NOTES ON CLUSTERING

Clustering is an unsupervised classification method in Machine Learning. There are various ways to "cluster" data depending on distance measures (euclidean, manhattan, etc.), centroid formulation, etc. The most common forms of clustering analysis are **K-means**, **KNN**, and **Hierachical Clustering**. Less known and possibly under-utilized are **Network Clustering** (aka "graphical network analysis" aka "community discovery"), **Bisectioning K-means**, and **Gaussian Mixture Models** (GMM)...

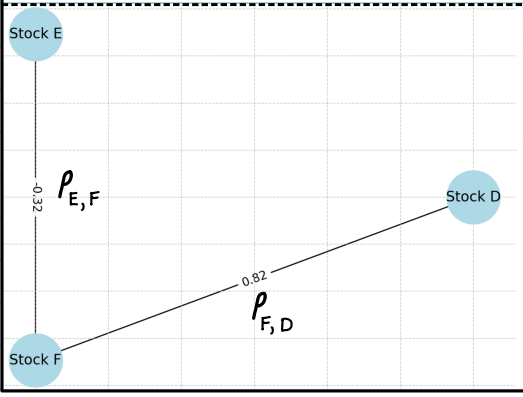
~ THE COMMON GOAL IS TO CATEGORIZE UNLABLED DATA BASED ON SIMILARITIES OBSERVED IN THE DISTANCES BETWEEN DATA POINTS ~

I Network Clustering aka "Graphical Network Analysis"

- potentially useful for moding expected returns & factor exposure

Step-by-Step Example:

1. **Stocks Selection**: Assume we have three stocks: A, B, and C.
2. **Price Data & Returns**:
   - Suppose we have closing prices over 5 days for simplicity.
   - We compute the daily returns (percentage change from one day to the next).
3. **Calculate Correlations**:
   - We calculate the correlation coefficients between the returns of each pair of stocks.
4. **Network Construction**:
   - Each stock is a node.
   - An edge between two stocks is drawn if their correlation exceeds a certain threshold (e.g., 0.5).
5. **Visualization**:
   - We visualize this network to see which stocks are closely related.



- **Nodes**: Represent the stocks (D, E, F).
- **Edges**: Only the significant correlations (above the threshold of 0.3) are shown:
  - Stock D and Stock F with a correlation of 0.82.
  - Stock E and Stock F with a correlation of -0.32.
- **Edge Labels**: Display the strength of the correlations.

This network offers insights into potential investment strategies:

- **Diversification**: Stock E, showing negative correlations with Stock F, might provide diversification benefits within a portfolio that includes Stock F.
- **Risk Management**: Understanding these correlations helps in managing the risks associated with potential market movements affecting these stocks differently.

II Bisectioning K-Means

- a form of hierachical clustering
  - Agglomerative clustering is "bottom up"
  - Divisive clustering is "top down"

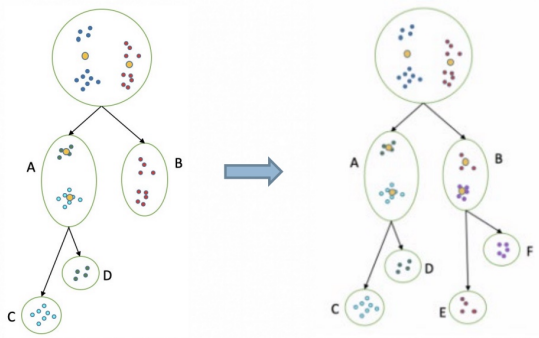
- faster than regular k-means and likely to produce different results
- iteratively partitions existing clusters into two new clusters until K clusters are obtained from the algorithm:

Bisecting K-Means Clustering

- Step 3: Measure intra-class distance for each cluster
  - *Xi* is each instance within a cluster
  - *X-bar* is the centroid of the cluster

$$\sum_{i=0}^{n} (X_i - \overline{X})^2$$

- Step 4: Select the cluster with the largest intra-class distance (i.e., the least dense one) and split it into two via Steps 2-3
- Stop until having K clusters.



19

Bisecting K-Means in PySpark

```
from pyspark.ml.clustering import BisectingKMeans

bkmeans = BisectingKMeans().setK(3).setFeaturesCol("scaledFeatures") .setMinDivisibleClusterSize(1)

bkm_model = bkmeans.fit(data)

predictions = bkm_model.transform(data)
```

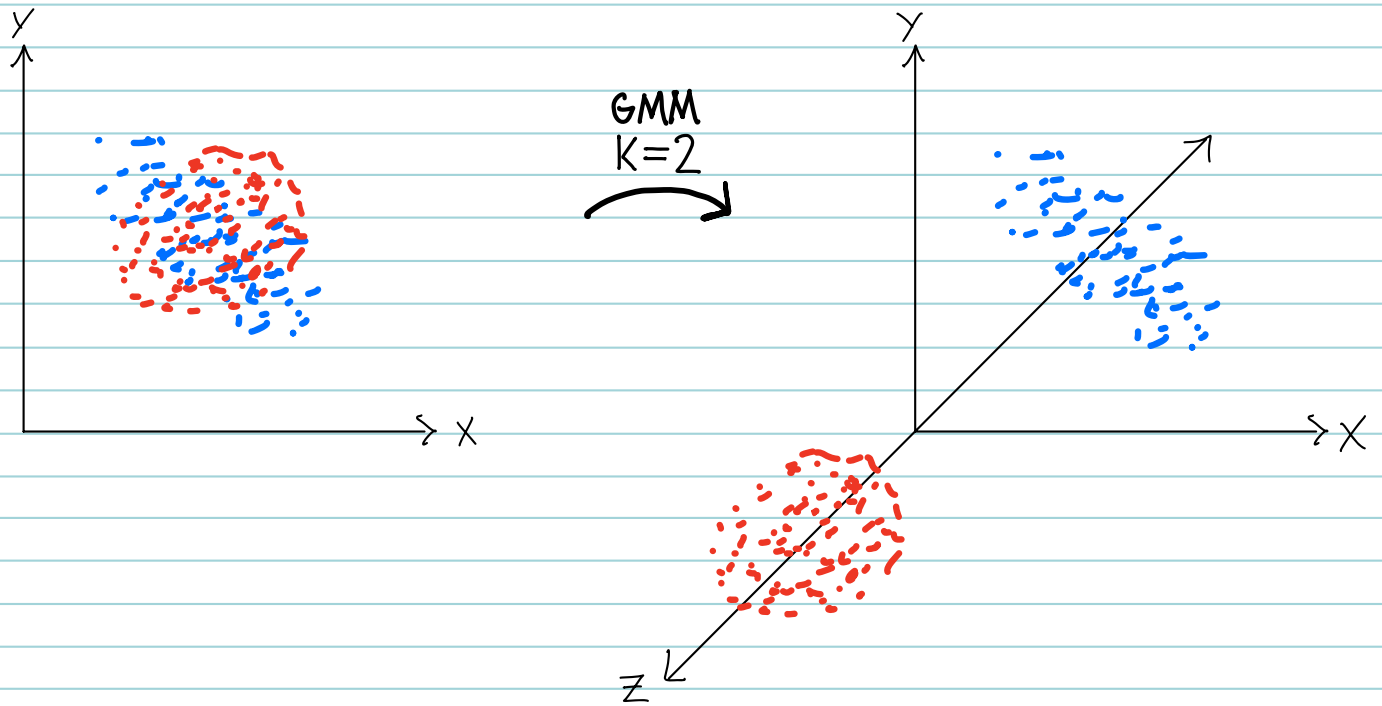
Clusters with this size will not be divided further

20

source: Prof. Xu Feiran BAIS6110_Module5_Hive_Advanced_Part3

III Gaussian Mixture Models (GMM)

- used to cluster high dimentional data which might overlap in two dimensions
- conceptually, GMM is the unsupervised version of an SVM classifier (SVM is a supervised machine learning algorithm capable of handling an arbitrarily large number of data dimensions).
- used to cluster high dimensional data which might overlap in two dimensions
- A Gaussian Mixture Model (GMM) does not create or add dimensions to the data; rather, it utilizes all existing dimensions of the data to model and separate the clusters. The dimensions are inherent features of your dataset. Each dimension is considered in determining how the data is organized into different clusters (components) within the model.

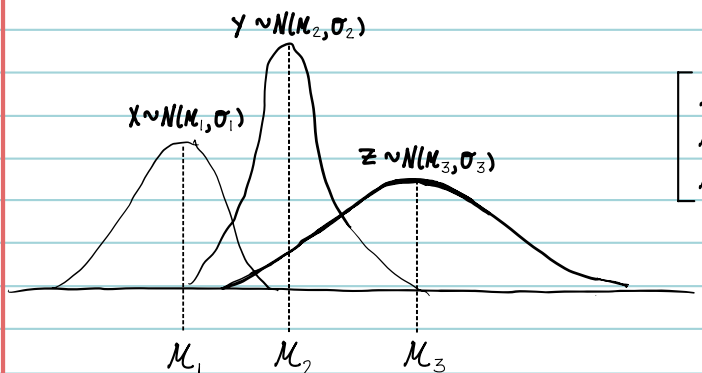


~ GMM FORMS CLUSTERS IN HIGHER DIMENSIONS TO SEPARATE OVERLAPPLING DATA ASSUMING THE MULTI-NORMAL DISTRIBUTION ~

$$(X, Y, Z) \sim N\left(\vec{\mu}, \sum_{3\times 3}\right)$$

variance-covariance matrix

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$



- The **dimensionality** refers to the number of features or variables in your dataset.
- K (the number of components) refers to how many distinct Gaussian distributions you are using to model the data, which reflects how many clusters you believe the data to be segmented into.

Three-Dimensional Data: The observations are in a three-dimensional space (x, y, z). Each Gaussian component in the GMM will have a mean that is a vector of three elements (one for each dimension) and a covariance matrix that is 3×3.