Cluster Analysis and other unsupervised learning methods

Introduction and motivation

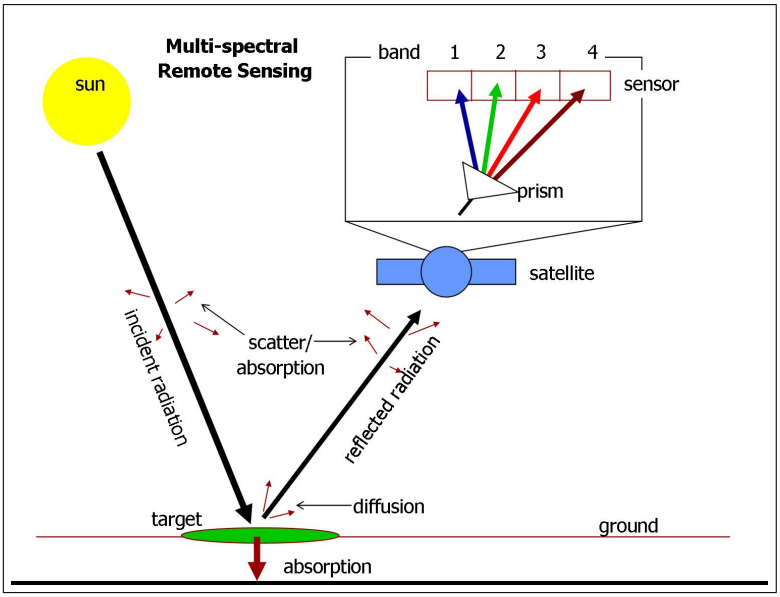
Werner Stuetzle
Department of Statistics
University of Washington

March 20, 2017

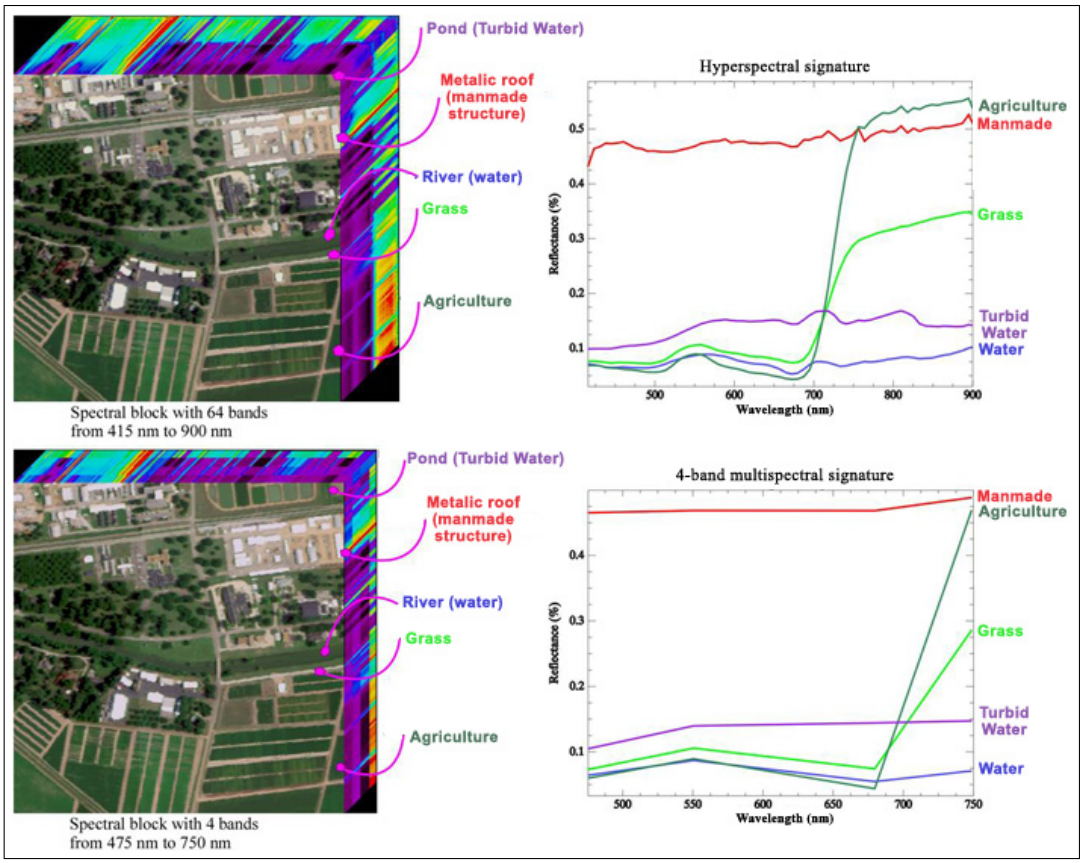
Unsupervised vs supervised vs learning

Example: Satellite image analysis

Given: Multispectral image of earth's surface with pixel spectra $\mathbf{x}_1, \ldots, \mathbf{x}_N$.



Vague Goal: Figure out what's on the ground.



Supervised learning problem:

Create rule that predicts the ground cover (forest, road, water, cornfield, ...) for the area corresponding to a pixel from the pixel's spectrum.

To make rule need training sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of pixels for which spectrum $\mathbf{x}_i$ and gound cover y_i are known.

Unsupervised learning problem:

Partition spectra $x_1, \ldots, x_N$ into groups such that spectra in each group are similar, and dissimilar from spectra in other groups.

Motivation: Could be helpful in selecting areas for which ground truth should be ascertained.

Example: Document analysis

Given: Collection of documents, mapped to vectors in some euclidean space (interesting problem all in itself).

Unsupervised learning problem:

Prepare digest or table of contents of the collection (maybe hierarchical).

Supervised learning problem:

Assign topic label to new document.

More generally:

In supervised learning we have a collection $\mathbf{X} = (X_1, \ldots, X_m)$ of predictor variables and a response variable Y.

We are given a training sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

Based on this training sample we want to

- Create a rule that predicts Y for new observations for which we know only $\mathbf{x}$
- Understand which predictors affect Y, and how
- Assess how well we do

For formally; we want to estimate (properties of) the conditional distribution $p(y \mid \mathbf{x})$. Marginal distribution $p(\mathbf{x})$ is of secondary interest.

In unsupervised learning we want to "understand" the distribution of a feature vector $\mathbf{X}$.

If features are Euclidean, can visualize observations as point cloud in m-dimensional space. May want to find "structure":

- Distinct groups
- Outliers
- Concentration near lower dimensional manifolds.
- All of the above

If feature are categorical $\Rightarrow$ maybe joint distribution is product of marginals, or there might be a few low order interactions.

Goal of unspervised learning is vague: *organize, categorize, summarize, require explanation.*

The term *unsupervised learning* comes form cognitive science. Cognitive science is interested in unsupervised learning because a lot of human learning is unsupervised (cats vs dogs, language).

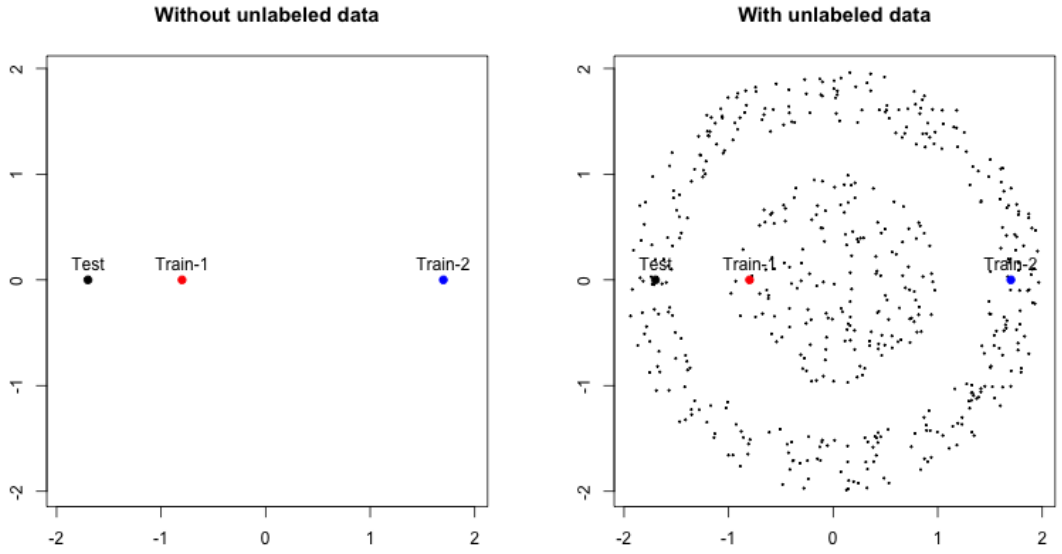
Semi-supervised learning

The goal of semi-supervised learning is the same as the goal of supervised learning. However, in semi-supervised learning we have labeled as well as unlabeled data.

Why might unlabeled data be useful when the goal is to predict labels?

Left plot: Two labeled training observations. The `Test` observation would be labeled "red" because it is more similar (closer to) `Train-1` than two `Train-2`.

Right plot: The cluster structure of the unlabeled observations suggests that the `Test` observation is more similar to `Train-2` and should be labeled "blue".



Semi-supervised learning is becoming increasingly important because it is much easier to acquire unlabeled data. Labeling data is often expensive.

Other examples for unsupervised learning problems / methods

Numerical Taxonomy

Biology: Organize life forms

Medicine: Organize diseases

Archaeology: Organize finds, like pottery shards

Linguistics: Organize languages

Market basket analysis

For each customer visit to store record set of items that customer purchased.

Identify sets of items that are more likely to co-occur than under independence.

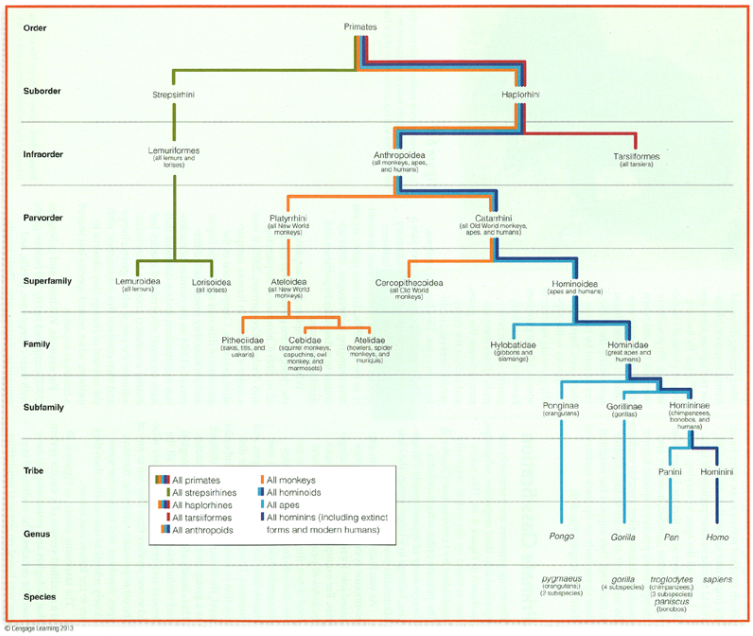
Motivation: Organize store so that frequently co-purchased items are close together (or not?)

Detecting communities in social networks

Nodes represent individuals

(Weighted) edges represent connections

Question: Are there communities — sets of nodes for which connections within are stronger than connections to the outside?



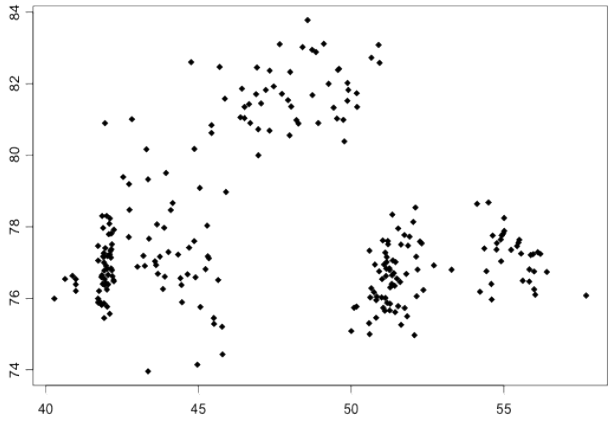
What is Clustering ?

Given:

Collection of n objects, characterized by feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, or adissimilarity matrix $d(i,j), 1 \leq i,j \leq n$, or a similarity matrix.

Goal:

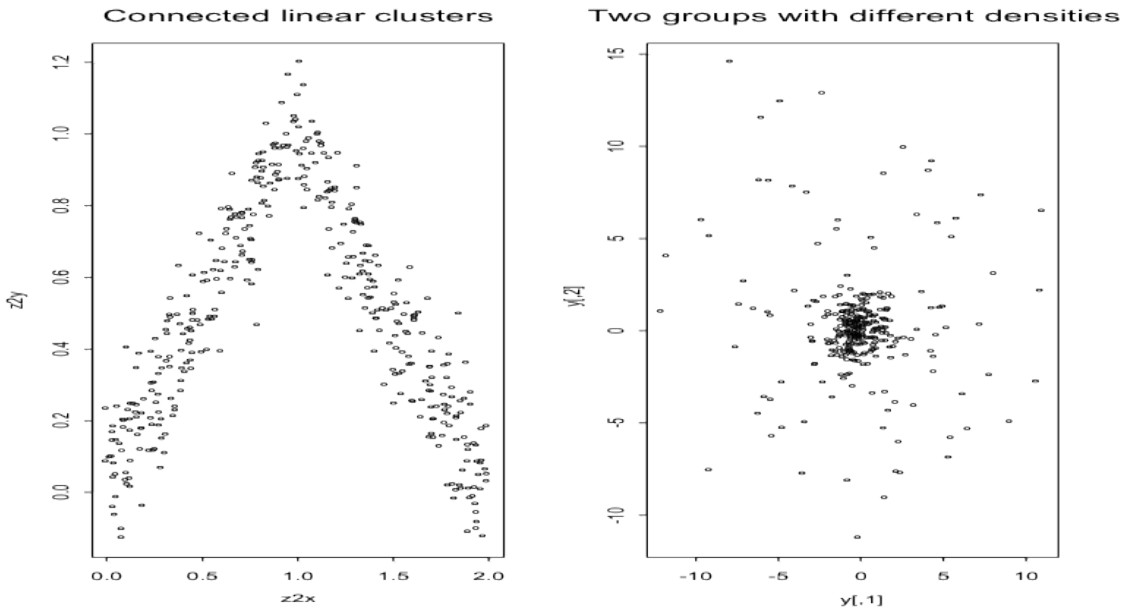
Detect presence of distinct groups, and assign objects to groups.



Common definition of "groups":

Contiguous, densely populated areas of feature space, separated by contiguous, relatively empty regions.

Definition does not capture all meanings of "group"

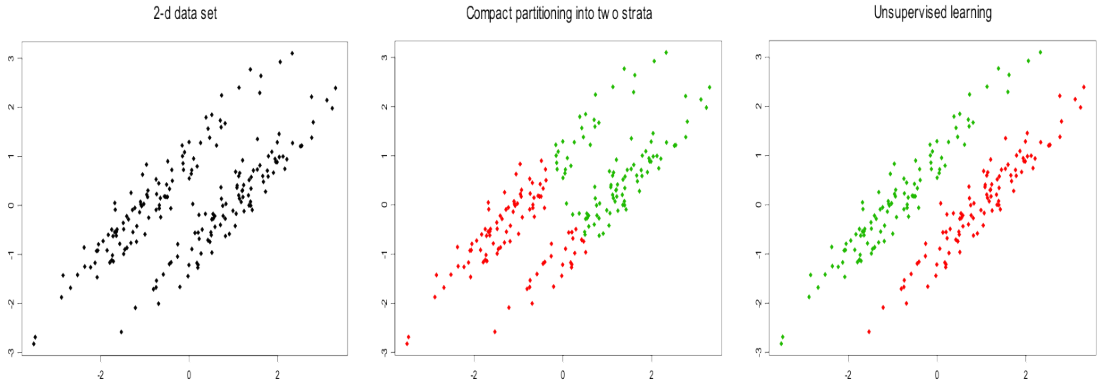


No automatic method can detect all kinds of "structure".

Important to distinguish between

Clustering: Detect presence of distinct groups;

Dissection: Partition collection of objects into compact strata.



Disection can make sense even if there are no distinct groups.

Example: Vector quantization.

Approaches to clustering

Statistical approach:

- Regard observed feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as a sample from some underlying distribution $p(\mathbf{x})$.
- Estimate properties of $p(\mathbf{x})$ indicating presence of groups, or lack thereof.

Ad-hoc approach: (for lack of a better word)

- Design algorithm based on heuristics

Advantages of statistical approach:

- Can compare methods: How well are we estimating whatever we are trying to estimate?
- Can reasonably ask how many groups there "really" are.

Variants of statistical approach

Parametric (model-based) clustering:

- Based on premise that each group g is represented by distribution $p_g(\mathbf{x})$ that is a member of some parametric $\Rightarrow$ $p(\mathbf{x})$ is a mixture

$$p(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \, p_g(\mathbf{x}) \, .$$

- Estimate the parameters of the group densities p_g, the mixing proportions π_g, and the number of groups G from sample.
- Assign observations to groups using Bayes' rule.

Nonparametric clustering:

- Based on premise that groups manifest themselves as multiple modes of $p(\mathbf{x})$.
- Estimate modes from sample.
- Partition feature space into "domains of attraction" of modes.