EXPLORE AI

Clustering Pt 2

GMM clustering + geopandas!

14 March 2023

Objectives

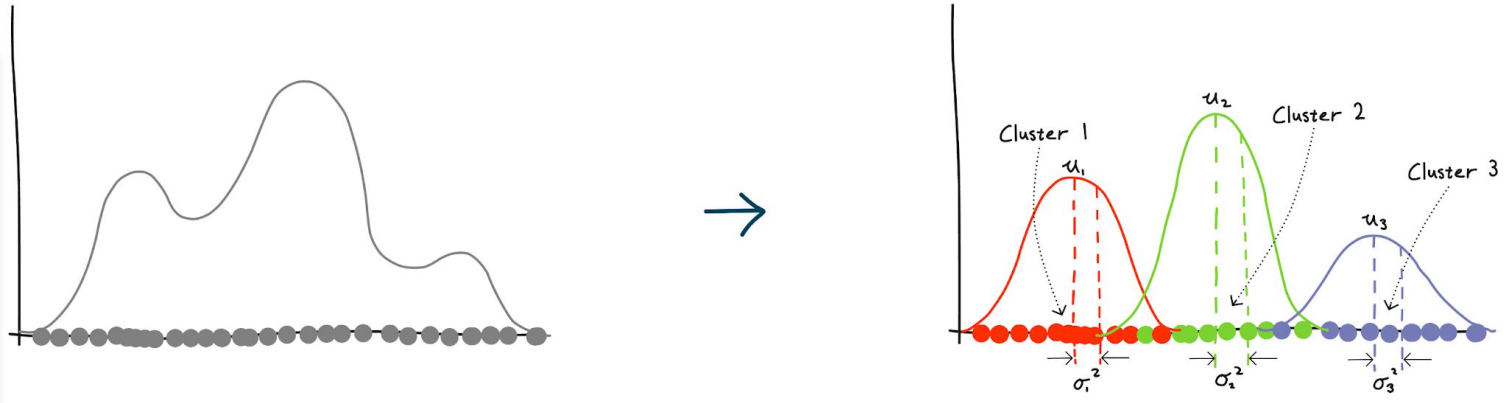
By the end of this session you should be able to...

- Distinguish between hard and soft clustering methods
- Understand the Gaussian Mixture Models algorithm
- Understand the logic/steps behind the code to apply clustering to a dataset in Python
- Recognise geometry (GeoSeries) parts of a GeoDataFrame

Gaussian Mixture Models (GMM)

GMM is an example of a **soft clustering** method, which means that instead of being in fixed clusters, points are assigned **probabilities** of being in each cluster

Soft clustering techniques keep all possibilities of cluster assignment - good for when clusters may overlap



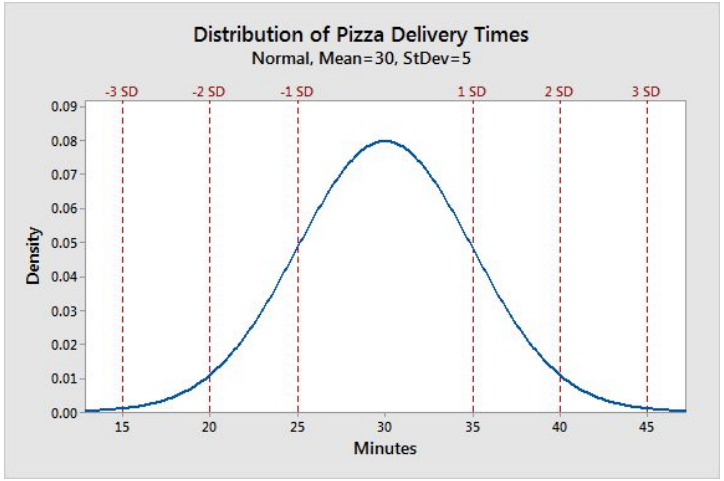
Gaussian Mixture Models (GMM)

Mixture of Gaussian distributions where we attempt to fit a Gaussian distribution to each cluster.

Assumption: each cluster can be characterised by a **mean** and a **variance** - but these are **unknown**.

- Objective of the GMM process: approximate these parameters as closely as possible

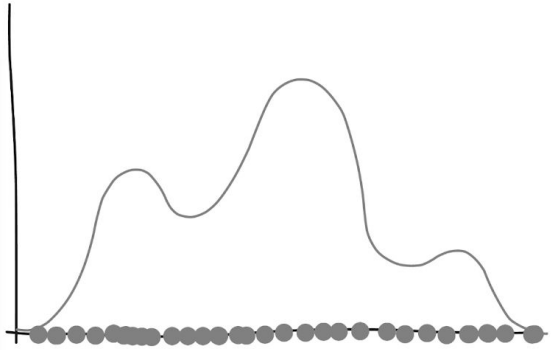
→ We also assume that the data was generated in Gaussian way, i.e., the data is normally distributed

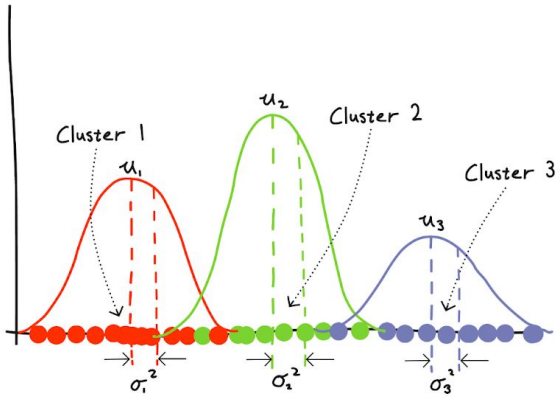


Gaussian Mixture Models (GMM)

Expectation-Maximisation Algorithm

1. Choose the number of clusters
2. Randomly initialise all Gaussian distributions
3. **Expectation**: Compute the likelihood
   a. Probability each data point belongs to each cluster
   b. "how likely is each data point under each Gaussian?"
4. **Expectation**: Compute the posterior
   a. Probability of each Gaussian/cluster
   b. "how likely is each Gaussian model for each data point?"
5. **Maximisation**: Update cluster assignments and Gaussian parameters
6. Repeat until stopping condition



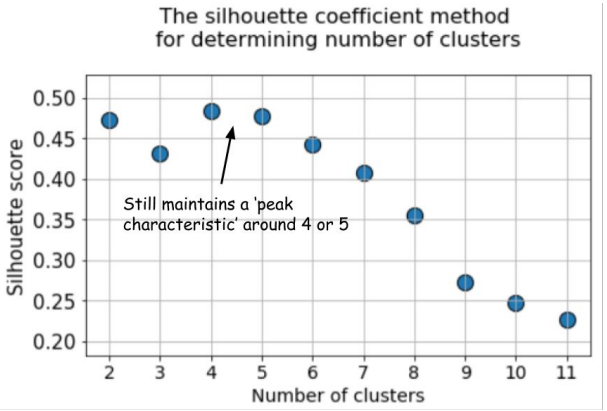


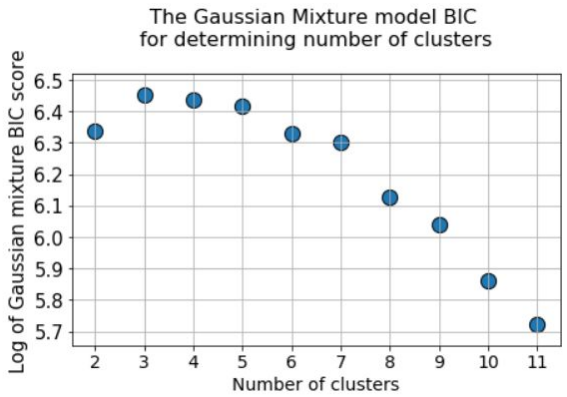
Gaussian Mixture Models (GMM)

Note: this is beyond the course content!

How do we decide on the number of clusters, K?

- Silhouette method
  - looks at how compact each cluster is, and how well-separated it is from others
- Bayesian Information Criterion (BIC)
  - a regularisation technique (often used in linear regressions!)
  - Penalizes a large number of Gaussians and tries to keep the model simple enough to explain the given data pattern





Source: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Gaussian Mixture Models (GMM)

Advantages

- Provides probability estimates rather than hard assignments
- Clusters can be of any ellipsoidal shape, not just circular ones.

Disadvantages

- The number of clusters K needs to be specified before
- Requires assumption of Gaussian (Normal) distributions across dimensions.

To the notebook!

Objectives

By the end of this session you should be able to...

- Distinguish between hard and soft clustering methods
- Understand the Gaussian Mixture Models algorithm
- Understand the logic/steps behind the code to apply clustering to a dataset in Python
- Recognise geometry (GeoSeries) parts of a GeoDataFrame