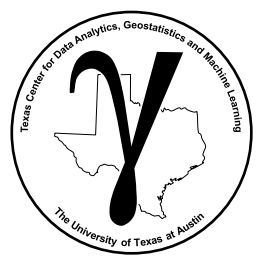


PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:

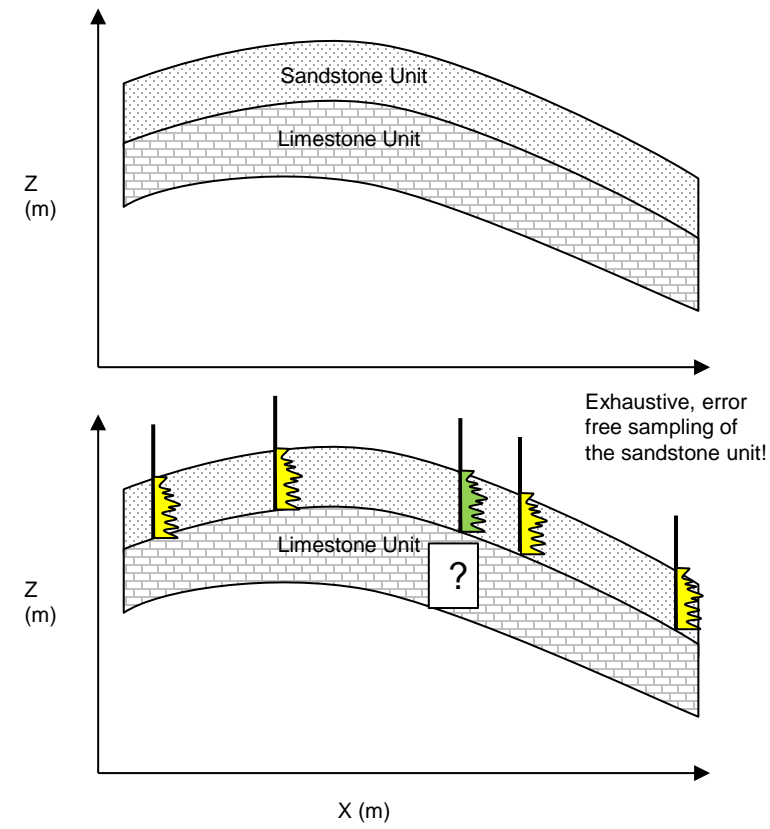
- **Prototype Methods**
- **K-means Clustering**
- **K-means Clustering Hands-on**
- **Other Clustering Methods**



Motivation for Clustering

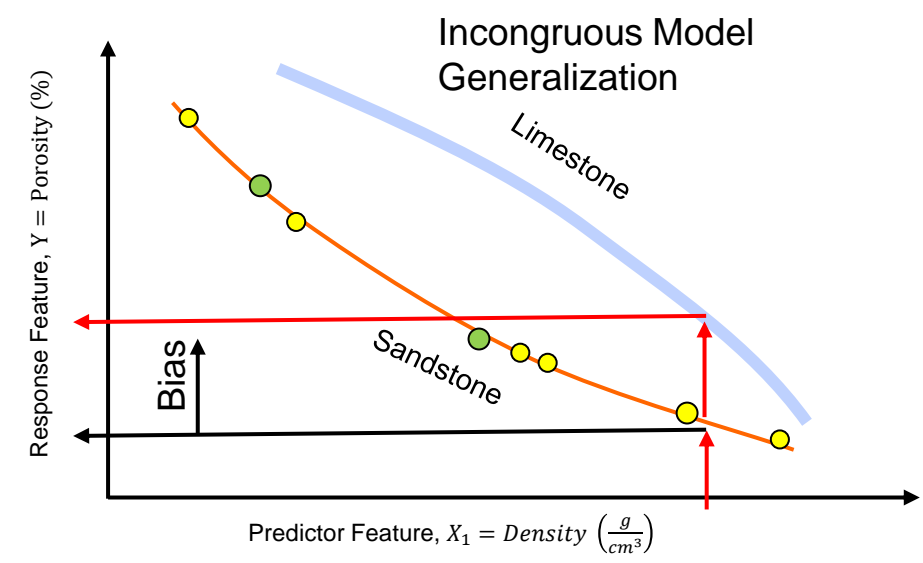
Consequence of mixing populations

We need to learn and segment distinct populations to improve our prediction models.

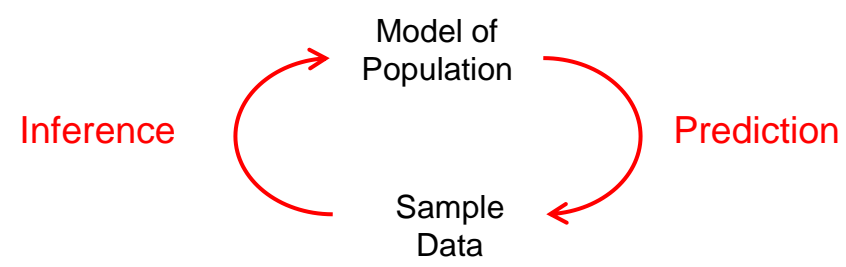


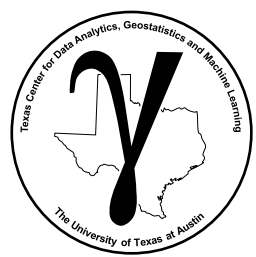
Building a model with sandstone data.

Exhaustive, error free sampling of the sandstone unit!



Then predicting for limestone data.





Motivation for Clustering

Our first inferential method for machine learning

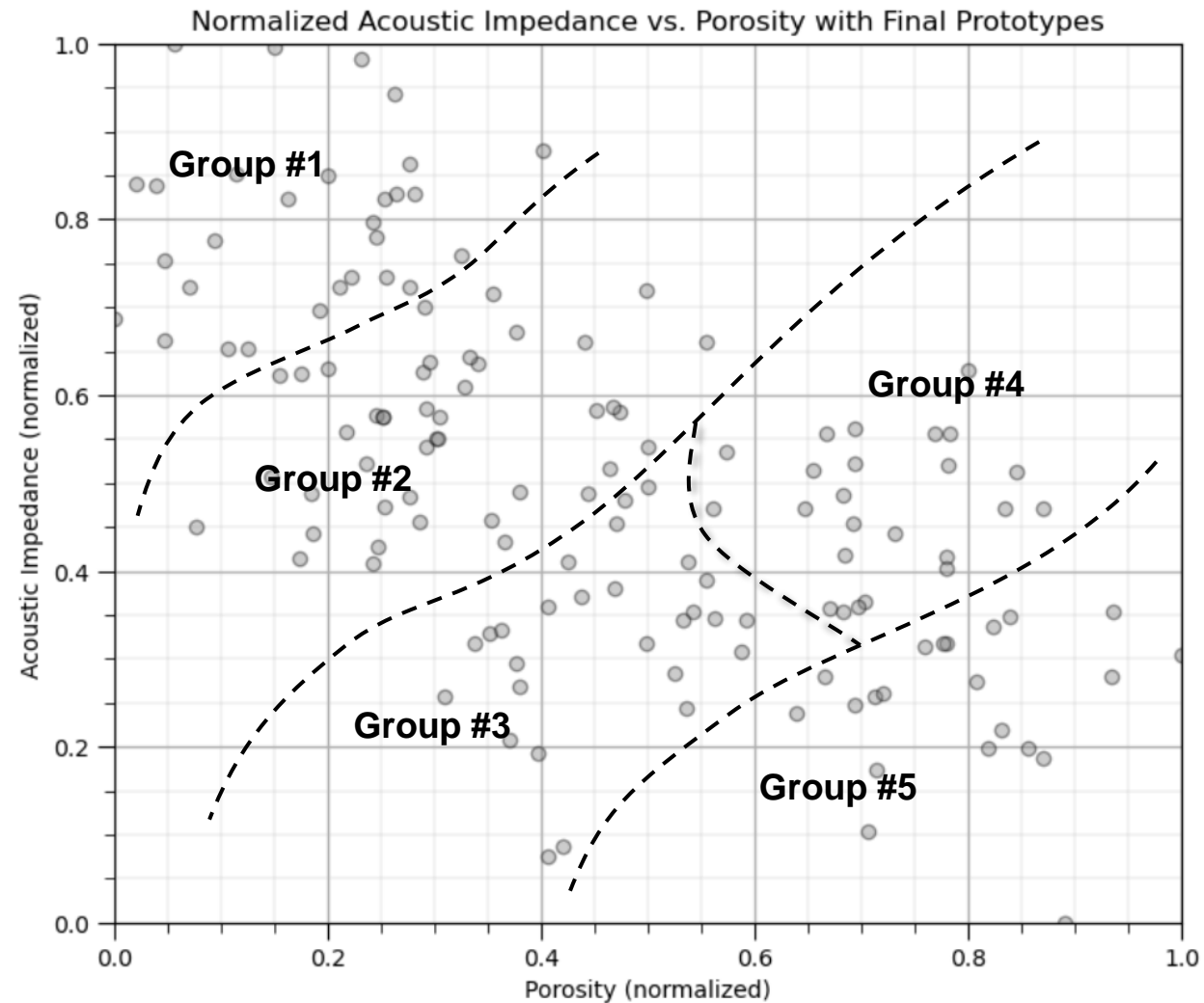
- Detect patterns (sample clustering) in predictor feature space

$$X_1, \dots, X_m$$

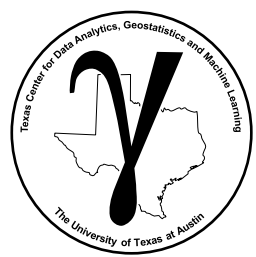
- Simple and visual method, gentle introduction

Teaches concepts like

- Feature transforms
- Distance metrics in feature space



Cluster analysis is not a predictive model, so it does not predict the group membership of new data (it doesn't give us the boundary / line as shown above). From Cluster Analysis chapter of Applied Machine Learning in Python e-book at, https://geostatsguy.github.io/MachineLearningDemos_Book/.



Motivation for Clustering

Our first inferential method for machine learning

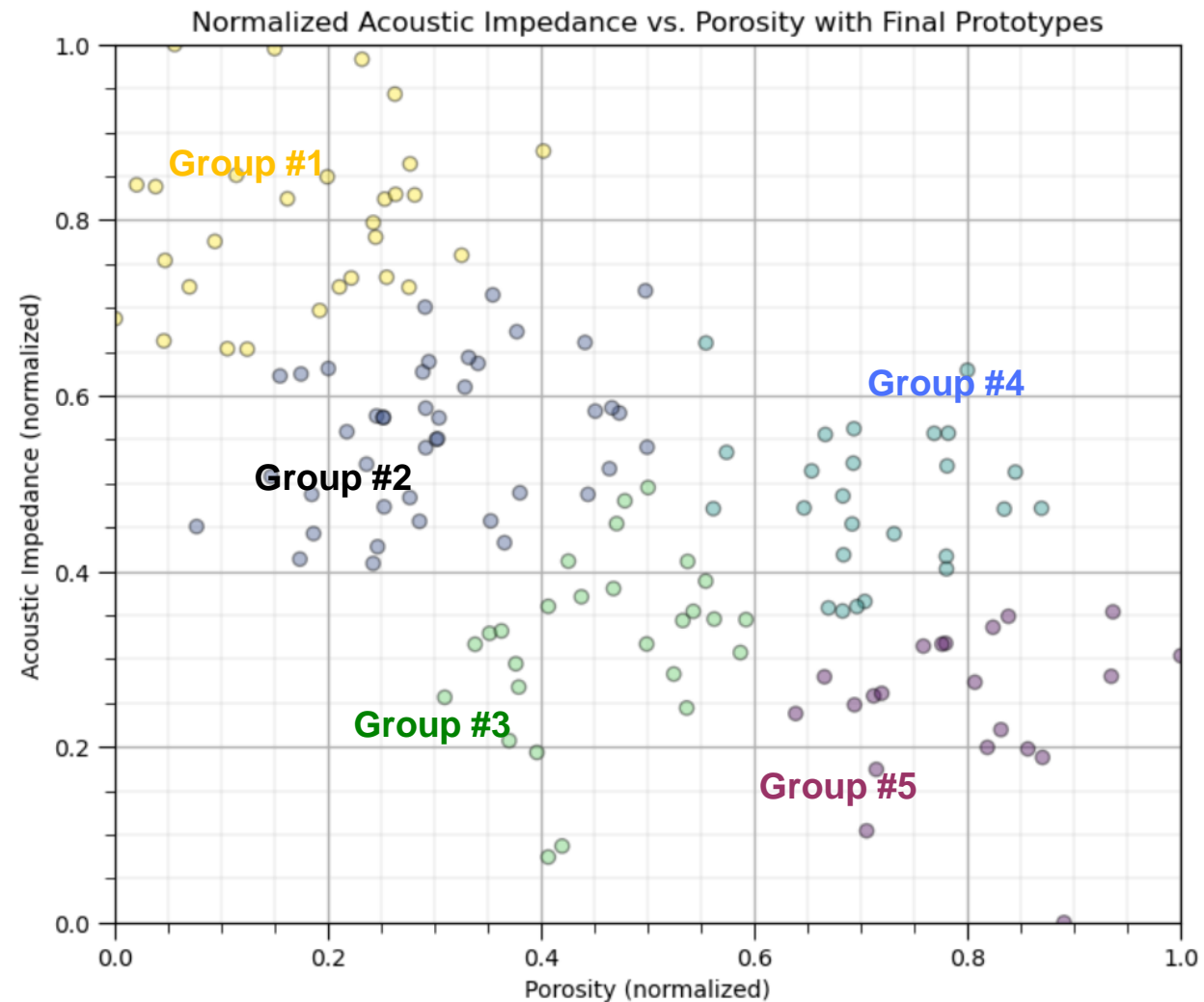
- Detect patterns (sample clustering) in predictor feature space

$$X_1, \dots, X_m$$

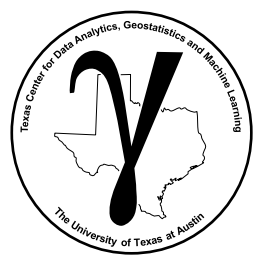
- Simple and visual method, gentle introduction

Teaches concepts like

- Feature transforms
- Distance metrics in feature space



Cluster analysis assigns group membership to data. Modified from Cluster Analysis chapter of Applied Machine Learning in Python e-book at, https://geostatsguy.github.io/MachineLearningDemos_Book/.



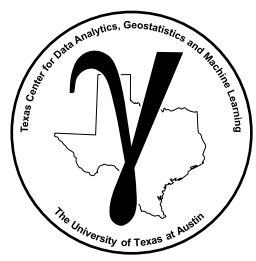
Unsupervised Learning

Unsupervised learning learns patterns in data from unlabeled data.

- No response features, Y , just predictor features.

$$X_1, \dots, X_m$$

- Machine learns by mimicry a compact representation of the data
- Captures patterns as feature projections, group assignments, neural network latent features, etc.
- We focus on inference of the population, the natural system, instead of prediction of response features.

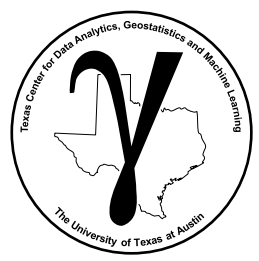


PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:

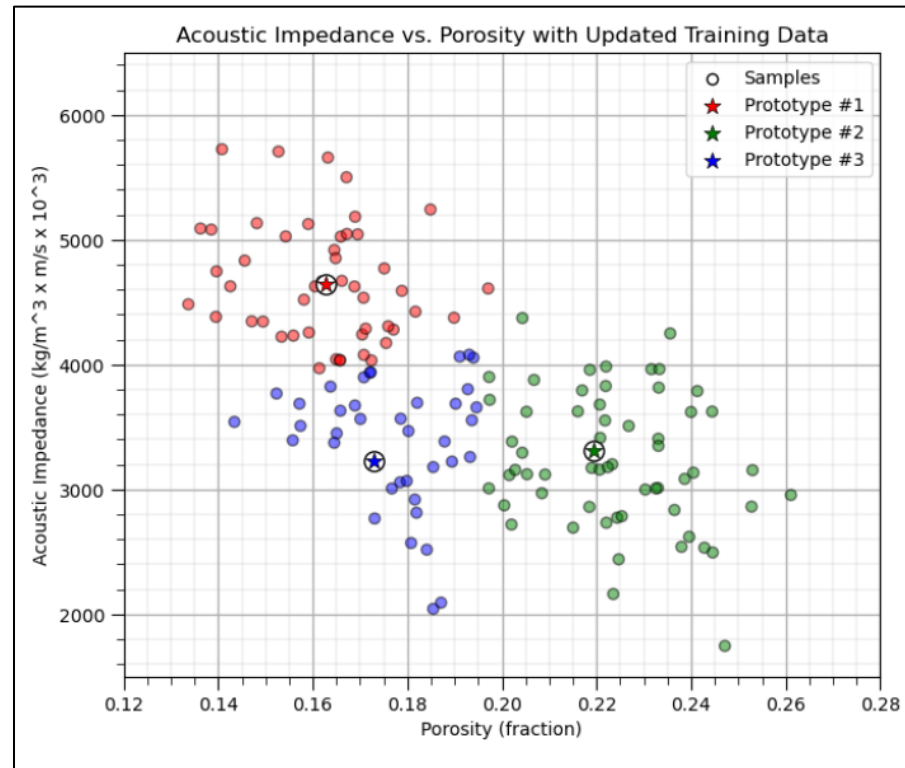
- **Prototype Methods**



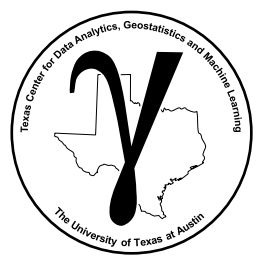
Prototype Methods

Represent the sample data with set of points in the feature space.

- Prototypes are typically not actual samples
- Sample data often assigned to the nearest (Euclidean) distance prototype



Data samples and prototypes, from
MachineLearning_clustering chapter of e-book.

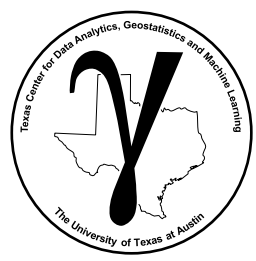


PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:

- **K-means Clustering**



Clustering

Clustering by Similarity:

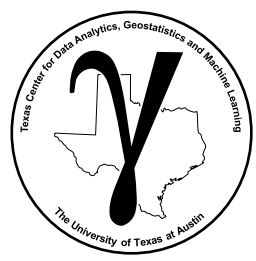
Assignment of groups (categorical assignment) to sample data such that the data within a group are more similar to each other than they are to data outside of the group.

- In general, minimize difference within a group and maximize difference between groups. Difference is general the distance separation in feature space.

More Complicated Group Detection:

Some clustering methods are designed to work with complicated group shapes/different size groups and even user supplied connections over networks.

- *Cluster assignment may go beyond a proximity in predictor feature space.*
- This first lecture we cover the simplest form of clustering. Next lecture we cover 2 advanced methods.



Clustering

General Comments:

One of the most important and powerful inferencial approaches

- known as clustering, cluster analysis, automatic classification, numerical taxonomy, typological analysis (Madhulatha, 2012)

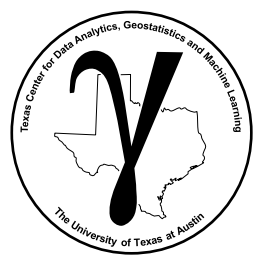
Two general approaches for clustering,

Partitional Clustering

- All clusters determined at once, k-means clustering
- Easy to update, for example, by modifying the prototype locations and recalculating the group assignments

Hierarchical Clustering

- Agglomerative (bottom-up) – start with n clusters and merge
- Divisive (top-down) – start with 1 cluster and divide
- Once a merge / split is made it cannot be undone, can't be updated



K-means Clustering

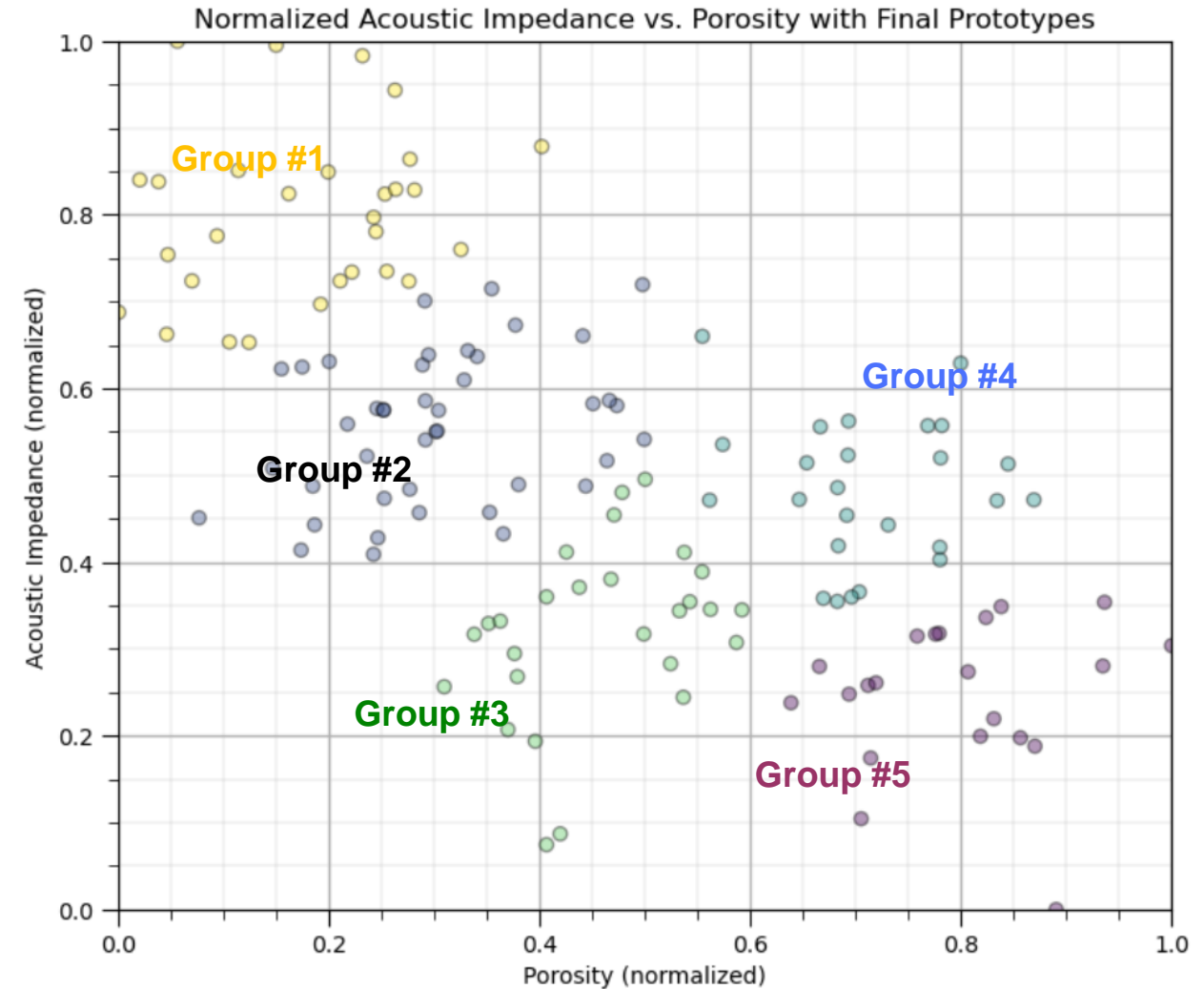
Assign clusters (groups) to all of the sample data in the feature space

Such that the groups are exhaustive:

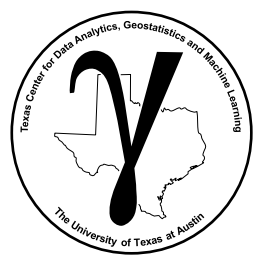
$$P(C_1 \cup C_2 \cup \dots \cup C_K) = 1.0$$

and the groups are mutually exclusive:

$$P(C_i \cap C_j \mid i \neq j) = 0.0$$



Cluster analysis assigns group membership to data. Modified from from Cluster Analysis chapter of Applied Machine Learning in Python e-book at, https://geostatsguy.github.io/MachineLearningDemos_Book/.



K-means Clustering

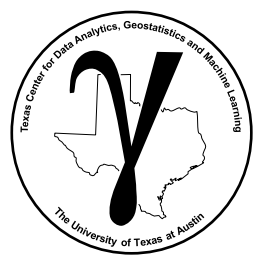
The groups are assigned to minimize the difference between the sample data in each group.

With K-means clustering we minimize this loss function,

$$J = \sum_{k=1}^K \sum_{i \in k} ||x_i - \mu_k||^2$$

where μ_k is the k cluster prototype, $i \in k$ are all data in the k cluster, and K is the total number of groups/clusters.

- $||x_i - \mu_k||^2 = \sqrt{\sum_{m=1}^M (x_i^m - \mu_k^m)^2}$ is Euclidian distance between data samples i its group prototype, μ_k , over all features, $m = 1, \dots, M$. Note, since $i \in k$, we only consider distances between data samples, i , and their k group's prototype.
- J is known as inertia or within-cluster sum of squares (WCSS).

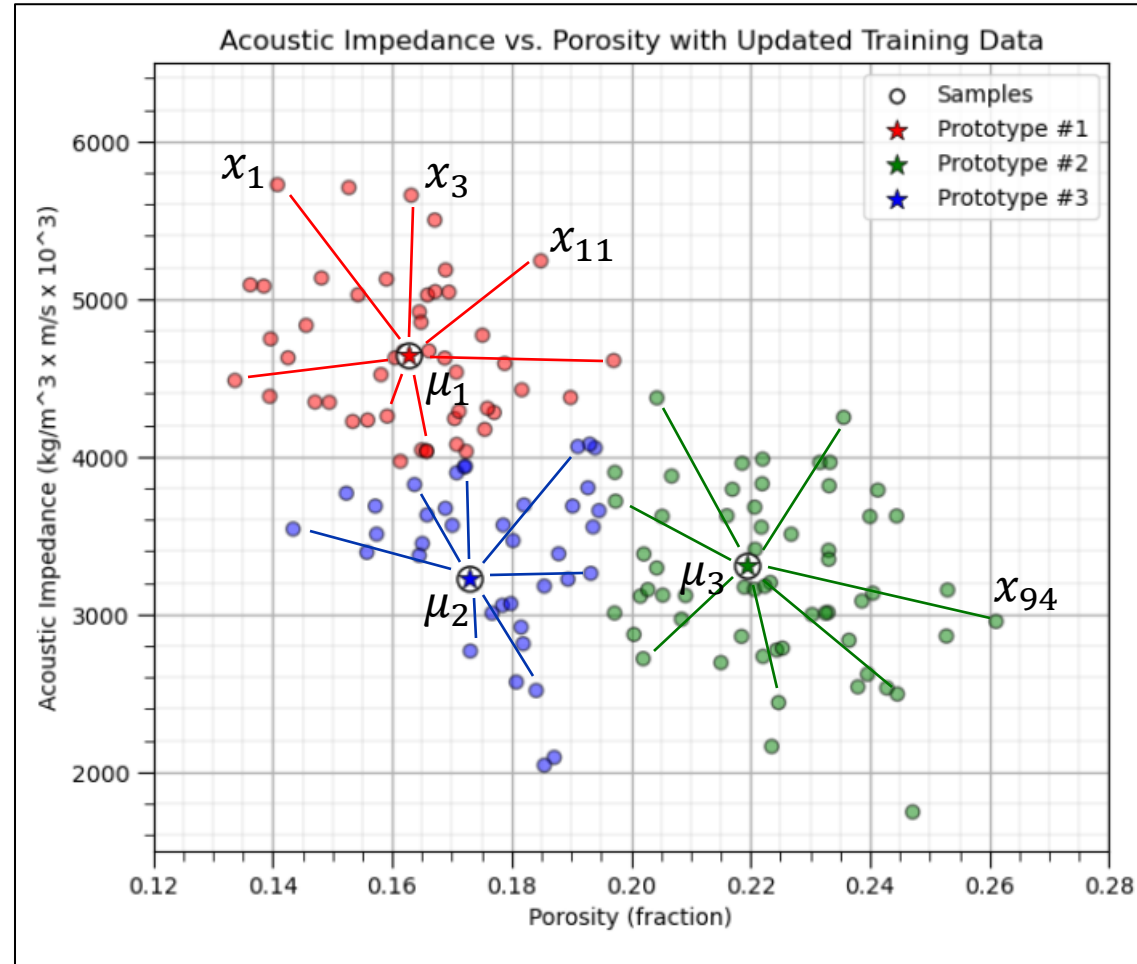


K-means Clustering

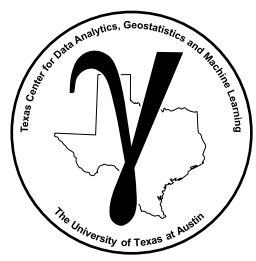
The groups are assigned to minimize the difference between the sample data in each group.

- Difference as m dimensional distance between each sample and the sample's group prototype.
- Training K-means clustering is minimizing inertia,

$$\min \left\{ \sum_{k=1}^K \sum_{i \in k} \|x_i - \mu_k\|^2 \right\}$$



Data samples and prototypes with a few distances indicated to illustrate the K-means clustering loss function, modified from MachineLearning_clustering chapter of e-book.



Loss Function Definition

The equation that is minimized to train model parameters.

For example,

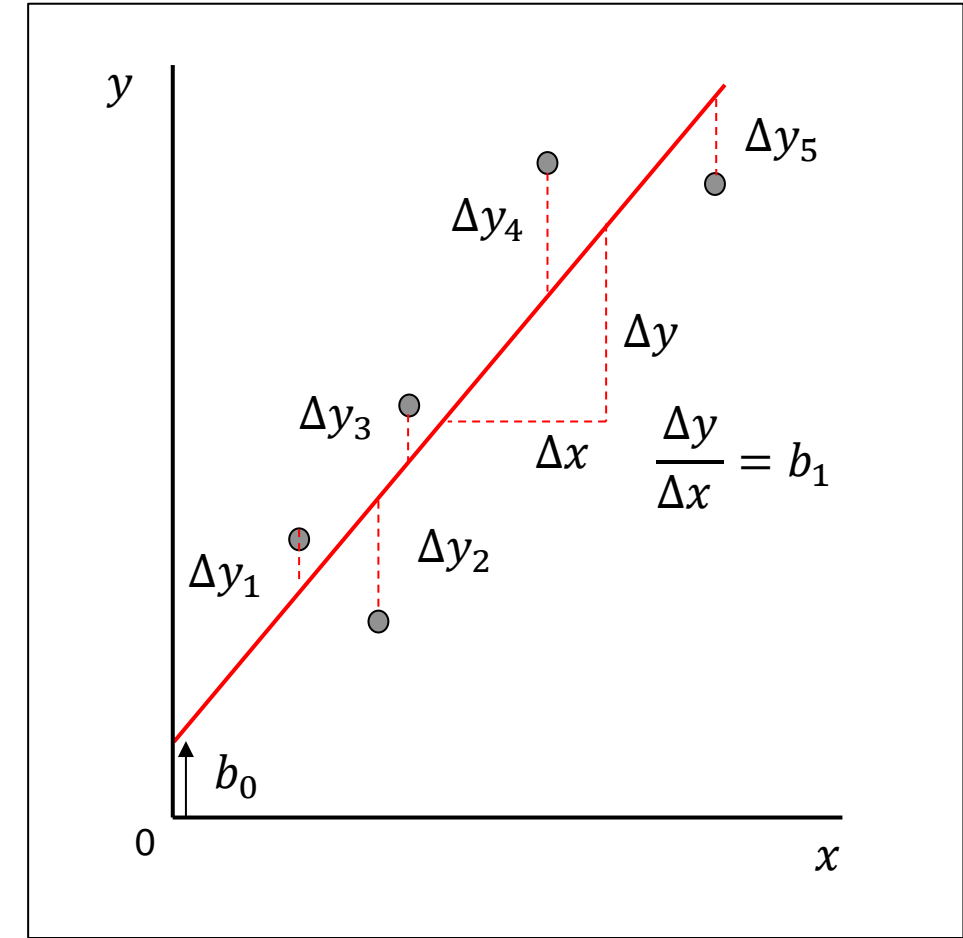
K-means clustering - inertia,

$$J = \sum_{k=1}^K \sum_{i \in k} ||x_i - \mu_k||^2$$

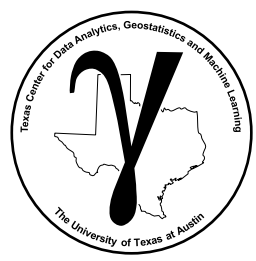
Linear regression – sum of square error,

$$SSE = \sum_{i=1}^n (\Delta y_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

We cannot train our model to training data without a loss function to specify the objective of the model.



Linear regression data, model and errors.



K-means Clustering

General Comments on K-means Clustering

- large solution space, K^n , where K is number of clusters and n is the number of data.

Computational Complexity of K-means clustering by checking all possible groups to minimize the loss function, known as brute force,

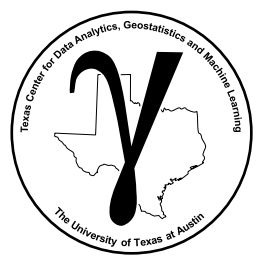
$$O(K^N \cdot n \cdot K)$$

- but the problem is practically solved iteratively with a **heuristic algorithm** (to be shown next)

$$O(I \cdot n \cdot K) \quad \text{where } I \text{ is the number of iterations}$$

- may converge to a local minimum (MacQueen, 1967), practically mitigated by seeding multiple random initial centroids and take the best converged solution

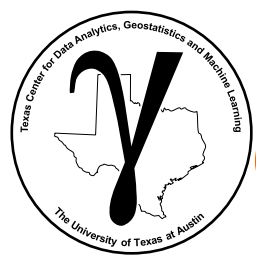
Note, we will define and explain computational complexity later with linear regression.



Heuristic Algorithm Definition

Heuristic Algorithm

- a shortcut solution to solve a difficult problem
- a compromise of optimality, accuracy for speed
- this general approach is common in machine learning, computer science and mathematical optimization

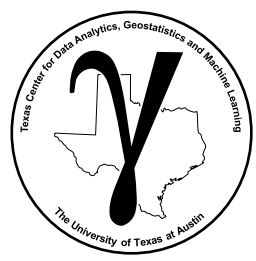


K-means Clusters and Curse of Dimensionality

General Comments on K-means Clustering with Many Features

Impact of the curse of dimensionality on cluster analysis:

- **Visualization:** clusters in high dimensionality space are impossible to visualize, even visualization of subspaces become meaningless. e.g. visualizing a subspace is not sufficient (e.g. visualize matrix scatter plots, 2 features at a time).
- **Concept of Distance:** distance becomes imprecise; therefore, the cluster assignments become imprecise!
- **Local Feature Relevance Problem:** groups may cluster based on unique combinations of features. With large numbers of features, it is common for all features not to be meaningful for all clusters.
- **Feature Redundancy:** results in cluster groups with arbitrary orientations that violate assumptions for methods such as K-means clustering.



K-means Clustering

General Comments on K-means Clustering

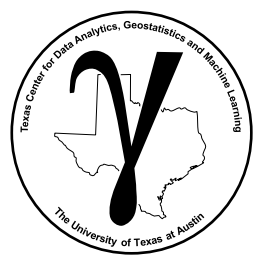
Clustering - of sample data in feature space to identify similar cases / distinct subsets in the multivariate data space.

Unsupervised - all predictor features X_1, \dots, X_m , no response feature labels

Inference – learning about the system, these groups, e.g., facies often explain 80% or more of the heterogeneity. **Some may argue this is not inferential (not linked to hypothesis testing), but in the subsurface, spatial phenomenon such tests are not generally practical.**

Prototype Method - represents the sample data with number of synthetic cases in the features space. For K-means clustering we assign prototypes.

Iterative Solution - the initial prototypes are assigned randomly in the feature space, the labels for each data sample are updated to the nearest prototype, then the prototypes are adjusted to the centroid of their assigned sample data, repeat until convergence.



K-means Clustering

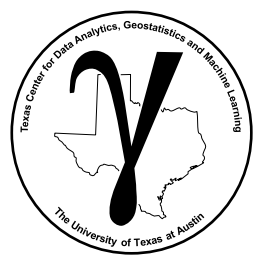
Advanced Comments on K-means Clustering

Feature Weighting

- the procedure depends on the 'distance' between data samples and prototypes in feature space. If the features have significantly different magnitudes, the feature(s) with the largest magnitudes and ranges will dominate the inertia loss, J .
- common approach is to standardize the features to avoid unintended feature weighting

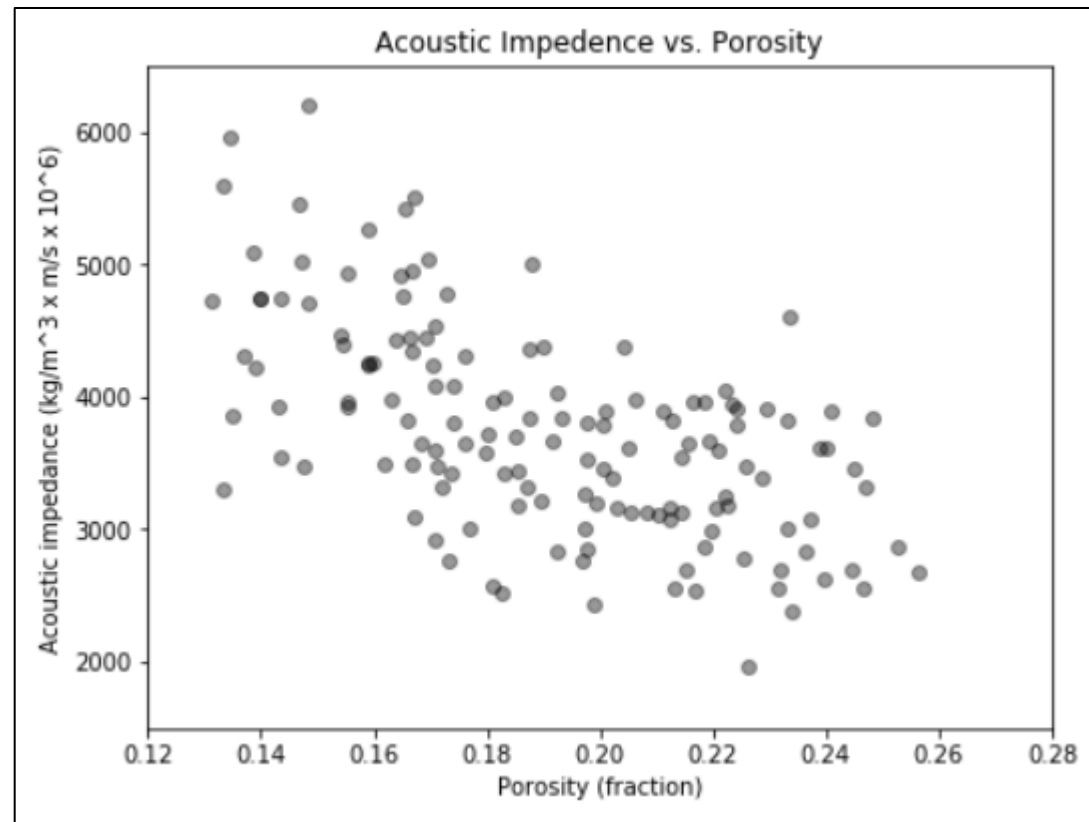
Supervised Learning Classification Variant

- applies multiple prototypes in each category to then inform a decision boundary for classification.

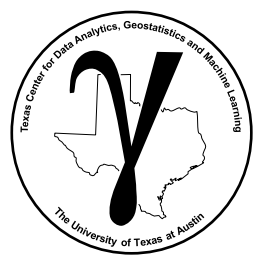


K-means Clustering

For example, given this sample data with porosity and acoustic impedance, find ‘ K ’ facies that:
For example, segment the porosity and acoustic impedance feature space.



Data samples, from
MachineLearning_clustering chapter of e-book.

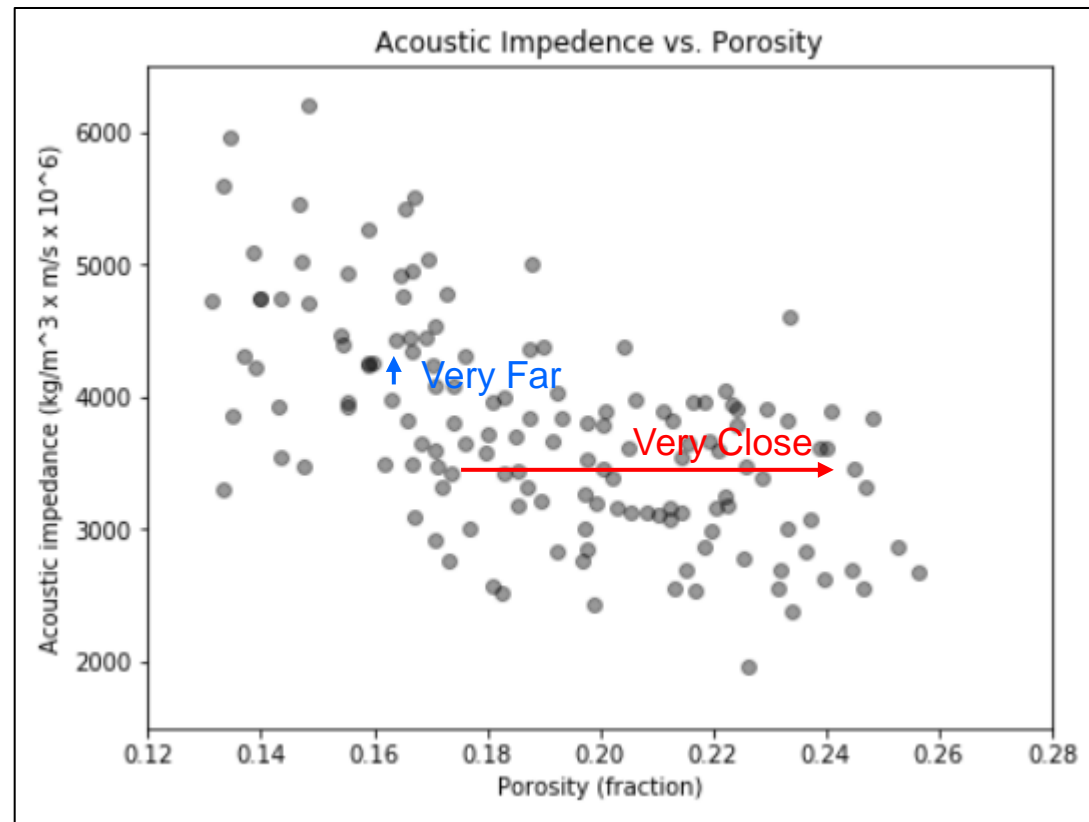


k-means Clustering

Feature Normalization

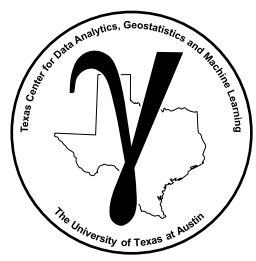
We require a measure of similarity,

Consider Euclidean distance with the original units, $d = \sqrt{\Delta Por^2 + \Delta AI^2}$



Data samples, from
MachineLearning_clustering chapter of e-book.

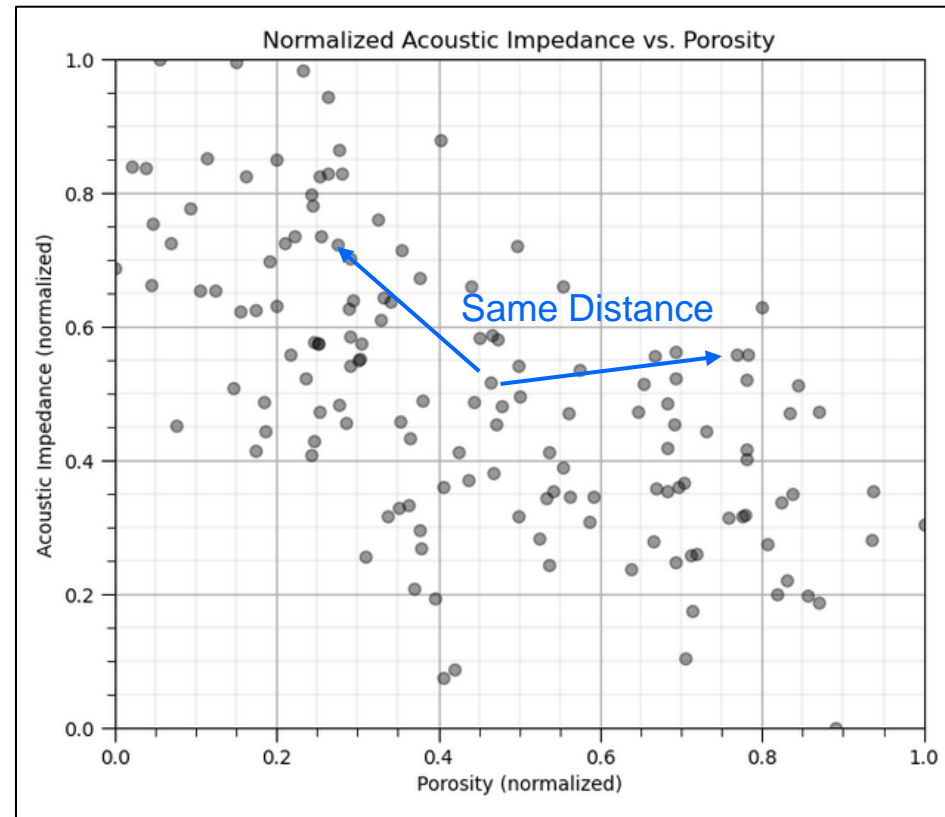
- For dissimilar units we require normalization, similar magnitude and range over all features.



k-means Clustering Feature Normalization

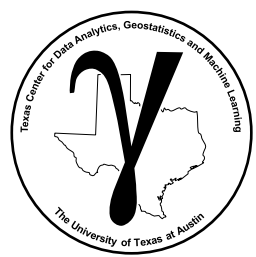
We require a measure of similarity,

Consider Euclidean distance with the normalized features, $d = \sqrt{\Delta Por^2 + \Delta AI^2}$



Normalized data samples, from
MachineLearning_clustering chapter of e-book.

For dissimilar units we require normalization, similar magnitude and range over all features.



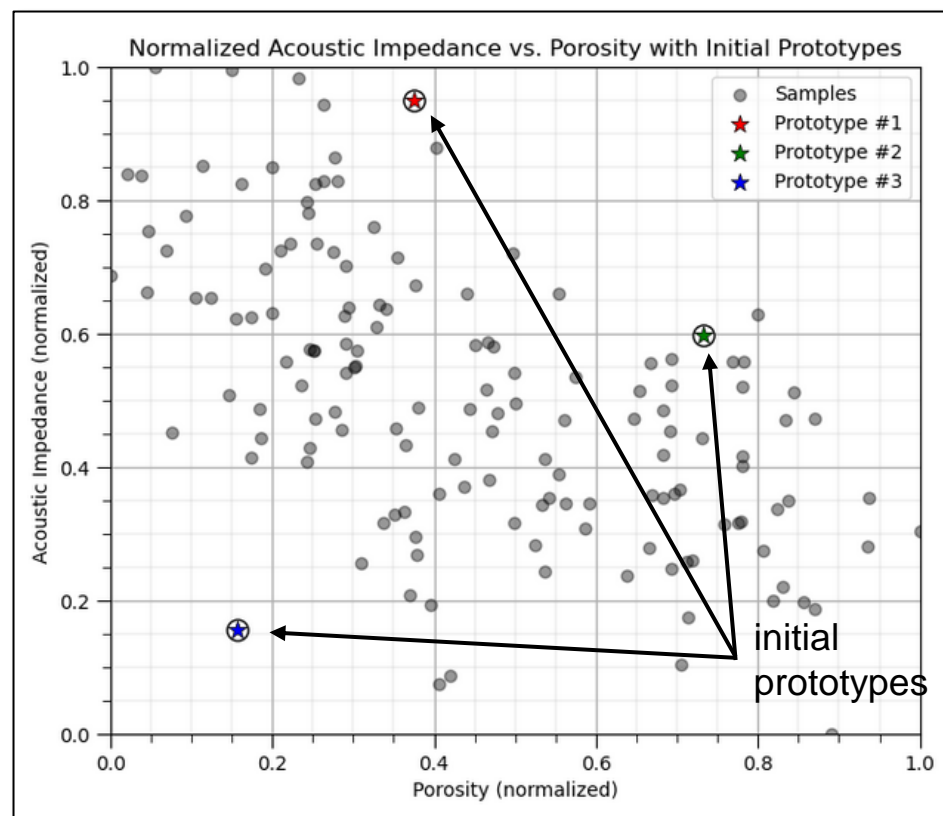
k-means Clustering Heuristic

Assign K prototypes in the feature space:

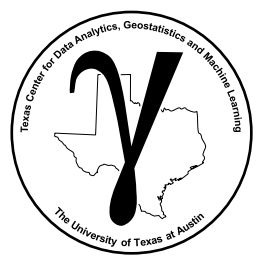
- random assignment (random value between min and max)

Note: the initial prototypes could be poor choices

- clustered, outside the sample data etc.



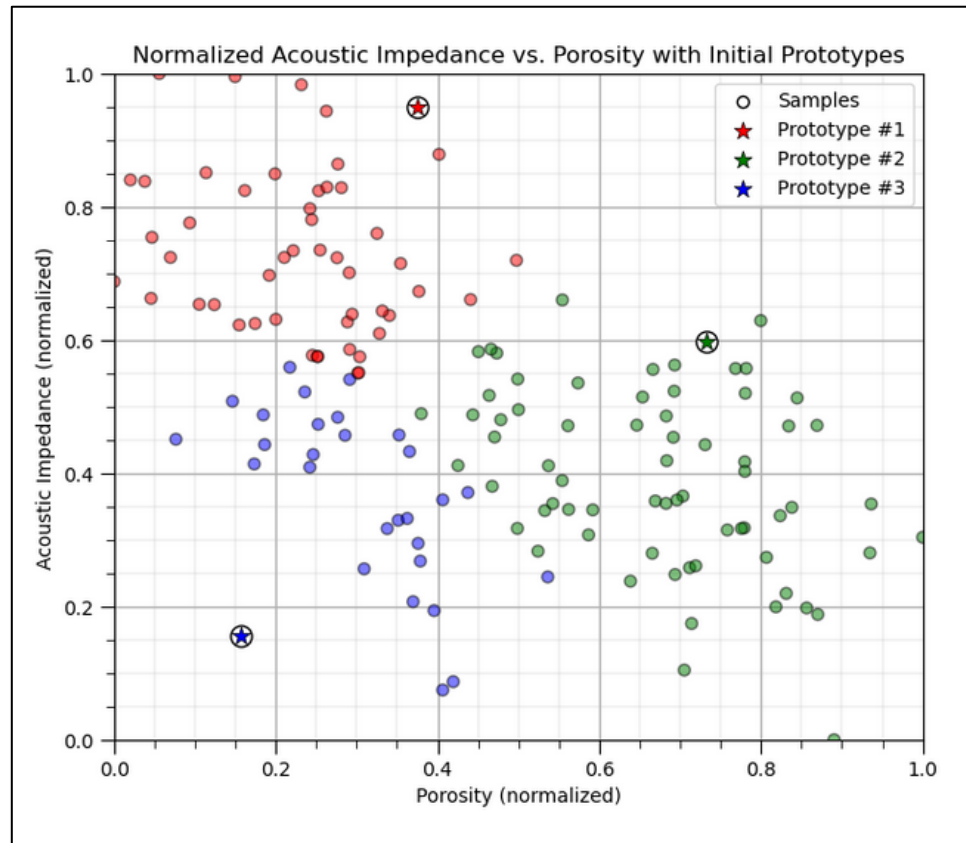
Normalized data samples and initial prototypes, from MachineLearning_clustering chapter of e-book.



k-means Clustering Heuristic

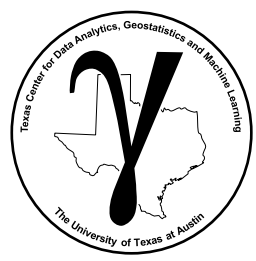
Assign sample data to the nearest prototype,

- we apply nearest Euclidian distance with the normalized features



Normalized data samples and initial prototypes and data assigned to nearest prototype, from MachineLearning_clustering chapter of e-book.

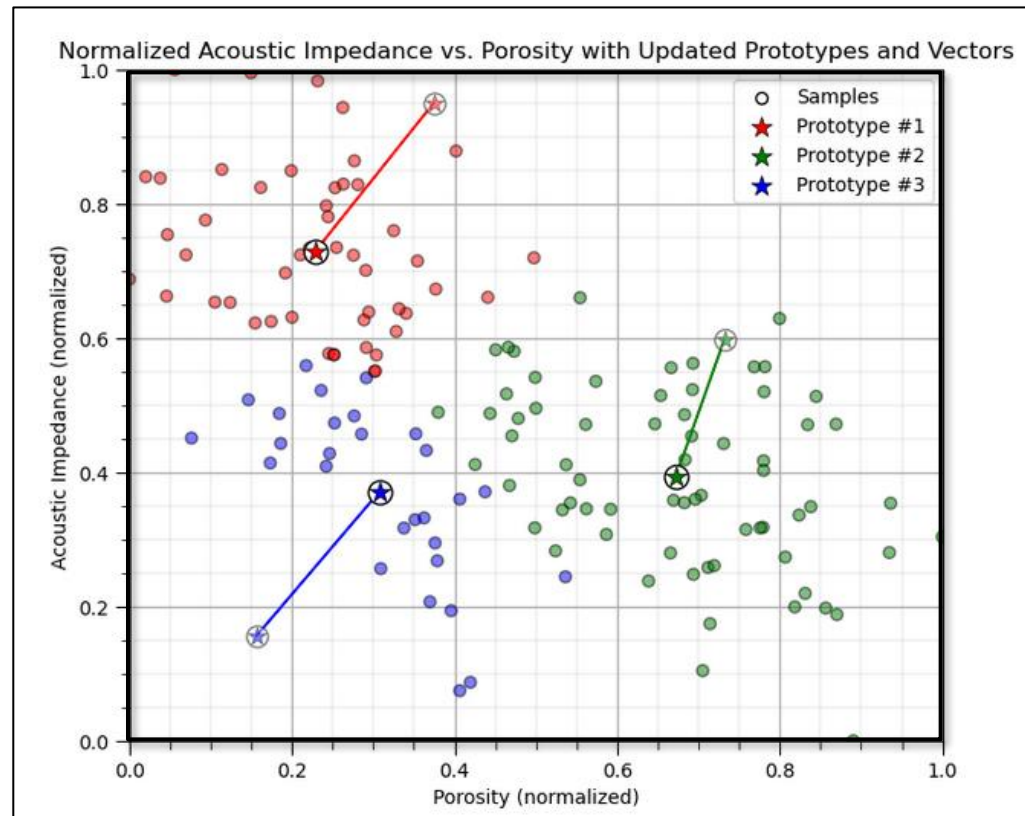
- Note, these initial clusters are not very good



k-means Clustering Heuristic

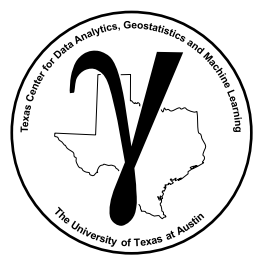
Update the prototypes to the centroids of the assigned sample data,

- vectors are included to show the update of the prototypes



Normalized data samples and updated prototypes, from MachineLearning_clustering chapter of e-book.

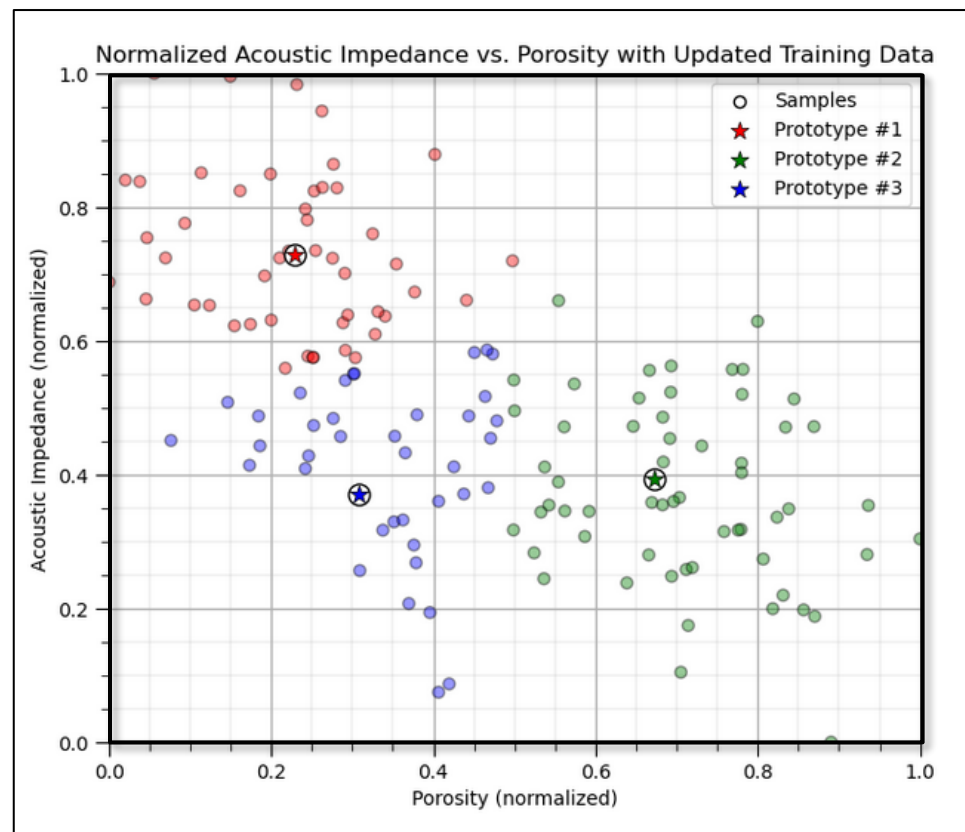
- The prototypes have improved.



k-means Clustering Heuristic

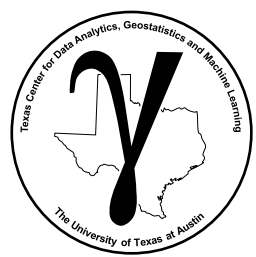
Update the sample data assignment to the nearest updated prototypes,

- once again, we apply nearest Euclidian distance with the normalized features



Normalized data samples and updated prototypes and updated data assignments, from MachineLearning_clustering chapter of e-book.

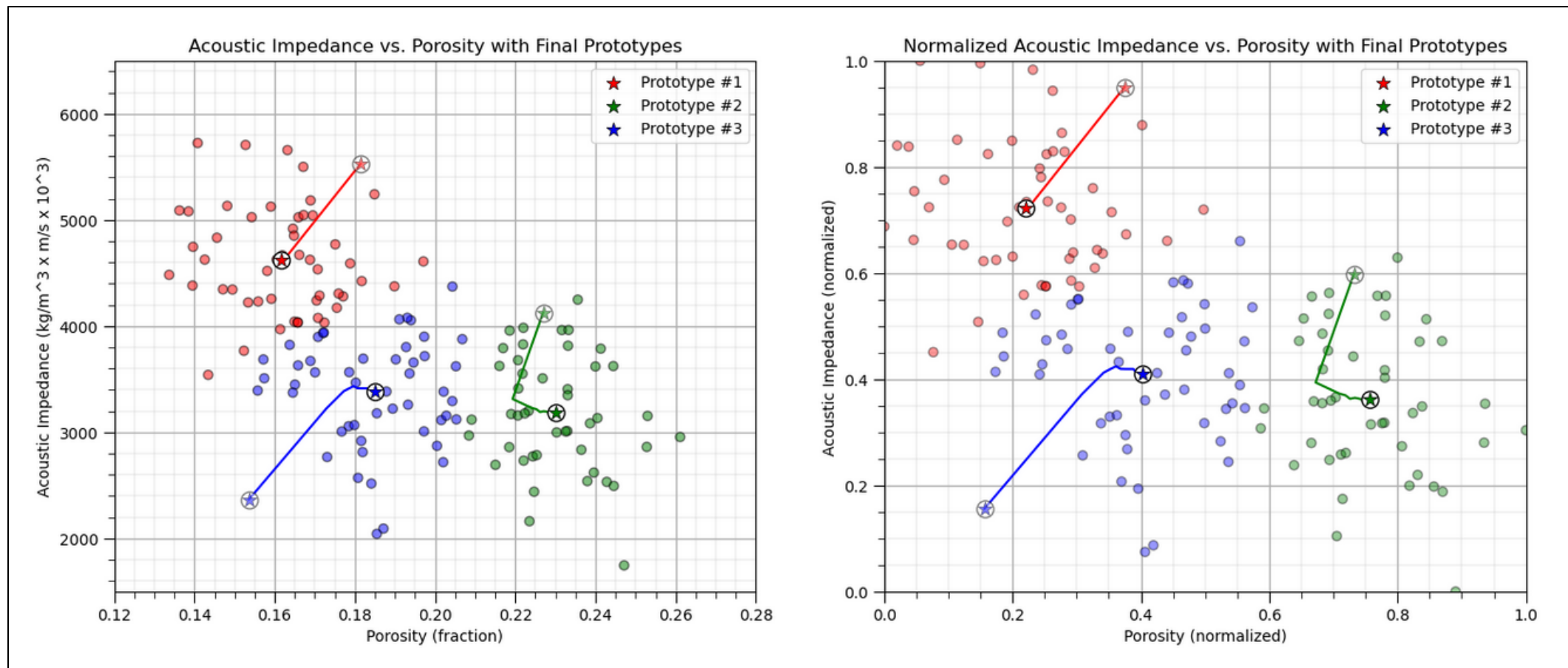
- The prototypes have improved.



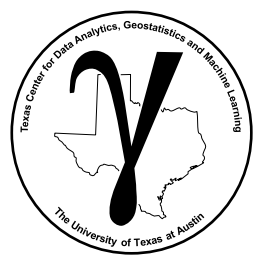
k-means Clustering Heuristic

Iterate until the centroid stop moving and the assignments stabilize,

- we now have 3 clusters that minimize the within group variation



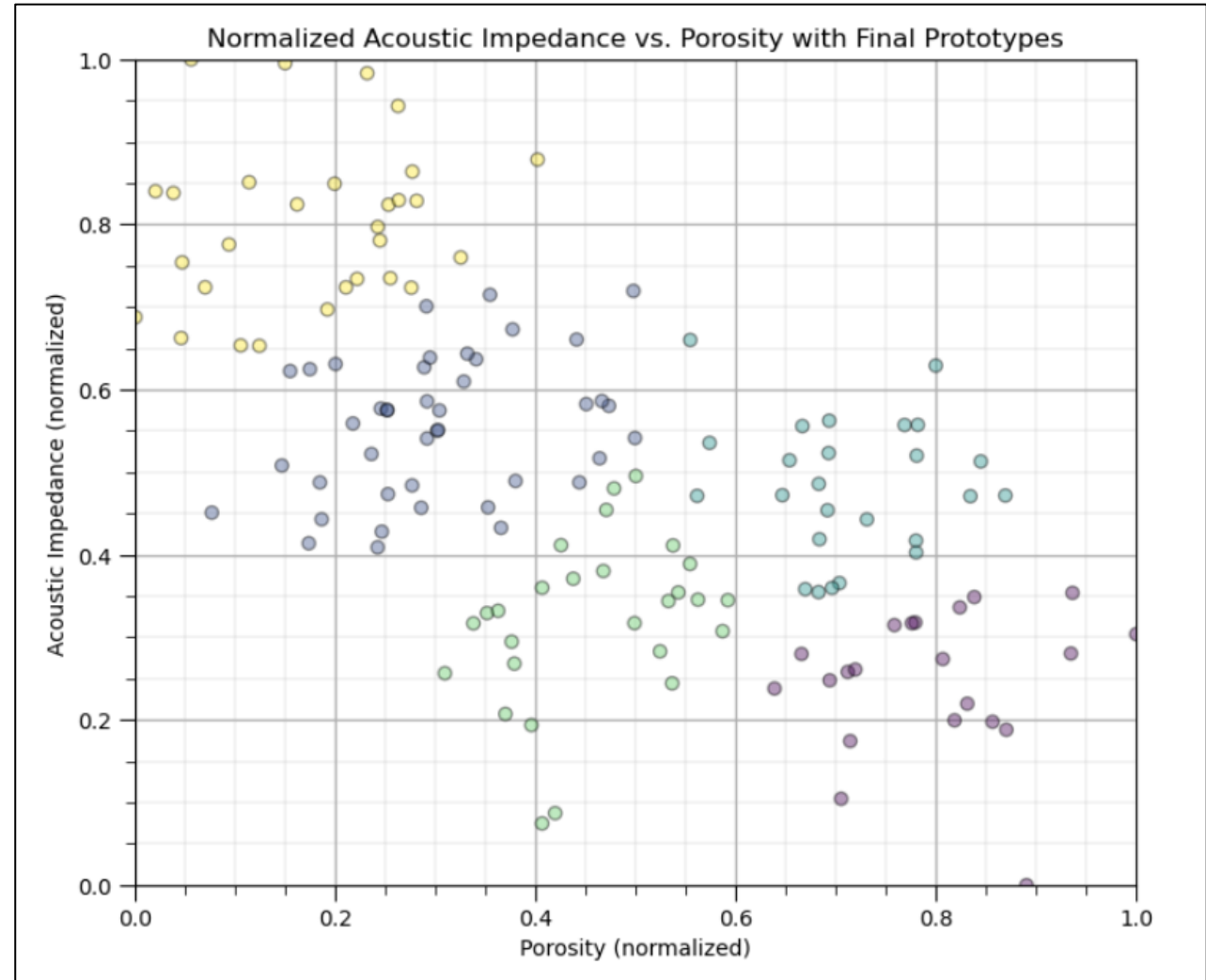
Final prototypes and group assignments after 3 iterations, from MachineLearning_clustering chapter of e-book.



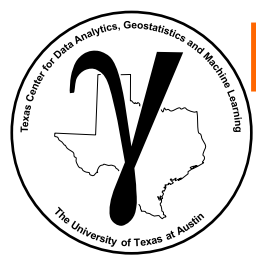
k-means Clustering Heuristic

The number of K clusters is an important decision:

- example with $K = 5$



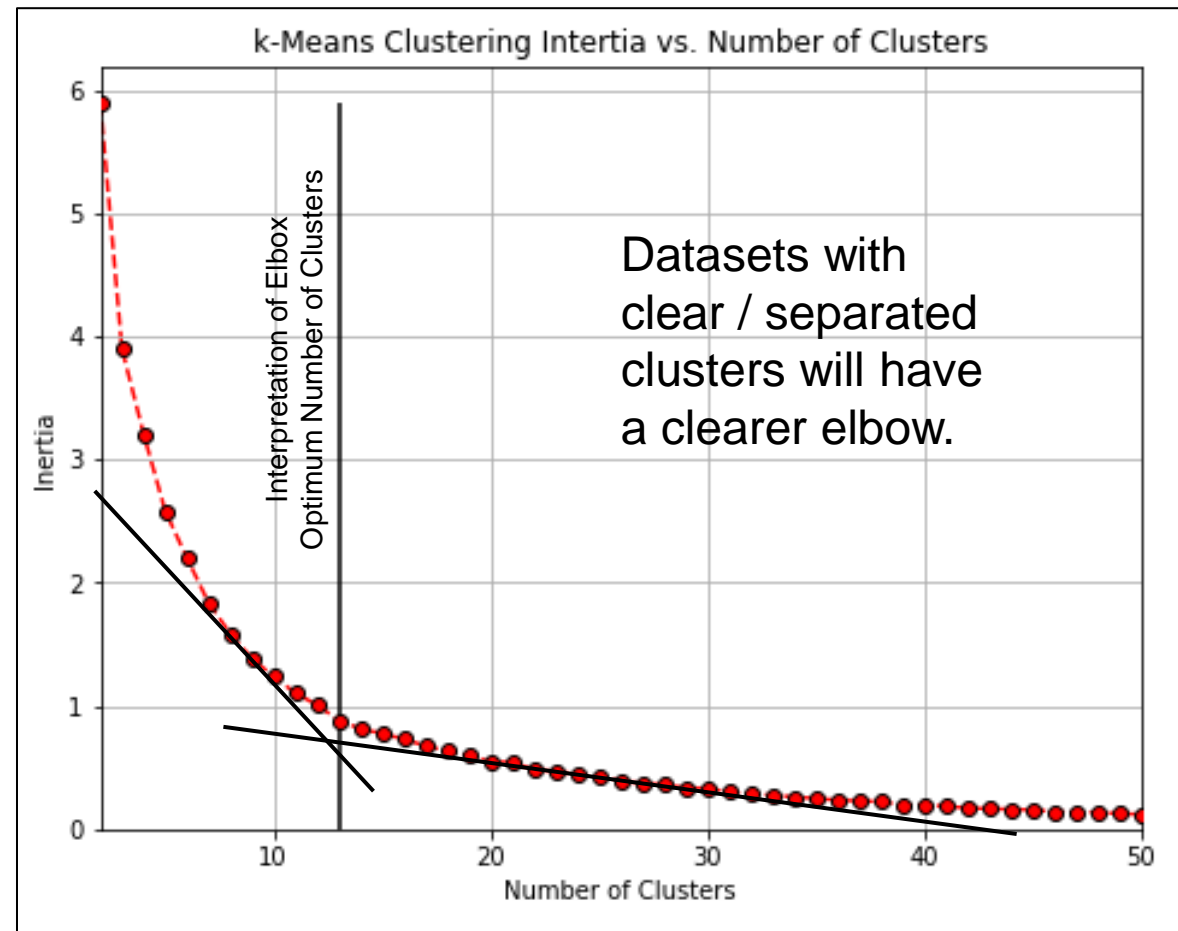
K-means clustering with $k=5$, from
MachineLearning_clustering chapter of e-book.



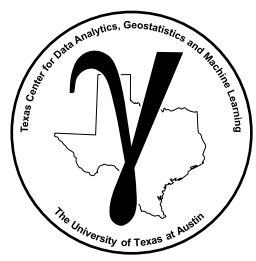
K-means Clustering Number of Clusters

Determination of the optimum number of clusters

- one method is known as the elbow method
1. Loop over k clusters, $k = 1, \dots, K$.
 2. Plot the loss function, inertia vs. the number of clusters.
 3. Identify the number of clusters at the elbow, point of 'diminishing returns'.



k-means clustering elbow plot with interpretation.



K-means Clustering Details

Minimizing the Inertia Metric

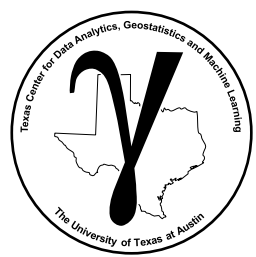
There exists a unique solution that will minimize the inertia, J .

$$J = \sum_{k=1}^K \sum_{i \in k} ||x_i - \mu_k||^2$$

Solution Uniqueness

The solution surface is typically not smooth, it is quite likely to be trapped in a local minima!

- the practical solution is to seed multiple initial prototypes and to retain the best solution.



k-means Clustering Details

How Stable is the k-means Clustering Solution Heuristic?

- I built out a Python dashboard to explore this.

Interactive k-Means Clustering Heuristic Demonstration

- select K , random number *seed* and step over the iterations of prototype location, observe the CDF of clustering inertia vs. the demonstrated case.

Michael Pyrcz, Professor, University of Texas at Austin

[Twitter](#) | [GitHub](#) | [Website](#) | [GoogleScholar](#) | [Book](#) | [YouTube](#) | [LinkedIn](#) | [GeostatsPy](#)

The Inputs

Observe the k-means clustering solution heuristic, e.g., step over iterations and change the random seed:

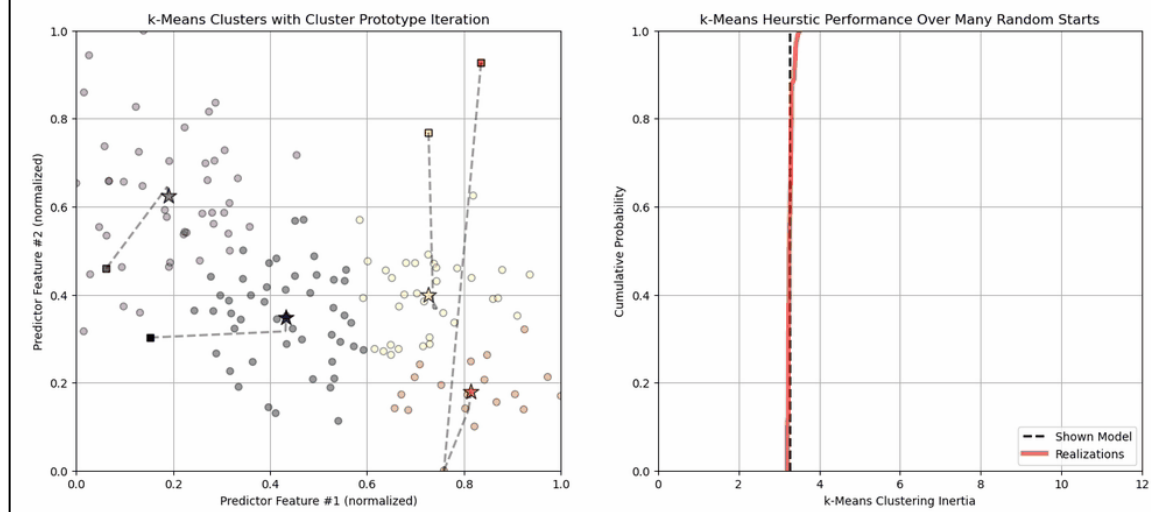
- K : number of k-means clusters, n_{iter} : number of iterations, *seed*: random number seed

```
1 display(ui2, interactive_plot)
```

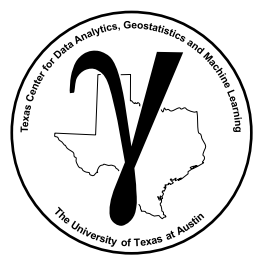
display the interactive plot

Machine Learning k-Means Clustering Heuristic Demo, Prof. Michael Pyrcz, The University of Texas at Austin

K n_{iter} *seed*



Exploration of k-means solution heuristic stability, file is
Interactive_kMeans_Clustering.ipynb.



k-means Clustering Assumptions

Assumptions of k-means clustering

1. spherical, convex, isotropic clusters

- minimize difference within clusters

2. equal variance for all features

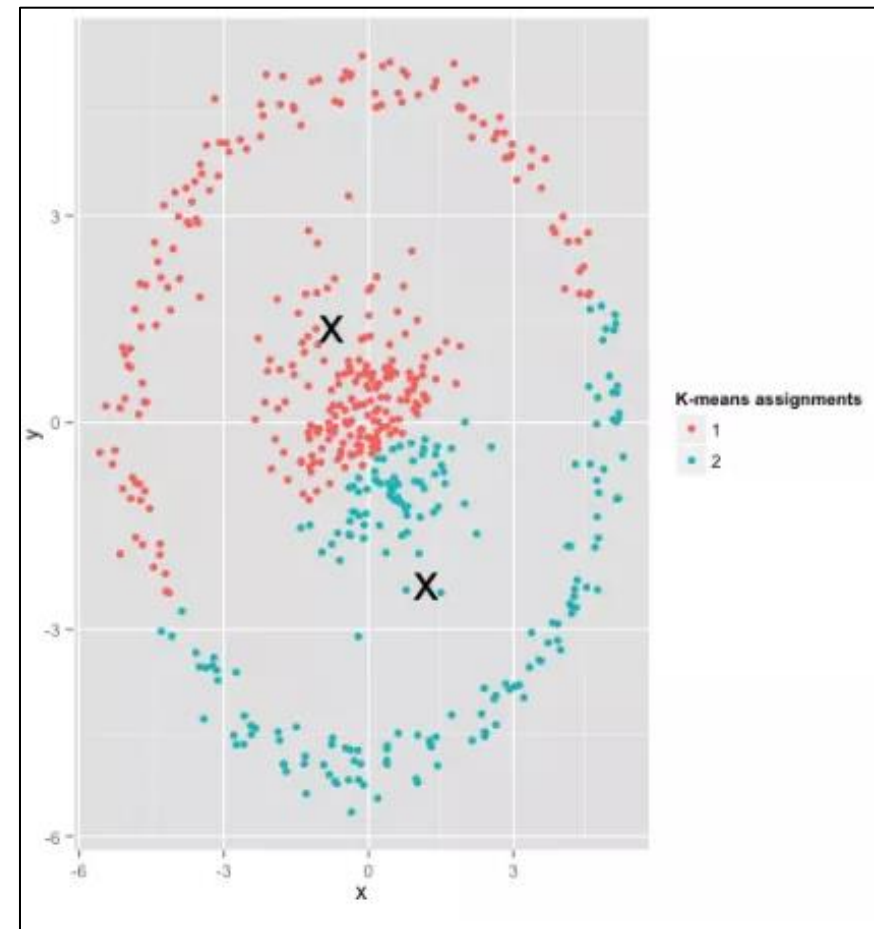
$$\sigma_{X_1}^2 = \sigma_{X_2}^2 = \dots = \sigma_{X_m}^2$$

- reliable measures of distance in feature space

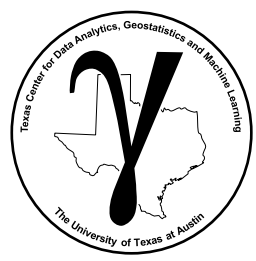
3. equal prior probability for all clusters:

$$P(x_\alpha = k) = \frac{1}{K} \quad \alpha = 1, \dots, n, k = 1, \dots, K$$

- no / naïve a priori clustering membership information (uniform distribution)



Nonspherical example figure from <https://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/>.

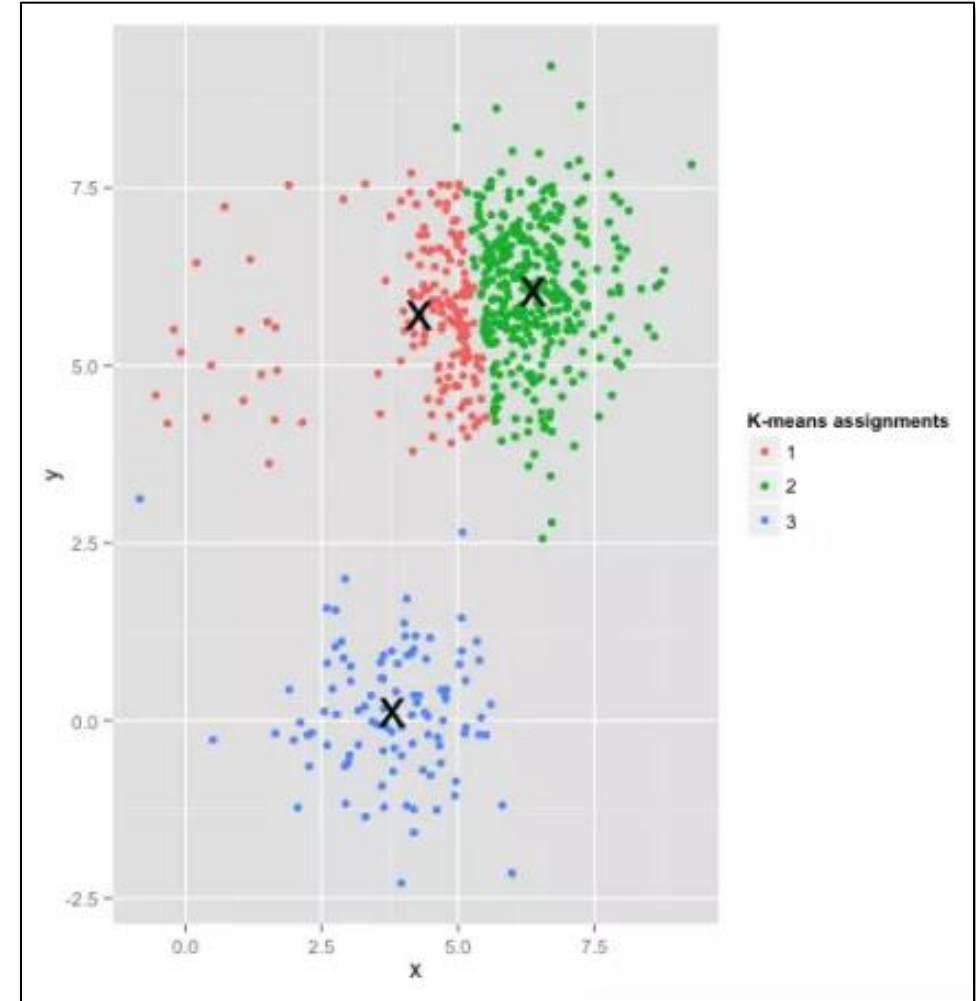


k-means Clustering Assumptions

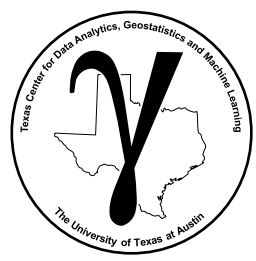
Assumptions of k-means clustering

4. similar sized / frequency clusters

- larger clusters are divided to minimize the overall variance within clusters
- clusters with few samples in feature space are overwhelmed!



Different cluster sizes from <https://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/>.

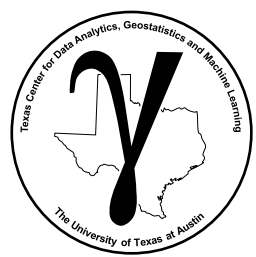


PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:


- **K-means Clustering Hands-on**



k-means Clustering Demonstration

Demonstration workflow with K-means clustering for unsupervised clustering / segmentation of sample data.

MachineLearning_clustering chapter of e-book.



Applied Machine Learning in Python: a Hands-on Guide with Code

Machine Learning Concepts

Workflow Construction and Coding

Probability Concepts

Loading and Plotting Data and Models

Univariate Analysis

Multivariate Analysis

Feature Transformations

Feature Ranking

Cluster Analysis

Density-based Clustering

Spectral Clustering

Principal Components Analysis

Multidimensional Scaling

Linear Regression

Ridge Regression

k-means Clustering

Michael J. Pyrcz, Professor, The University of Texas at Austin

[Twitter](#) | [GitHub](#) | [Website](#) | [GoogleScholar](#) | [Book](#) | [YouTube](#) | [Applied Geostats in Python e-book](#) | [LinkedIn](#)

Chapter of e-book “Applied Machine Learning in Python: a Hands-on Guide with Code”.

Cite this e-Book as:

Pyrcz, M.J., 2024, Applied Machine Learning in Python: a Hands-on Guide with Code, https://geostatsguy.github.io/MachineLearningDemos_Book.

The workflows in this book and more are available here:

Cite the MachineLearningDemos GitHub Repository as:

Pyrcz, M.J., 2024, MachineLearningDemos: Python Machine Learning Demonstration Workflows Repository (0.0.1). Zenodo. DOI [10.5281/zenodo.13835318](https://doi.org/10.5281/zenodo.13835318)

By Michael J. Pyrcz
© Copyright 2024.

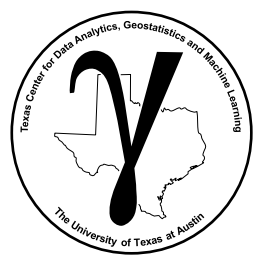
This chapter is a tutorial for / demonstration of **k-means Clustering**.

YouTube Lecture: check out my lectures on:

- [Introduction to Machine Learning](#)
- [Cluster Analysis](#)
- [Issues with k-Means Clustering](#)

Contents

- Motivation for Cluster Analysis
- Inferential Machine Learning
- k-Means Clustering
- The k-Means Clustering Demonstration
- Load the required libraries
- Declare Functions
- Set the Working Directory
- Loading Tabular Data
- Summary Statistics for Tabular Data
- Normalize the Features
- Extract Features of Interest
- Infer Model Parameters
- Visualize the Training Data
- Calculating k-Means Clustering By-hand
- k-Means Clustering with the scikit-learn Function
- Selecting the Optimum Number of Clusters
- Clustering without Normalization / Standardization
- Comments
- The Author:
- Want to Work Together?
- More Resources Available at: [Twitter](#) | [GitHub](#) | [Website](#) | [GoogleScholar](#) | [Book](#) | [YouTube](#) | [Applied Geostats in Python e-book](#) | [LinkedIn](#)

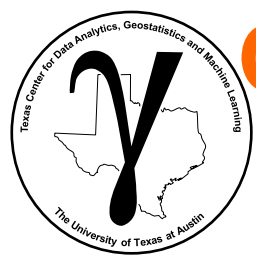


PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:

- Other Clustering Methods



Other Clustering Methods

K-medoids clustering, another partitional method

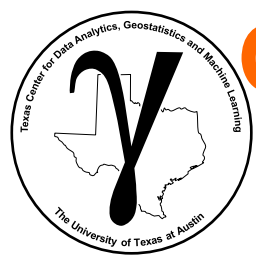
- Use actual data samples for prototypes, most central data sample to group, instead of centroid
- Swap randomly if loss function is reduced,

K-medoids Loss Function

$$C = \sum_{k=1}^K \sum_{i \in k} |x_i - x_k|$$

Sum of Dissimilarities

- where $|x_i - x_k|$ is any distance / dissimilarity measure, and x_k is the most central data sample in the k group, found as the sample that minimizes the loss.
- since prototypes are not based on a centroid (requiring averaging) any distance measure can be used, and even categorical features can be included.
- Less sensitive to outliers than K-means clustering, but other limitations of k-means clustering remain and generally slower, with higher computational complexity.

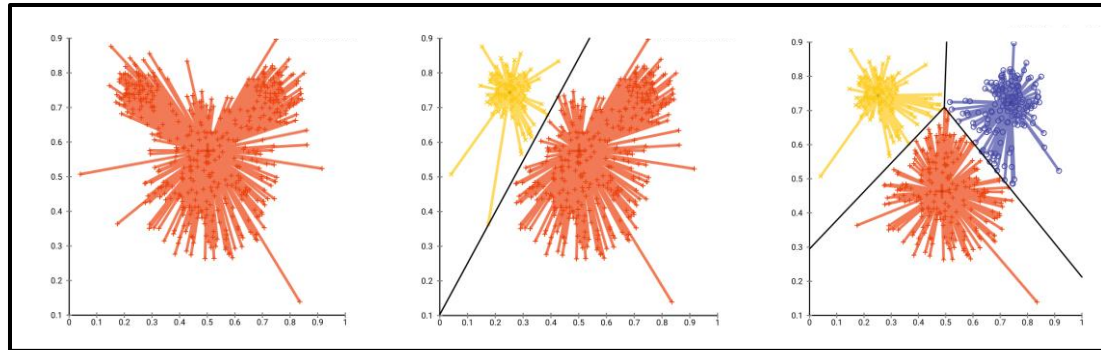


Other Clustering Methods

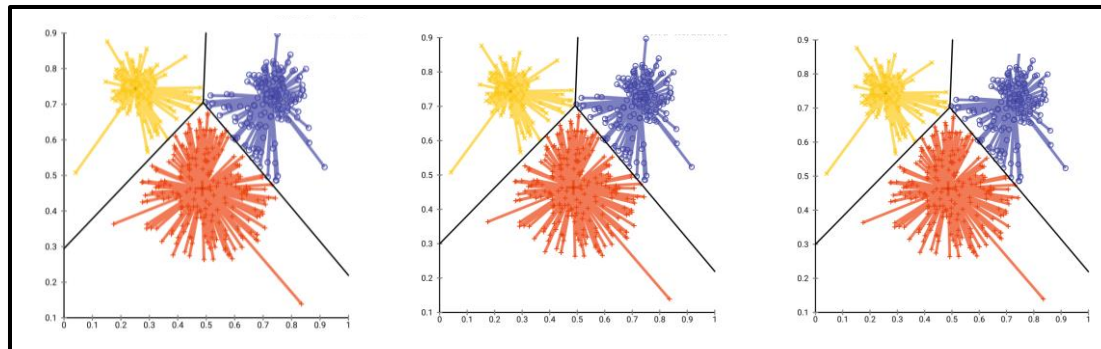
k-medoids clustering, also partitional method

The workflow:

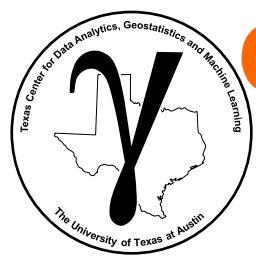
- Sequentially select k medoids from the data, assign data to nearest medoid, proceed greedily to minimize cost, \mathcal{C} .



- Loop over k medoids and all non-medoid data consider swap
- Make swap if \mathcal{C} is reduced. Loop until no swap found to reduce cost, \mathcal{C} .



Images from
https://upload.wikimedia.org/wikipedia/commons/e/e1/K-Medoids_Clustering.gif



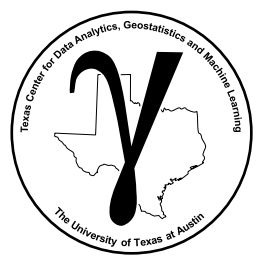
Other Clustering Methods

DBSCAN – Density-based Spatial Clustering of Applications with Noise, hierarchical agglomerative method

- Density-based clustering – cluster is a maximal set of density connected points, grows over complicated connected patterns
- Quite sensitive to the parameters:
 - maximum radius of the neighbourhood
 - minimum number of points to form a cluster

Workflow

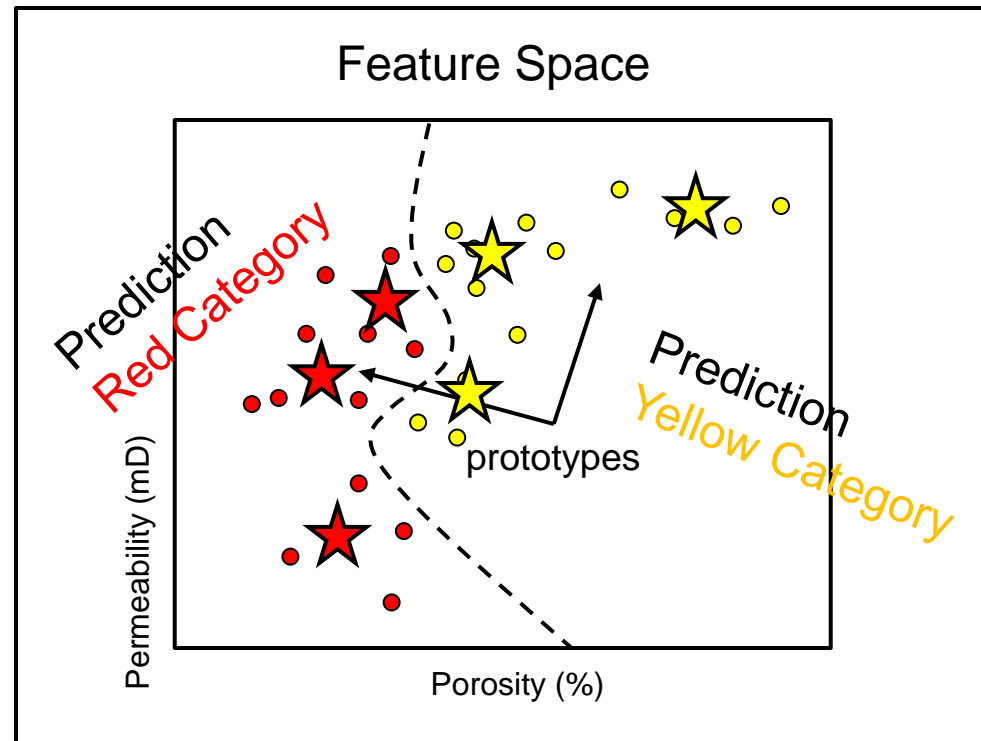
- Visit sample data at random
- Check neighbourhood:
 - if enough points, start a new cluster or add to an adjacent cluster
 - if not enough points, assign as an outlier
- Stop when all sample data are visited
- Advantages: unequal, non-spherical clusters and outliers



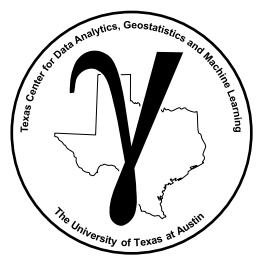
K-means for Classification

Not a clustering method, but based on the k-means clustering method

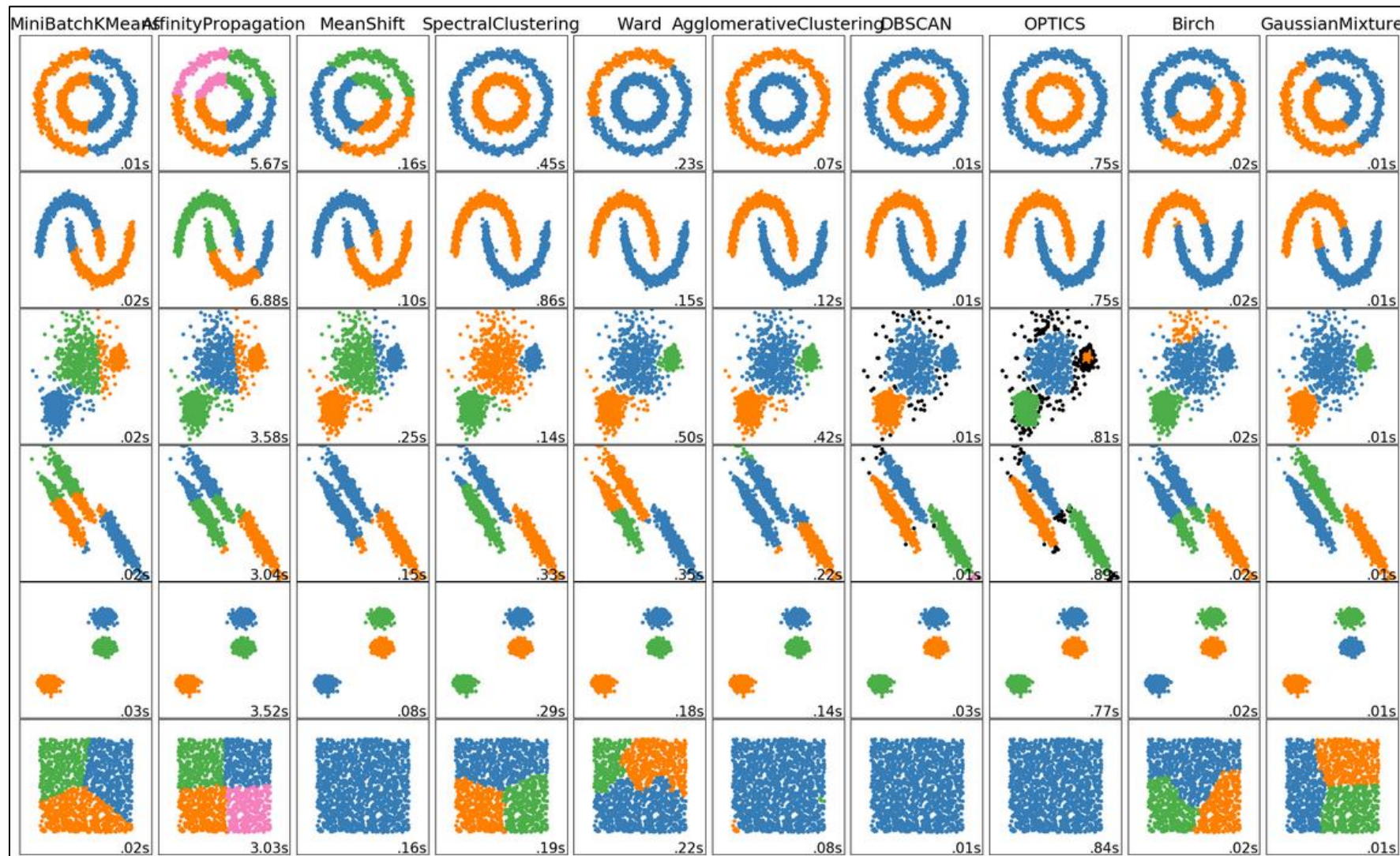
- Given a set of labelled data (e.g. facies labels included with continuous features)
- Apply k-means clustering in each category
- Predictions in features space assigned to the nearest (Euclidean) distance prototype



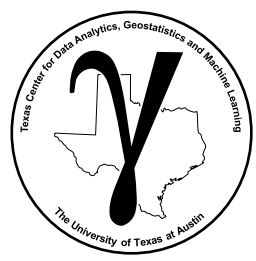
Schematic of K-means classification.



Other Clustering Methods



Comparison of clustering methods from https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html



PGE 383 Subsurface Machine Learning

Lecture 7: Clustering

Lecture outline:

- **Prototype Methods**
- **K-means Clustering**
- **K-means Clustering Hands-on**
- **Other Clustering Methods**