Unsupervised Learning

Chantri Polprasert

CPDSAI

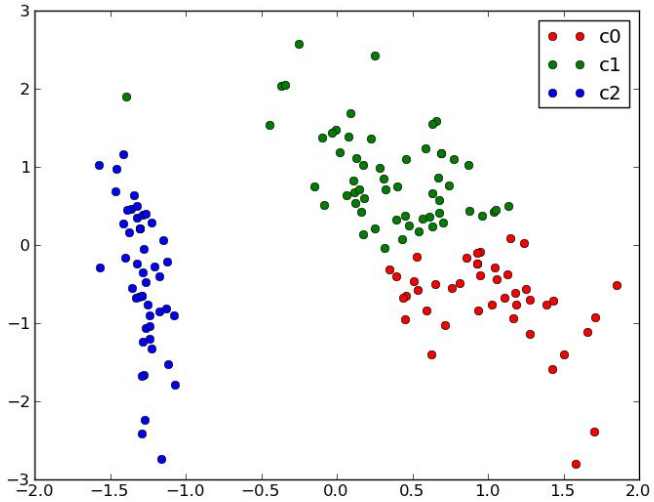
Recall

- The goal in *supervised* setting is to predict ***Y*** (either continuous or categorical) using *p* features $\boldsymbol{X_1, X_2,..., X_p}$ measured on n observations.
- In *unsupervised* learning, we do not have a response(target) variable Y so the goal is to discover interesting things (patterns, clusters, outliers) from the measurements on $\boldsymbol{X_1, X_2,..., X_p}$

Goal of Unsupervised Learning

The goal is to discover interesting things about the measurements:

- Is there an informative way to visualize the data?
- Can we discover subgroups among the variables or among the observations?
- Can we discover interesting patterns, relationships, or associations among items or variables in a dataset?
- Can be treated as a part of the EDA process



KMeans cluster assignments on 2D PCA iris data

Practical Applications

Techniques for unsupervised learning are of growing importance in a number of fields. For instance:

- Search for subgroups among breast cancer patients in order to gain a better understanding of the disease
- Understand customer buying behavior to identify groups of shoppers with similar browsing and purchase histories (for targeted ads)
- Representing a high-dimensional data set (e.g. gene expression) in smaller dimensions

Practical Applications

Techniques for unsupervised learning are of growing importance in a number of fields. For instance:

- Search for subgroups among breast cancer patients in order to gain a better understanding of the disease (**clustering**)
- Understand customer buying behavior to identify groups of shoppers with similar browsing and purchase histories (for targeted ads) (**association**)
- Representing a high-dimensional data set (e.g. gene expression) in smaller dimensions (**dimensionality reduction**)

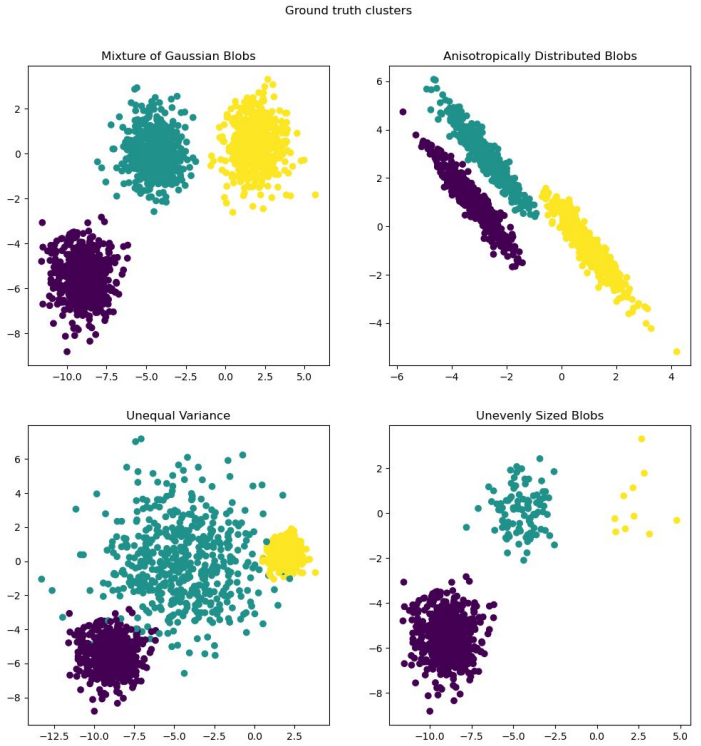
Challenges and Advantages of Unsupervised Learning

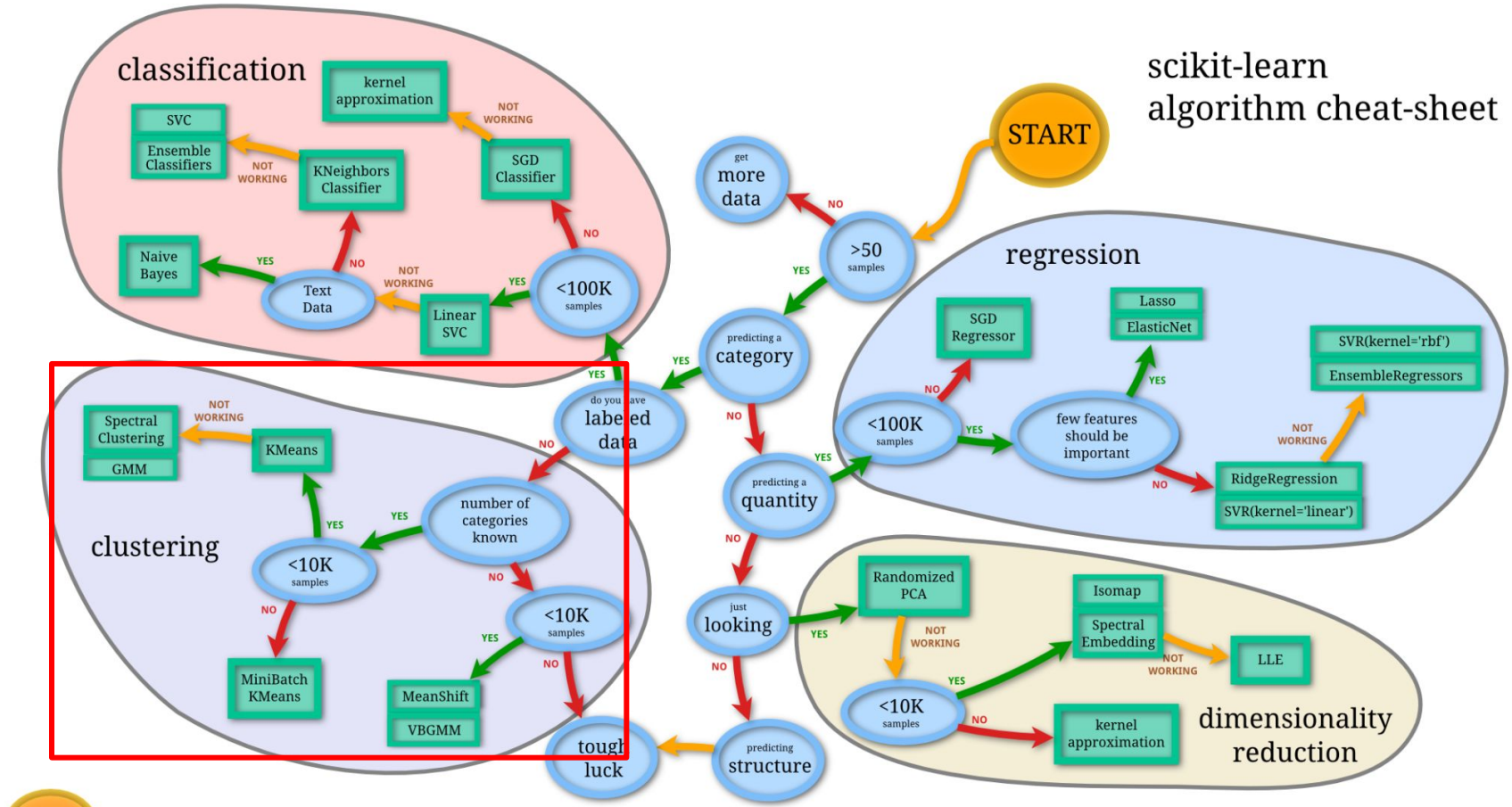
- **Challenge**: Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- **Advantage** It is often easier to obtain unlabeled data—from a lab instrument or a computer—than labeled data, which can require human intervention

Types of Unsupervised Learning

Unsupervised learning is utilized for three main tasks

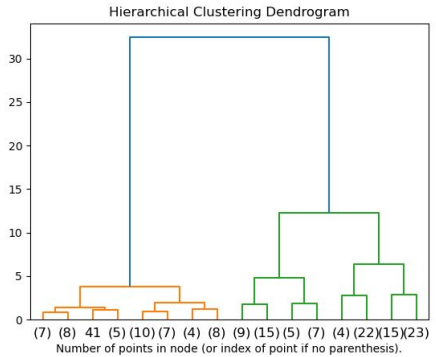
1. Clustering
2. Dimensionality reduction
3. Association

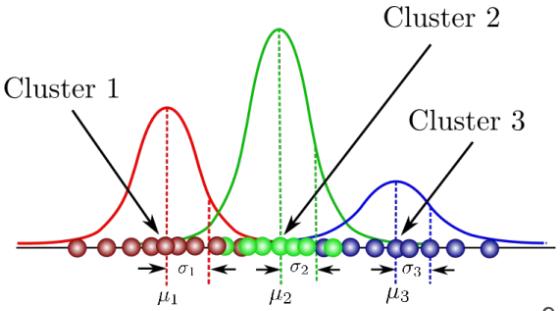



scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
NO
predicting a quantity
YES
NO
just looking
YES
NO
predicting structure
tough luck
do you have labeled data
YES
NO
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
YES
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
Back

Clustering

- The goal of clustering is to find groups (or clusters) such that all observations are
  - more **similar** to observations inside their group and
  - more **dissimilar** to observations in other groups.
- Types of clustering algorithms include:
  - Hierarchical: Agglomerative (bottom-up) or Divisive (top-down)
  - Partitioning (e.g. k-means “hard” clustering)
  - Probabilistic Clustering (e.g. Finite Mixture models)





K-means Clustering

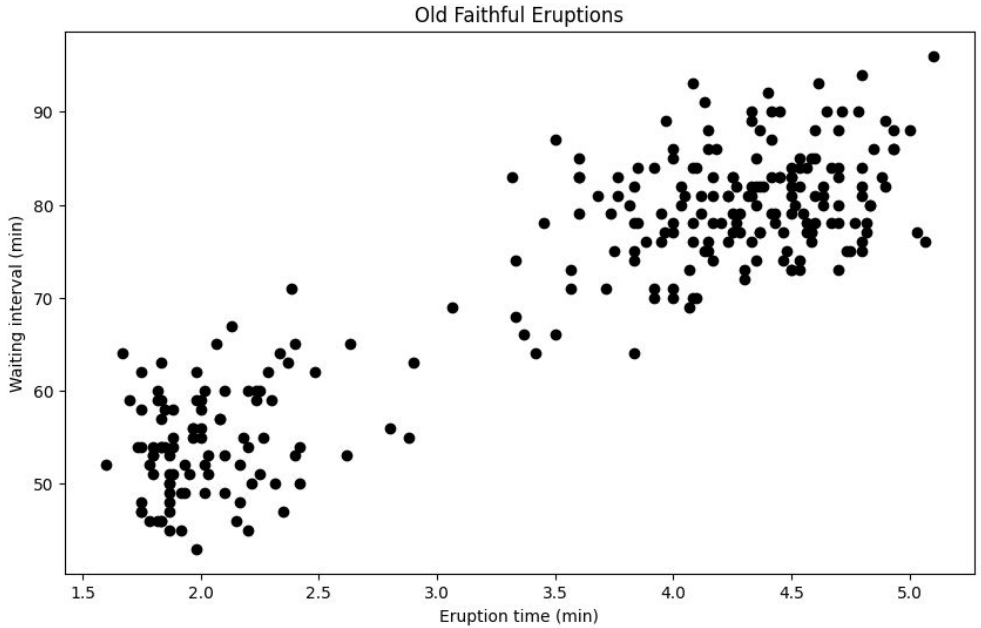
- *k*-means clustering is a popular method that requires the user to provide ***k***, the number of groups they are looking for
- In this algorithm, each observation will belong to the cluster whose mean is closest.

Algorithm

1. Randomly select (the number of groups) points in your data. These will serve as the first centroids
2. Assign all observations to their closest centroid (in terms of Euclidean distance). You now have ***k*** groups.
3. Calculate the means of observations from each group; these are your new centroids.
4. Repeat 2) and 3) until nothing changes anymore (each loop is called an *iteration*).

Old Faithful

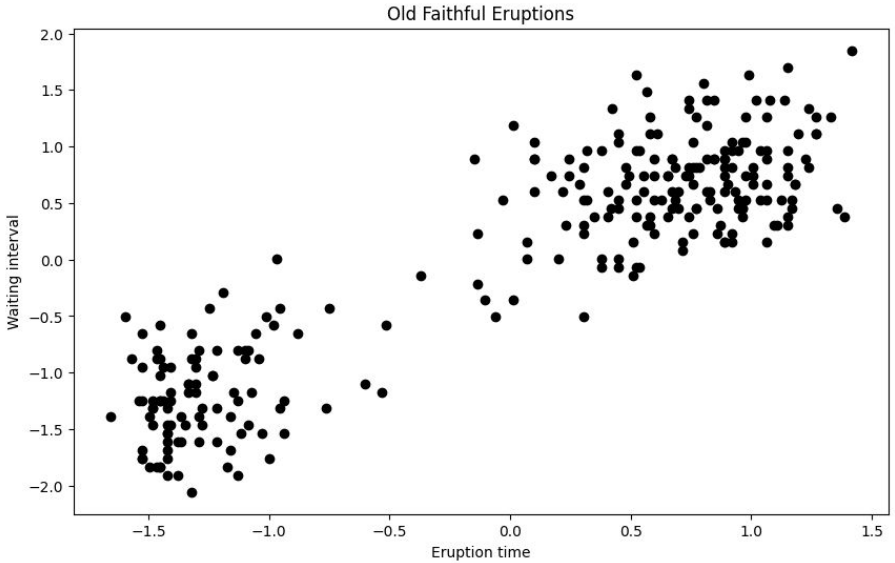
Old Faithful is a cone geyser in Yellowstone National Park in Wyoming, United States. It is a highly predictable geothermal feature and has erupted every 44 minutes to two hours since 2000



How many clusters?

Old Faithful

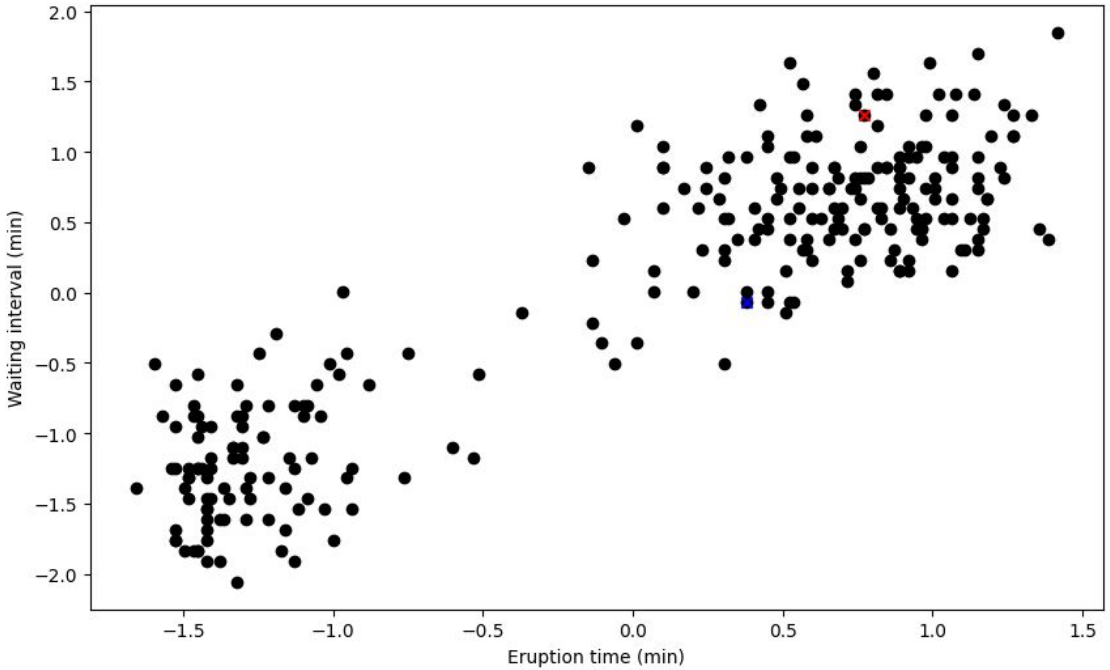
Old Faithful is a cone geyser in Yellowstone National Park in Wyoming, United States. It is a highly predictable geothermal feature and has erupted every 44 minutes to two hours since 2000



Normalized version

Step1 (k=2)

Randomly select 2 centroids

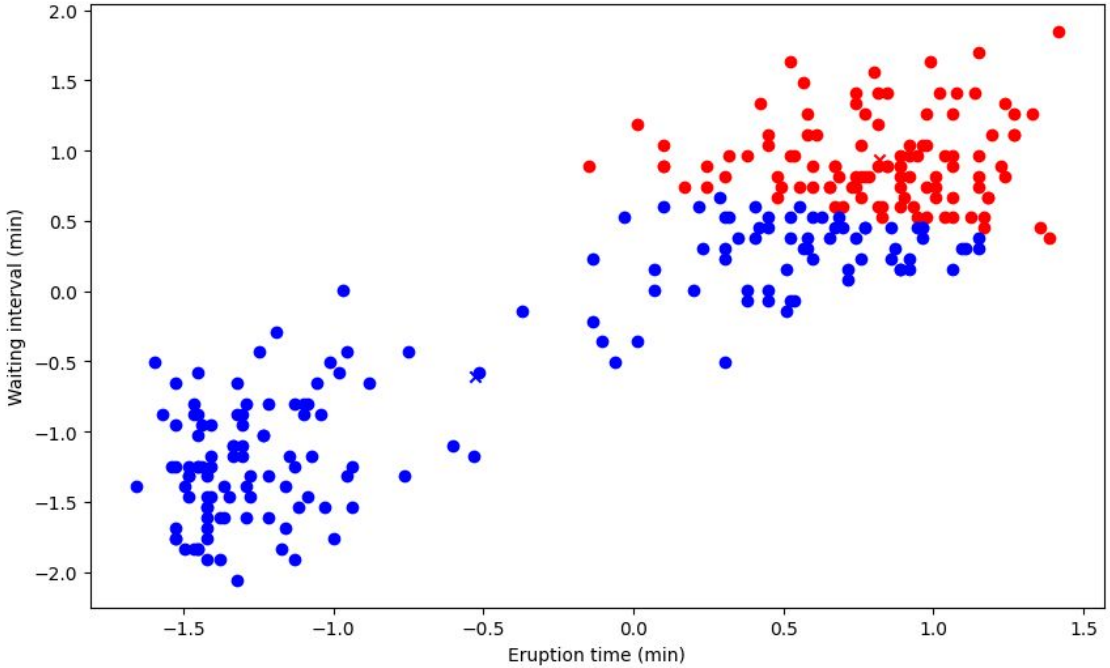


Centroids

	eruptions	waiting
101	0.771736	1.260353
228	0.376747	-0.066106

Step2

Assign observations to their closest centroid

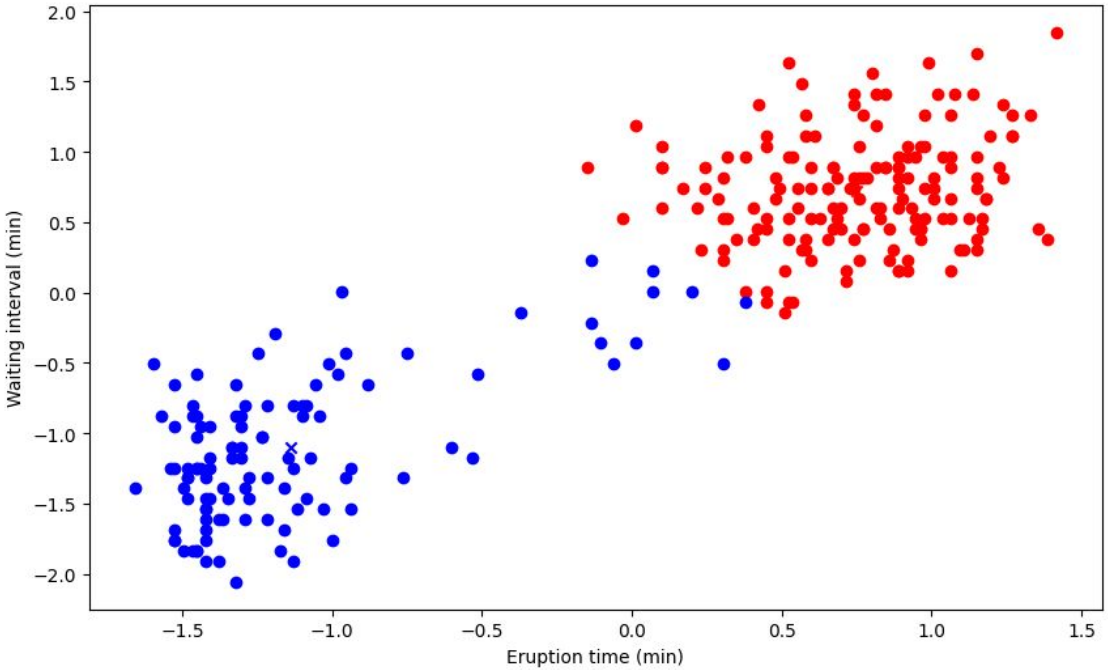


Centroids

Cluster	eruptions	waiting
1	0.817084	0.934592
2	-0.529867	-0.606069

Step3

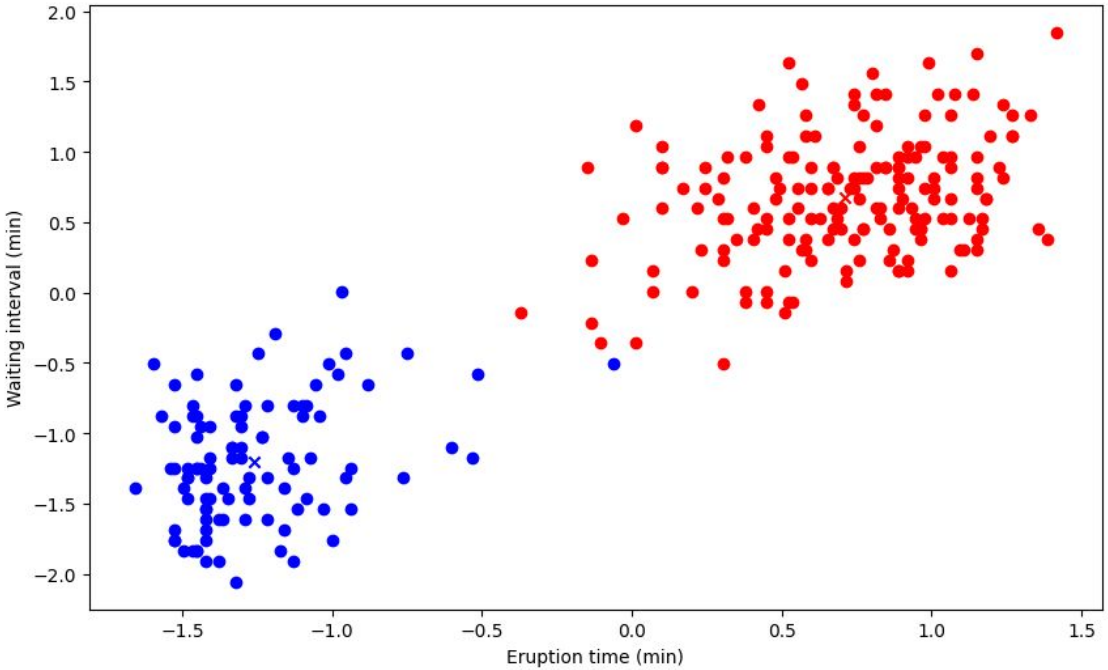
Recalculate the means/centroids of each group



Cluster	eruptions	waiting
1	0.751248	0.725635
2	-1.140784	-1.101891

Step4

Repeat 2) and 3) until nothing changes anymore



Cluster	eruptions	waiting
1	0.709703	0.676745
2	-1.260085	-1.201567

Why does -means work?

1. Randomly select (the number of groups) points in your data. These will serve as the first centroids.
2. Assign all observations to their closest centroid (in terms of the Euclidean distance). You now have ***k*** groups.
3. Calculate the means of observations from each group (These are your new centroids)
4. Repeat 2) and 3) until nothing changes anymore (2. and 3. are recursively finding the minimum within-group sum of squared distances between points and their centroids.)

Comments

Advantages:

- Computationally efficient, even for large data sets.
- Only n×k matrices needed
- Relatively straightforward
- Often provides clearer groups than other methods (e.g. Hierarchical Clustering)

Drawbacks:

- Stochastic, i.e. random (as opposed to deterministic)
- Can return local optimal rather than global
- not for categorical input features
- Groups will be found no matter what.

Conclusions

- **Definition**: Unsupervised learning is a machine learning approach that analyzes unlabeled data to uncover hidden patterns and structures without prior guidance.
- **Key Techniques**: Common methods include clustering (e.g., K-means) for grouping similar data points or dimensionality reduction (e.g., PCA) for simplifying datasets while retaining essential information.
- **Applications**: Used in various fields such as customer segmentation, anomaly detection, image and text clustering, and genomics for pattern recognition and insights.
- **Advantages**: Enables exploratory data analysis, revealing insights from complex datasets without the need for labeled data, and is scalable for large volumes of information.
- **Challenges**: Results can be difficult to interpret and evaluate due to the absence of predefined labels, making it harder to assess model performance compared to supervised learning.

References

- https://scikit-learn.org/1.5/unsupervised_learning.html
- https://www.kaggle.com/code/niteshhalai/old-faithful-data-visualisation-and-modelling
- https://neptune.ai/blog/clustering-algorithms