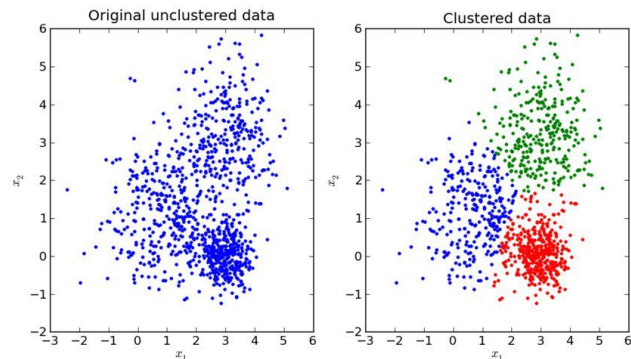
Unsupervised Learning Overview

Introduction to Unsupervised Learning

SUPERVISED LEARNING is what we've been engaged with up to now: we have **labelled** data (house prices, personality types) and we use our other (feature) data to try and predict these outcomes as closely as possible, in a way that generalises well to future emerging data.

In **UNSUPERVISED LEARNING**, we don't have those outcomes: we are dealing with **unlabelled** data and our job is to try to extract meaning from it. So we're looking to detect patterns, structures and/or insights from the data..

Unsupervised Learning



Unsupervised Learning applications

CLUSTERING

- Think about the way Netflix recommends which movies you should watch next based on what you've seen up to this point. It is able to use your previous viewing experiences as a guide to put you into a "group" of Netflix viewers. Based on that information it can infer future viewing patterns simply using enough data of this type.

ANOMALY DETECTION

- In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

DIMENSIONALITY REDUCTION

- A lot of the data we would like to use for machine learning is often high-dimensional. An image can have millions of pixels, and if we needed to run an algorithm through all of them, this creates a computational nightmare. So dimensionality reduction helps us retain the most important information using the least amount of information (think about the main features of an image such as the outer lines of a person's face, instead of each pixel).

What is expected of you?

Look out for the following upcoming materials over the next four weeks

	Pre-processing	Train	Test
Dimensionality Reduction	ISLR Chapter Blogs Videos	PCA Advanced DR Geospatial Analysis	Dimensionality Reduction
Clustering	Blogs Videos	K-means Hierarchical GMMs	Clustering
Recommender Systems	Blogs Videos	Recommender Systems	
		Loading data into S3	

Dimensionality Reduction Techniques

A widely encountered problem in machine learning is that of dimensionality. The problems with increasing or high levels of dimensionality are:

- More storage space required for the data;
- More computation time required to work with the data; and
- More features means more chance of feature correlation, and hence feature redundancy.

The goal of dimensionality reduction is to reduce the number of features in a dataset while minimising the amount of data loss. Three methods will be covered this sprint:

Principal Component Analysis

- data can be mapped to some lower number of dimensions, whilst retaining the maximum amount of variance

Multidimensional Scaling

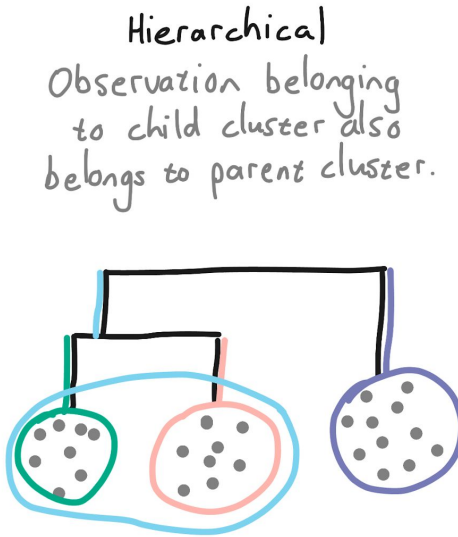
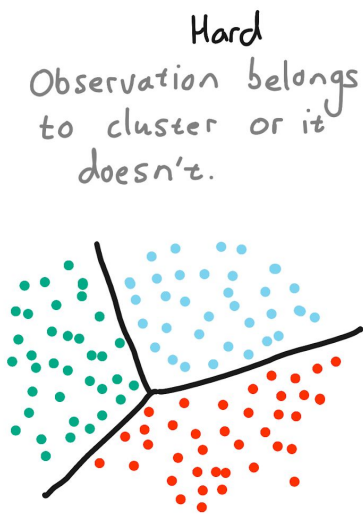
- map features to a low-dimensional space, while preserving the distances between observations

t-distributed Stochastic Neighbour Embedding

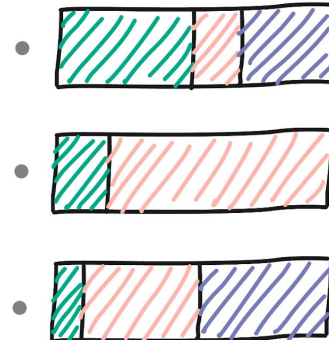
- Non-linear transformation which preserves both local and global structure, however it is computationally expensive

Clustering Algorithms

Clustering is the process of grouping similar data points together such that data points in the same groups are more similar to other data points in that group than those in other groups. The aim is to divide groups with similar characteristics and assign them to clusters. This will give us insights into the underlying patterns of the different groups/clusters. In this course, we will explore three types of clustering:







Soft/Fuzzy
Observation can belong to each cluster to varying degrees.



Recommender Systems

Utility Matrix

		Items			
					
Users	Bob	✓			✓
	Xolisa	✓	✓		
	Joanne			✓	✓
	Jon	✓		?	

Recommender systems are the unsung heroes of our modern technological world.

Search engines, online shopping, streaming multimedia platforms, news-feeds - all of these services depend on recommendation algorithms in order to provide users the content they want to interact with.

At a fundamental level, these systems operate using similarity. In **content-based filtering** this similarity is measured between items based on their properties, while **collaborative filtering** uses similarities amongst users to drive recommendations.

In this Sprint, we will implement and extend these algorithms ourselves - learning knowledge which, if mastered, will be extremely valuable to your career.