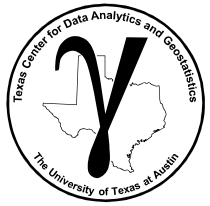
Geostatistics and Machine Learning

Clustering



- **Prototype Methods**
- **K-means Clustering**

Introduction

Data Analytics

Inferential Methods

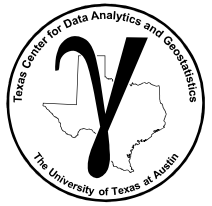
Predictive Methods

Advanced Methods

Conclusions

Michael Pyrcz, The University of Texas at Austin

Geostatistics and Machine Learning
Clustering



- **Prototype Methods**

Introduction

Data Analytics

Inferential Methods

Predictive Methods

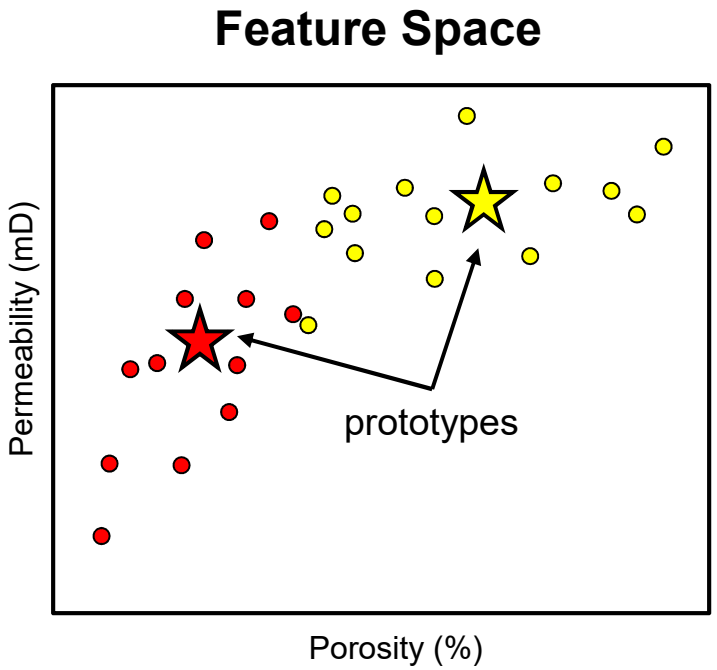
Advanced Methods

Conclusions

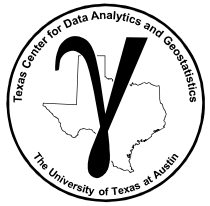
Michael Pyrcz, The University of Texas at Austin

Prototype Methods

- Represent the training data with set of points in the feature space.
- Prototypes are typically not actual samples
- Training data often assigned to the nearest (Euclidean) distance prototype



Geostatistics and Machine Learning
Clustering



- **K-means Clustering**

Introduction

Data Analytics

Inferential Methods

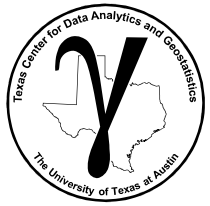
Predictive Methods

Advanced Methods

Conclusions

Michael Pyrcz, The University of Texas at Austin

K-means Clusters



General Comments on K-means Clustering

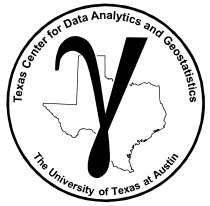
Clustering - of training data in feature space to identify similar cases / distinct subsets in the multivariate data space.

Unsupervised Learning - the training data are not labeled and are assigned

Prototype Method - represents the training 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 training sample are updated to the nearest prototype, then the prototypes are adjusted to the centroid of their assigned training data, repeat until convergence.

K-means Clusters



General Comments on K-means Clustering

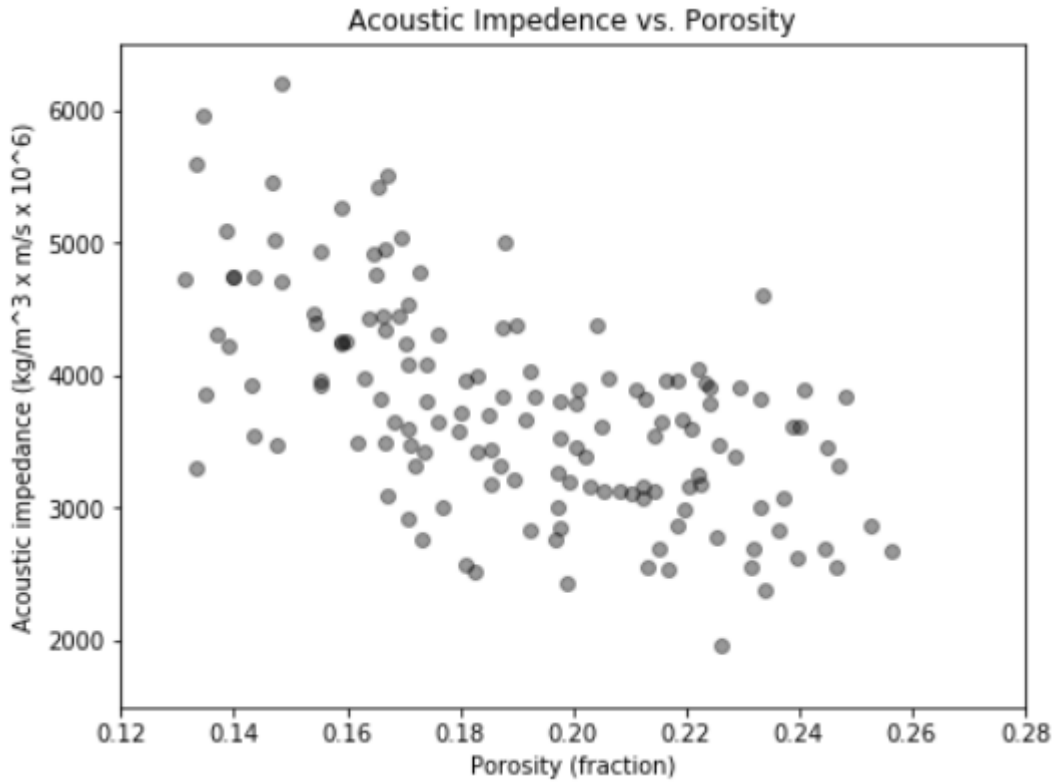
Feature Weighting - the procedure depends on the 'distance' between training samples and prototypes in feature space. If the features have significantly different magnitudes, the feature(s) with the largest magnitudes and ranges will dominate the process. One approach is to sandardize the variables. Also, by-feature weighting may be applied.

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 training data with porosity and acoustic impedance, find ‘K’ facies that:

- segment the porosity and acoustic impedance feature space.

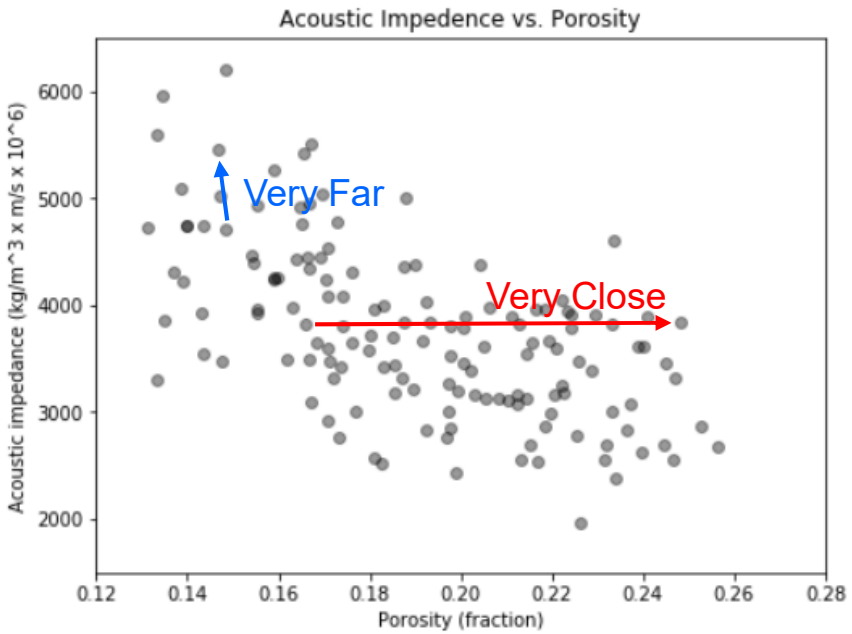


K-means Clustering Feature Normalization

We will require a measure of similarity:

- consider Euclidean distance with the original units

$$d = \sqrt{\Delta Por^2 + \Delta AI^2}$$



For dissimilar units we require normalization

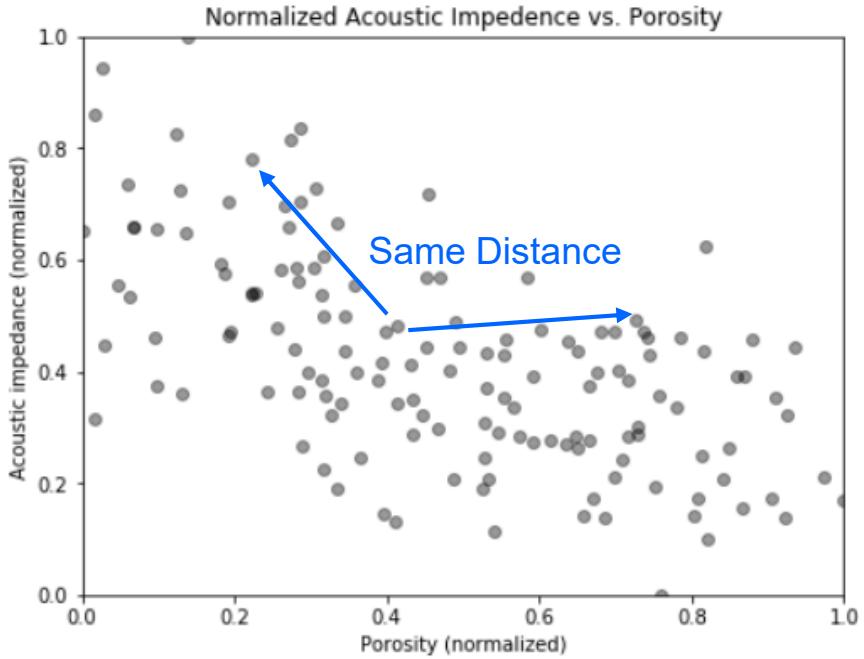
– similar magnitude and range over all features.

K-means Clustering Feature Normalization

We will require a measure of similarity:

- consider Euclidean distance with the normalized features

$$d = \sqrt{\Delta N(Por)^2 + \Delta N(AI)^2}$$



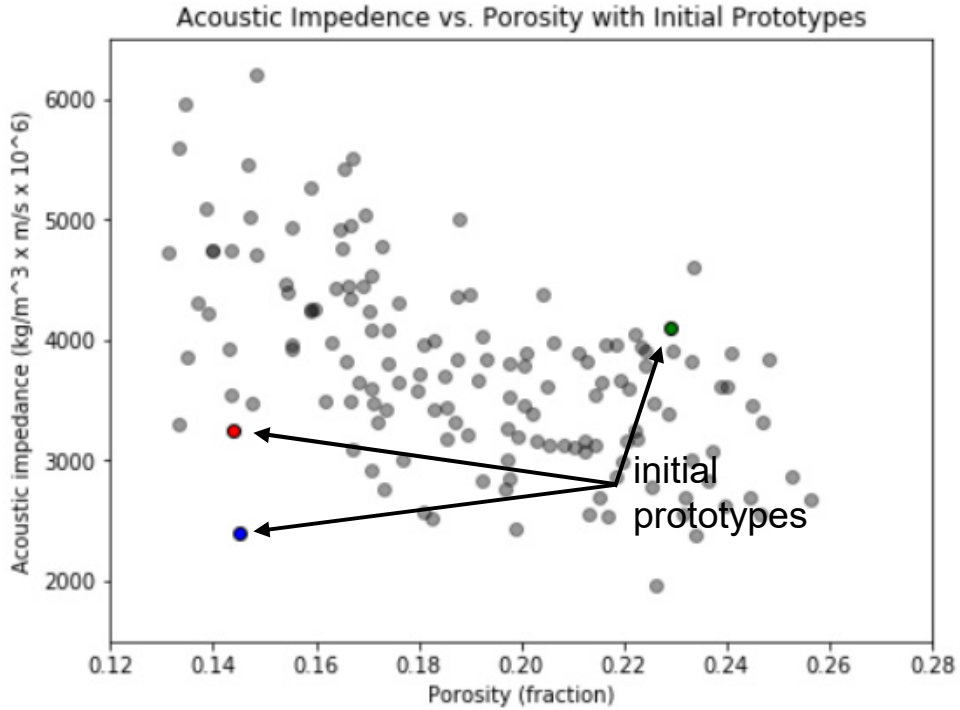
For dissimilar units we require normalization

– similar magnitude and range over all features.

K-means Clustering Workflow

Assign K prototypes in the feature space:

- random assignment (random value between min and max)



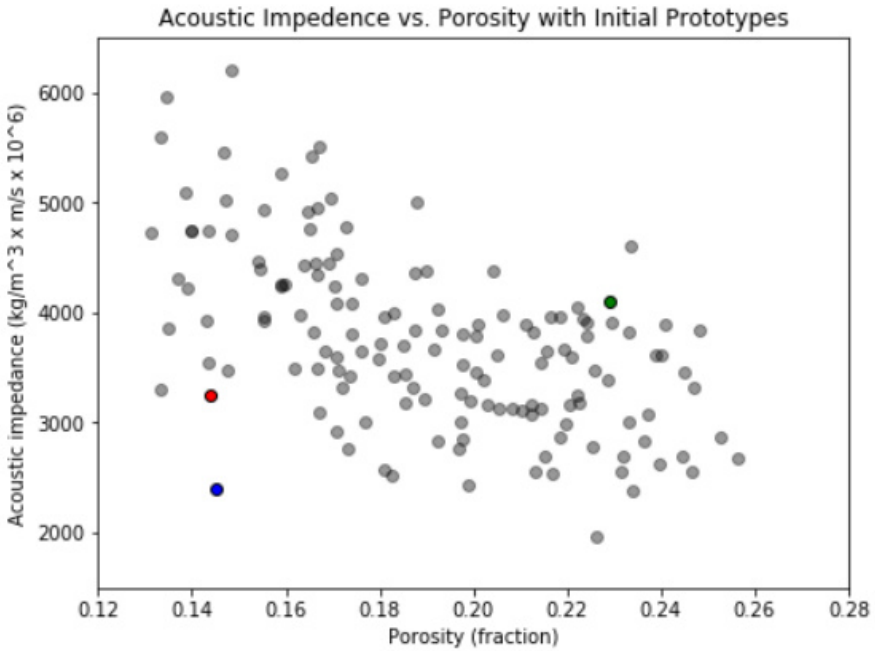
Note: the initial prototypes could be poor choices

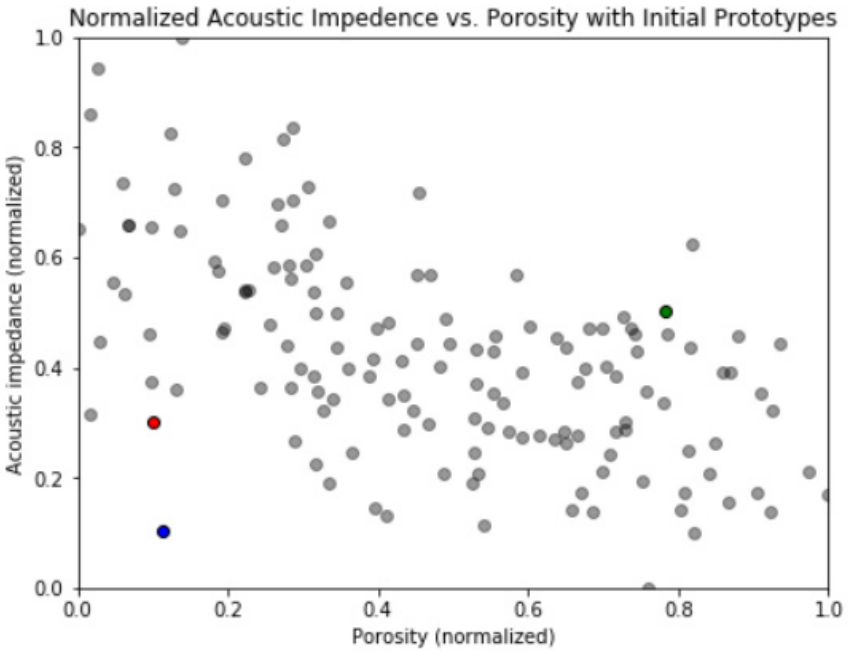
– clustered, outside the training data etc.

K-means Clustering Workflow

Assign K prototypes in the feature space:

- as a reminder this is happening in the normalized features space

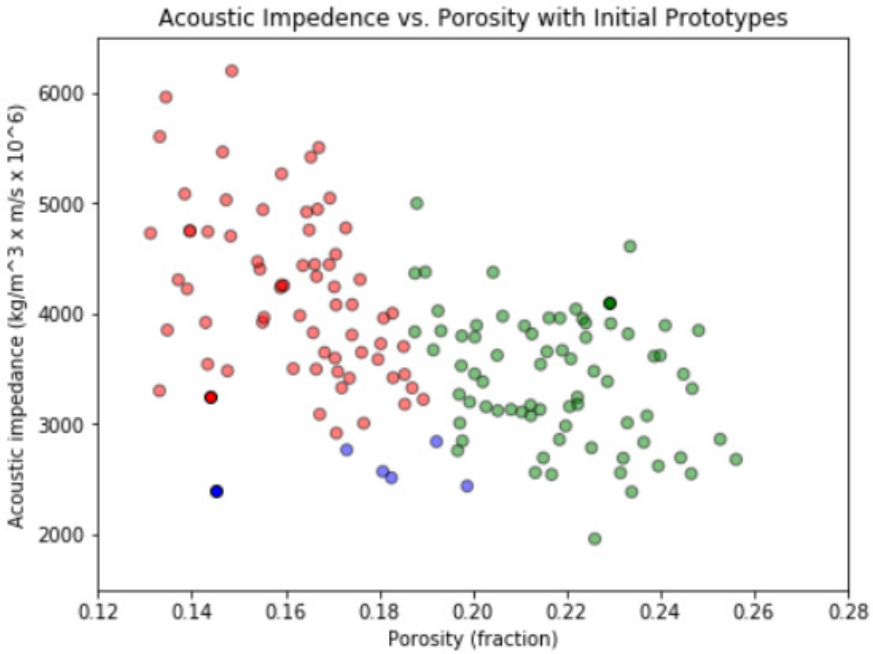


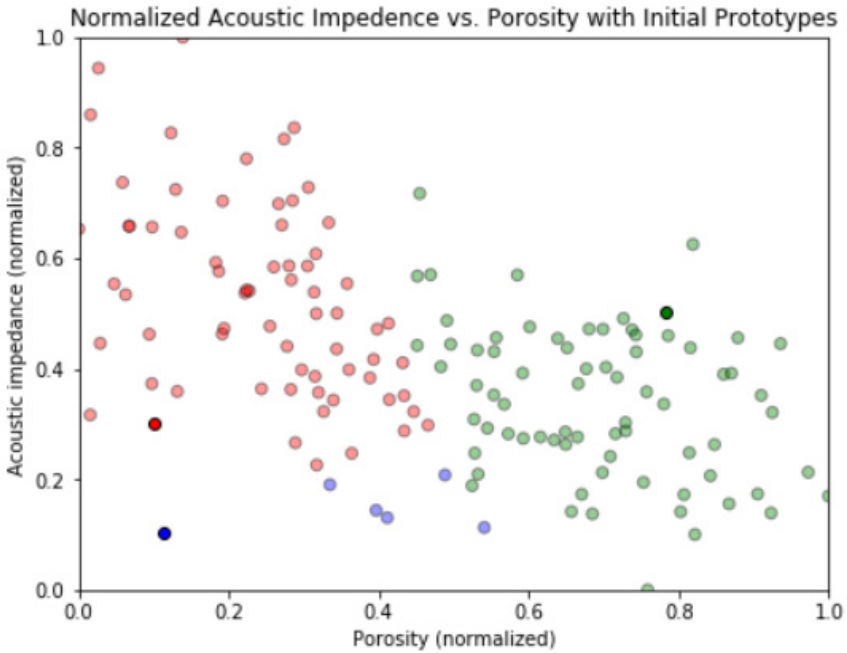


K-means Clustering Workflow

Assign training data to the nearest prototype:

- we apply nearest Euclidian distance with the normalized features



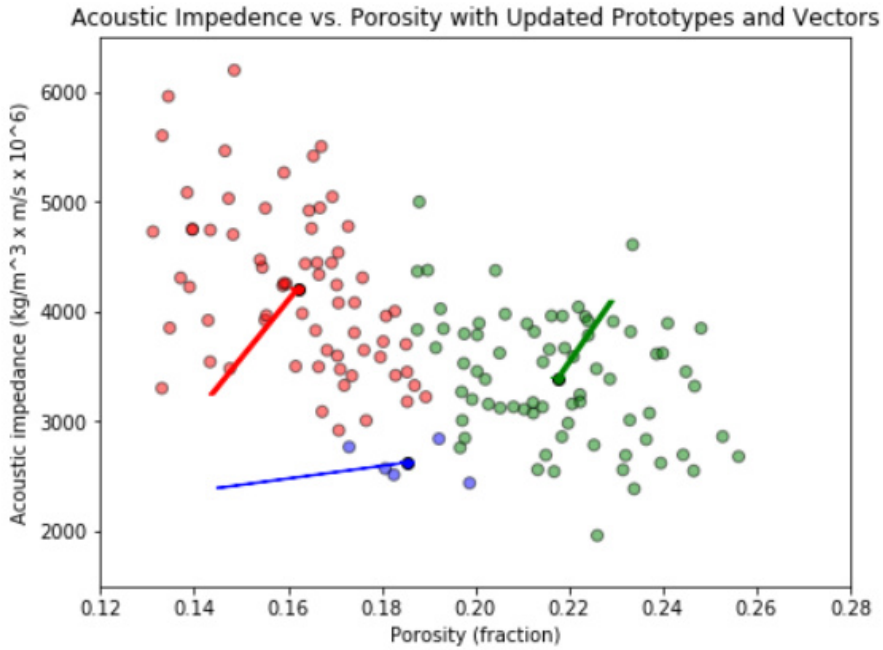


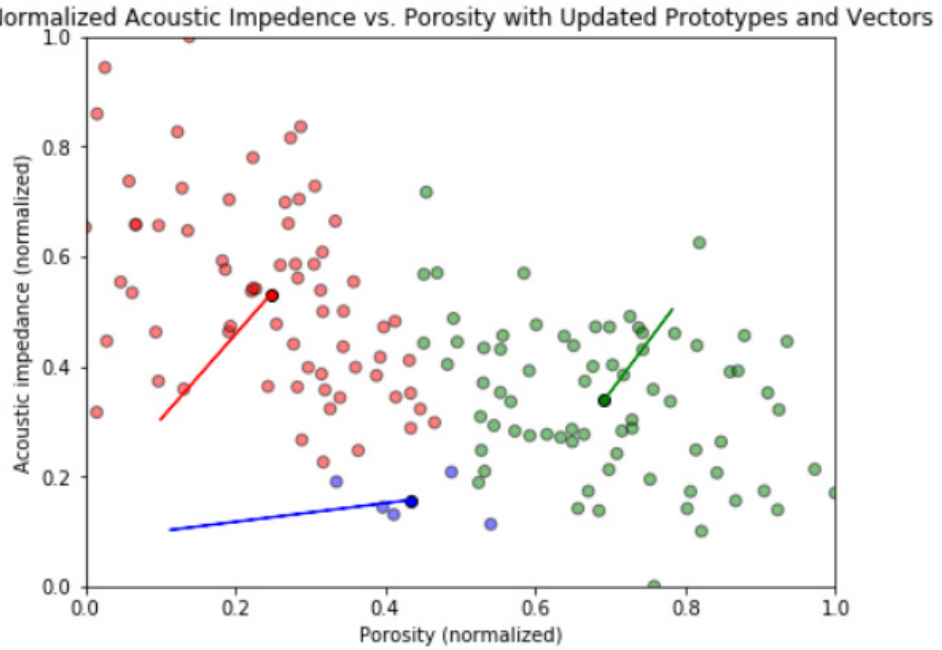
- Note, these initial clusters are not very good

K-means Clustering Workflow

Update the prototypes to the centroids of the assigned training data:

- vectors are included to show the update of the prototypes



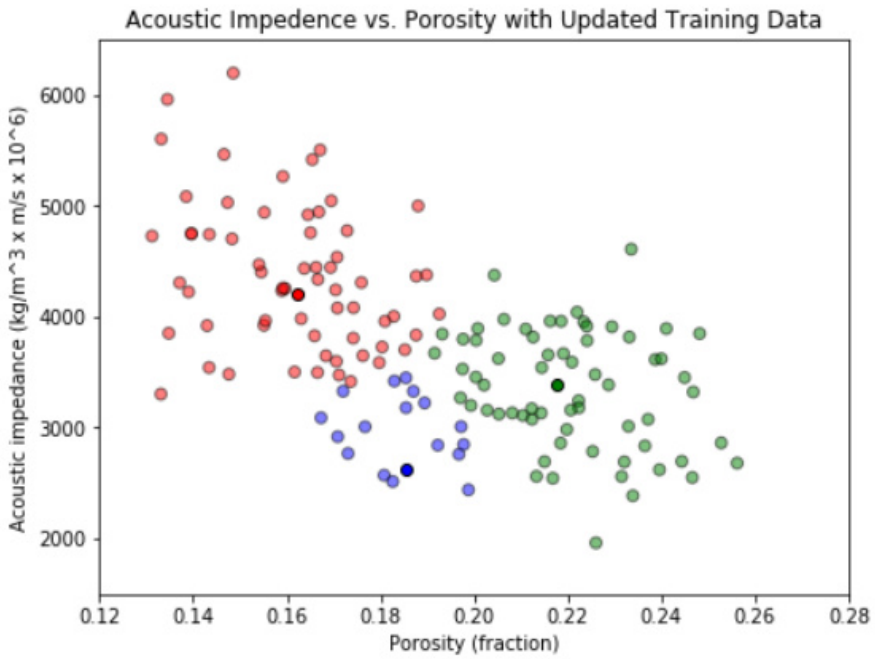


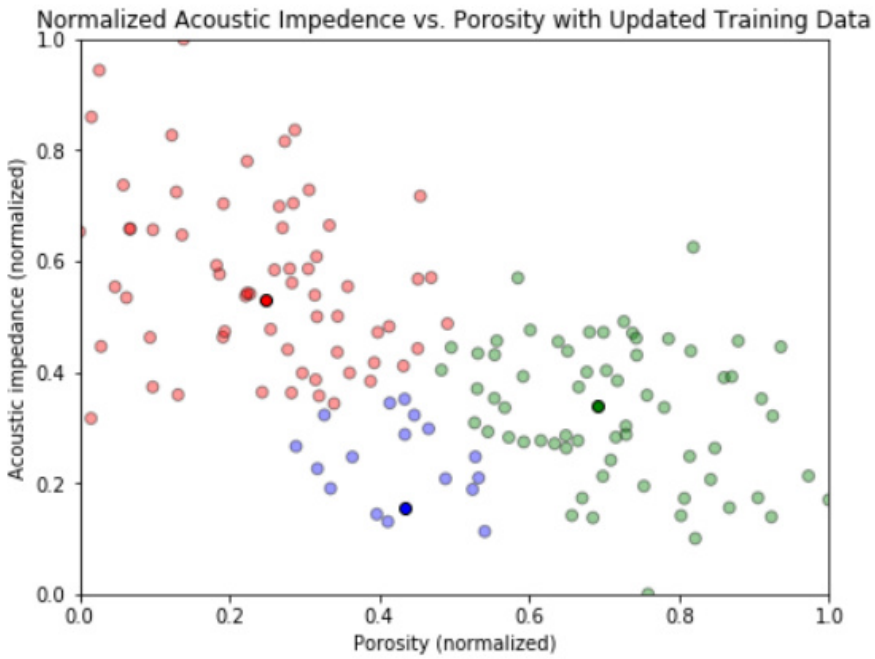
- The prototypes have improved.

K-means Clustering Workflow

Update the training data assignment to the nearest updated prototypes:

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



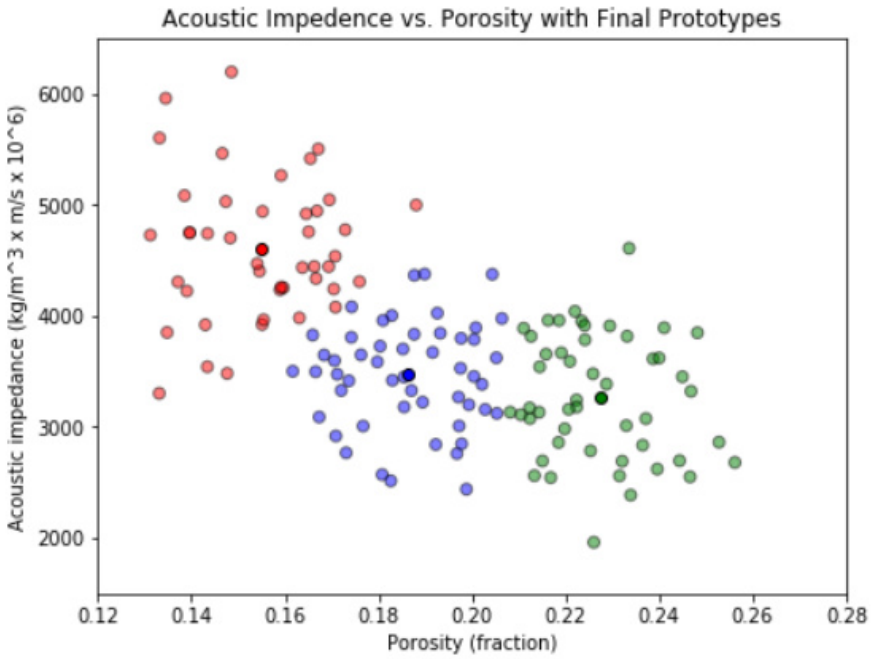


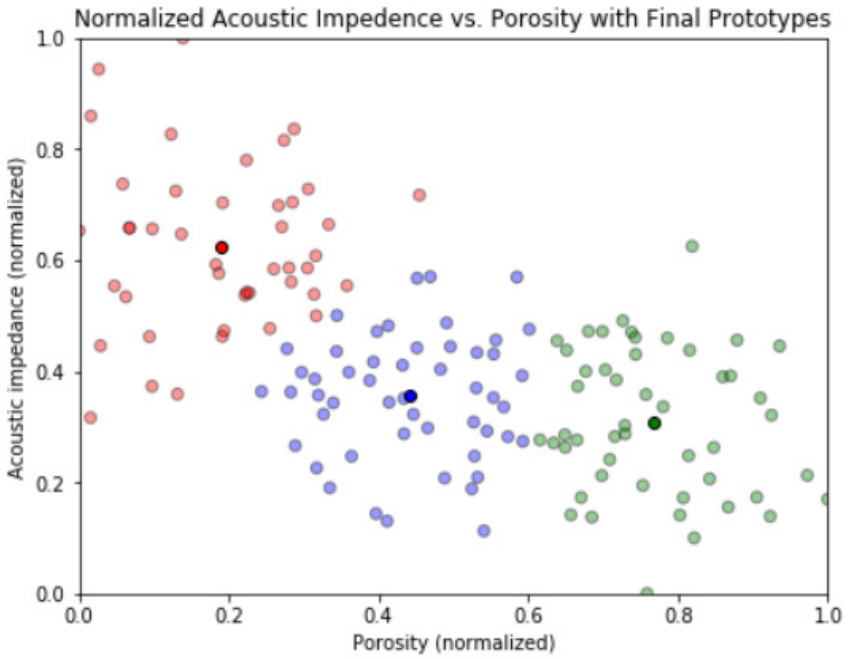
- The prototypes have improved.

K-means Clustering Workflow

Iterate until the centroid stop moving and the assignments stabilize:

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

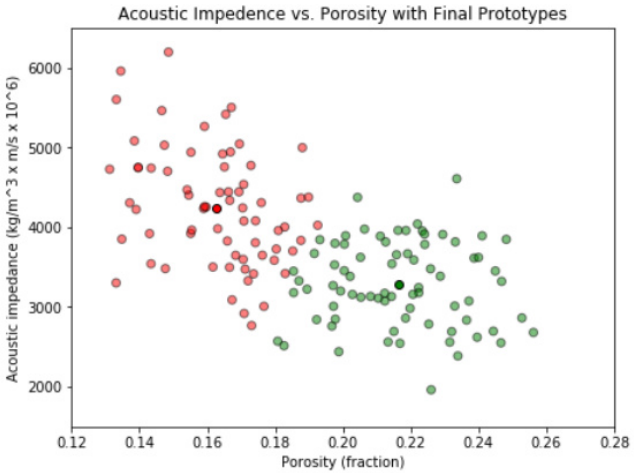


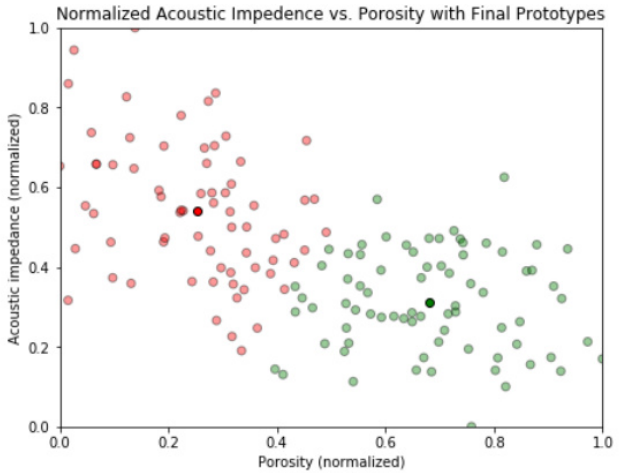


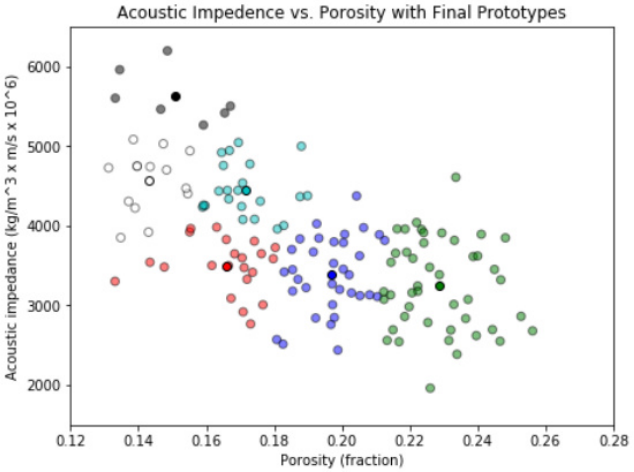
K-means Clustering Workflow

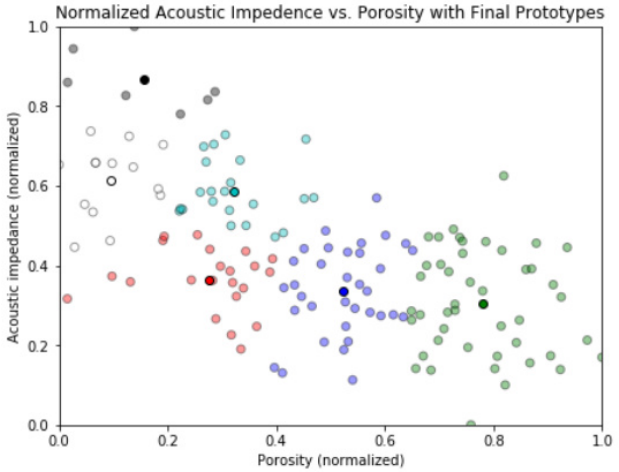
The number of K clusters is an important decision:

- examples with K = 2 and 7









K-means Clustering Demonstration

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

Subsurface Data Analytics

K-means Clustering

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | GitHub | Website | GoogleScholar | Book | YouTube | LinkedIn

PGE 383 Exercise: K-means Clustering for Subsurface Data Analytics in Python

Here's a simple workflow, demonstration of K-means clustering for subsurface modeling workflows. This should help you get started with building subsurface models that integrate uncertainty in the sample statistics.

The code was modified from the great tutorial provided by Ben Keen. The original tutorial is avaiable at here. I really liked the way Ben decomposed K-means clustering into the primary steps of:

- assign initial random prototype with labels
- assign samples to the nearest prototype label
- update prototype based on centroids of samples belonging to this prototype
- iterate until no sample assignments change

k-Means Clustering

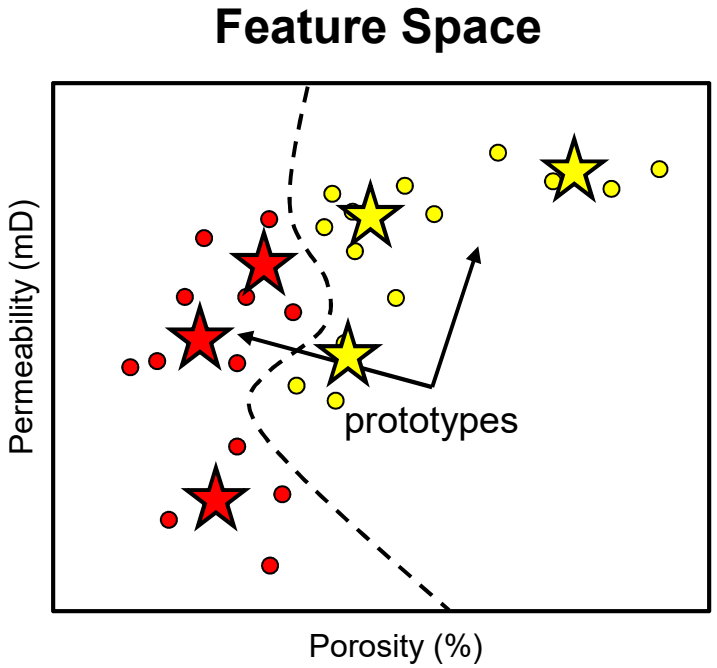
The K-means clustering approach is primaryly applied as an unsupervised method for classification:

- Prototype Method - represents the training data with number of synthetic cases in the features space. For K-means clsutering we assign K prototypes.
- Iterative Solution - the initial prototypes are assigned randomly in the feature space, the labels for each training sample are updated to the nearest prototype, then the prototypes are adjusted to the centroid of their assigned training data, repeat
- Unsupervised Learning - the training data are not labeled and are assigned K labels based on their proximity in the feature space. The idea is that similar things should be in the same category.
- Feature Weighting - the procedure depends on the 'distance' between training 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 process. One approach is to sandardize the variables. Also, by-feature weighting may be applied.
- Supervised Learning Variant - applies multiple prototypes in each category to then inform a decision boundary.

File SubsurfaceDataAnalytics_clustering.ipynb at https://git.io/fjWQn.

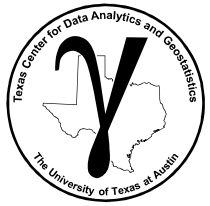
K-means for Classification

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



Geostatistics and Machine Learning

Clustering



- **Prototype Methods**
- **K-means Clustering**

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

Michael Pyrcz, The University of Texas at Austin