ML Week

Clustering and Anomalies

Jeff Abrahamson

23–24 novembre 2016

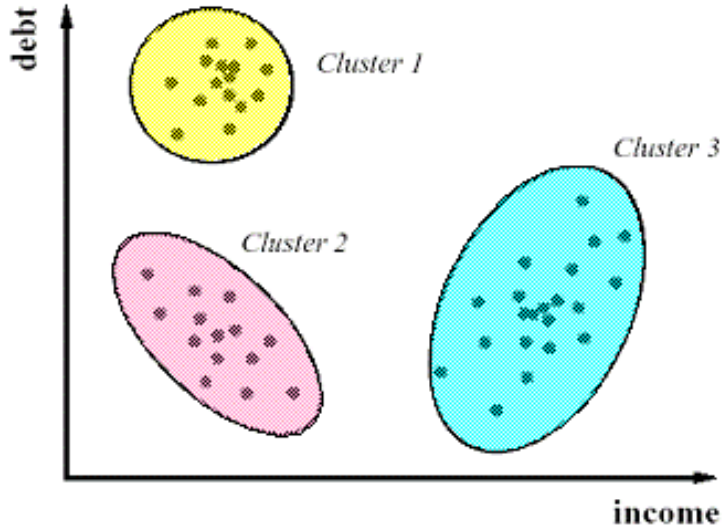
Clustering

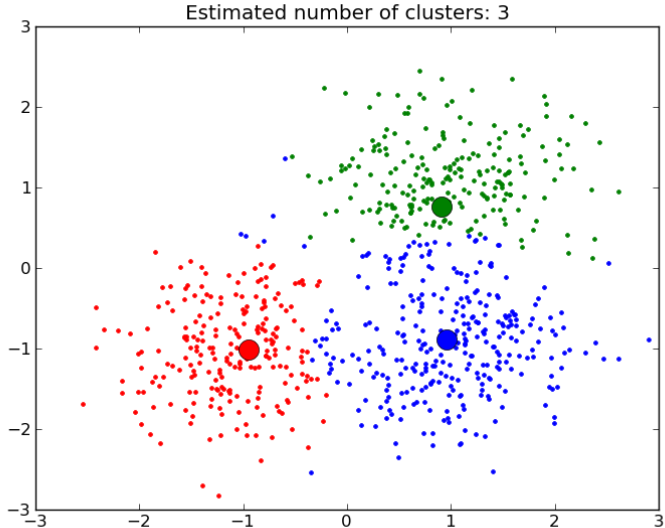
The Problem

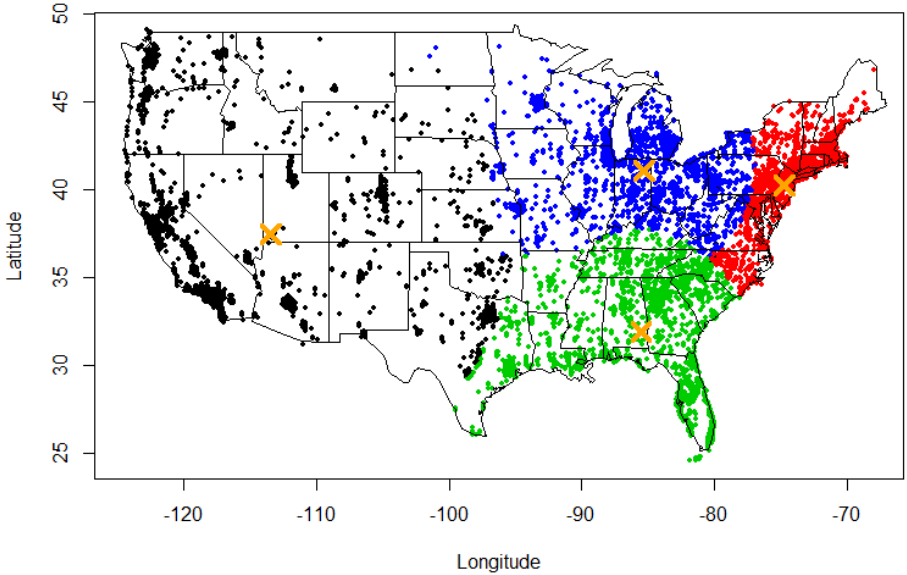
Have points $d = \{d_1, \ldots, d_n\}$.

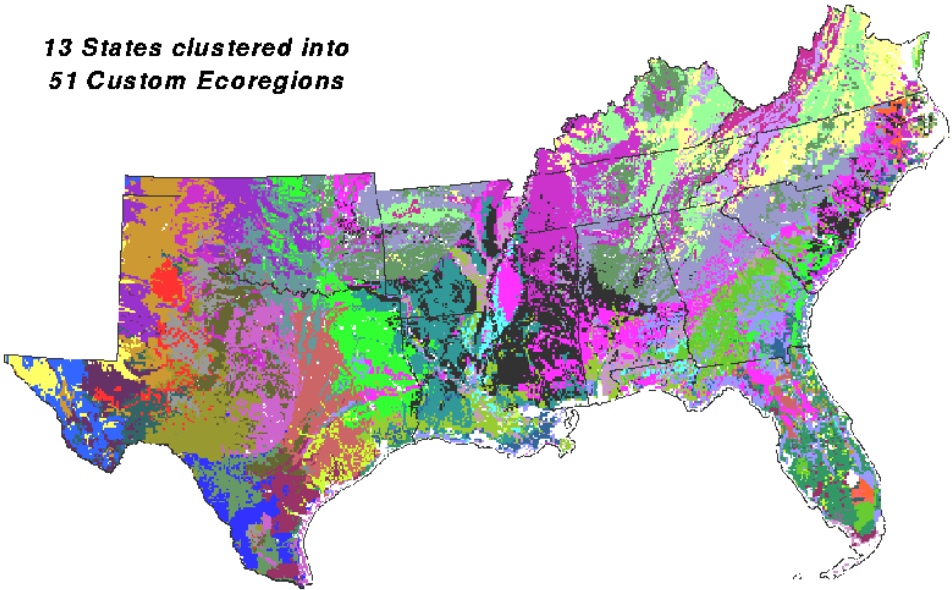
Have number of clusters k.

Want: an assignment of points to clusters









The Algorithm

1. Assign points to clusters at random
2. Repeat until stable:
    1. Compute centroids of each cluster
    2. Assign points to nearest centroid

Cost function

$$\text{cost} = \sum_i \sum_j |x_j - \mu_i|$$

Silhouette coefficient

Points $d = \{d_1, \ldots, d_n\}$

Clusters $K = \{k_1, \ldots, d_k\}$.

Cluster k_{d_i} is the cluster of d_i.

Silhouette coefficient

Points $d = \{d_1, \ldots, d_n\}$

Clusters $K = \{k_1, \ldots, d_k\}$.

Cluster k_{d_i} is the cluster of d_i.

Let a_i be the average dissimilarity of d_i to all points in its cluster.

Let b_i be the least average dissimilarity of d_i to any cluster other than k_{d_i}

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

$$s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

Silhouette coefficient

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

$$s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

So $s_i \in [-1, 1]$

Silhouette coefficient

s_i near 1 $\iff$ d_i well clustered

s_i near 0 $\iff$ d_i on the border between two clusters

s_i near -1 $\iff$ d_i well clustered

Silhouette coefficient

Consider $\overline{s_i}$ over $i \in k_j$ for cluster k_j

Silhouette coefficient

Consider $\overline{s_i}$

video time

Anomaly Detection

Introduction to Anomaly Detection

- Supervised
- Unsupervised

Introduction to Anomaly Detection

Supervised anomaly detection:

- Training data: normal, abnormal
- Train a classifier

So reduced to existing problem of supervised classification.

Introduction to Anomaly Detection

Unsupervised anomaly detection:

- Mostly, this is clustering
- Increasingly, this is neural networks in advanced applications

Introduction to Anomaly Detection

Applications:

- Intrusion detection (physical or electronic)
- Fraud detection
- Health monitoring (people, animals, machines)

Introduction to Anomaly Detection

Techniques:

- Density: kNN, local outlier factor
- SVM
- Clustering: *k*-Means

Introduction to Anomaly Detection

kNN techniques and variations

- Voronoi diagrams
- aNN

LOF

- Measure average density using kNN
- Points with low local density are suspect outliers
- There is no good thresholding technique

k-Means

Examples

ping times

Examples

httpd response times

Examples

single/multiple host access abuse (DOS/DDOS)

Examples

bank card fraud

Examples

spam

Questions?

ml-week.com/1