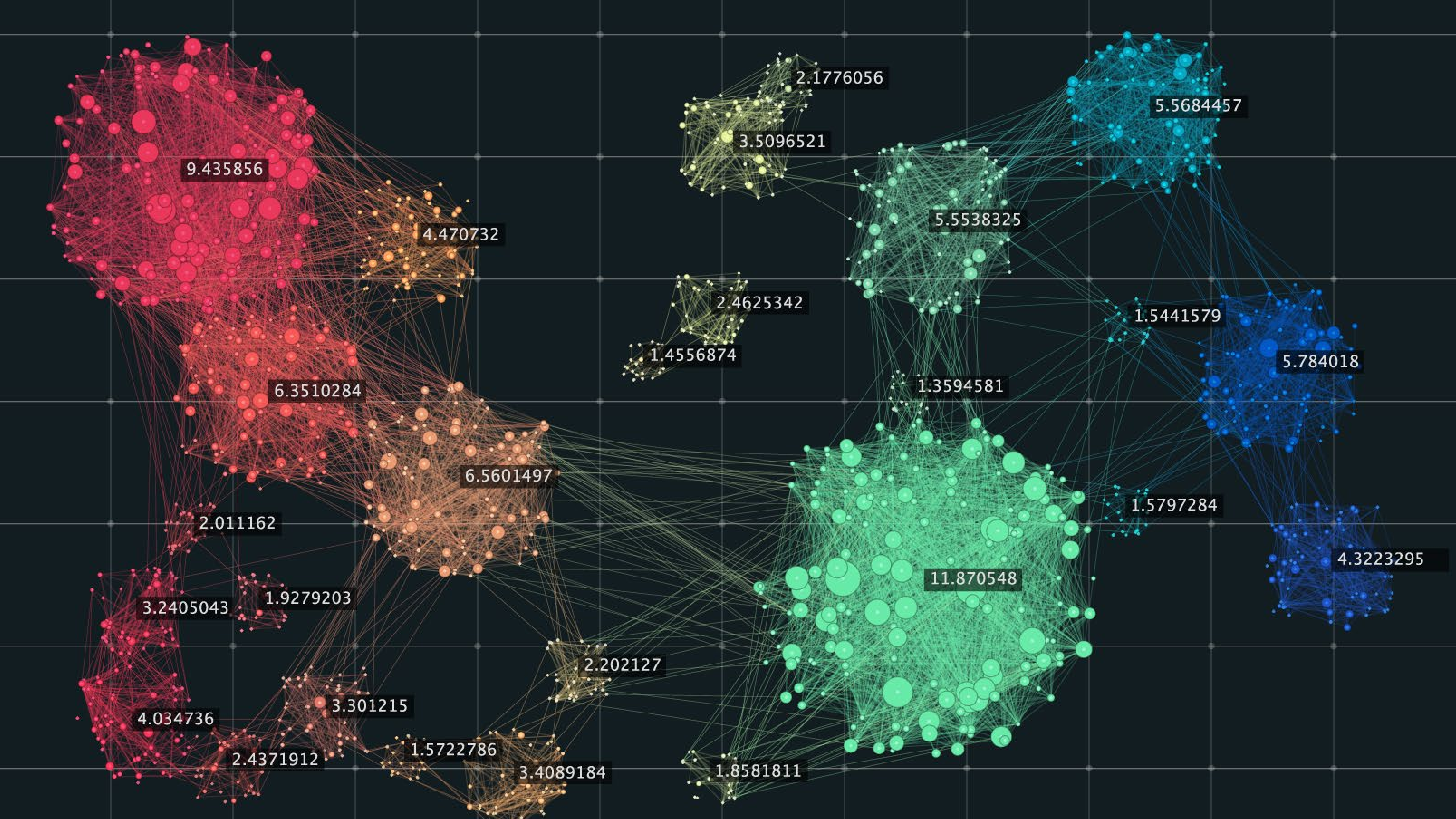
Clustering Algorithms

Xavier Morera

Helping developers understand and work with data

@xmorera www.xaviermorera.com / www.bigdatainc.org

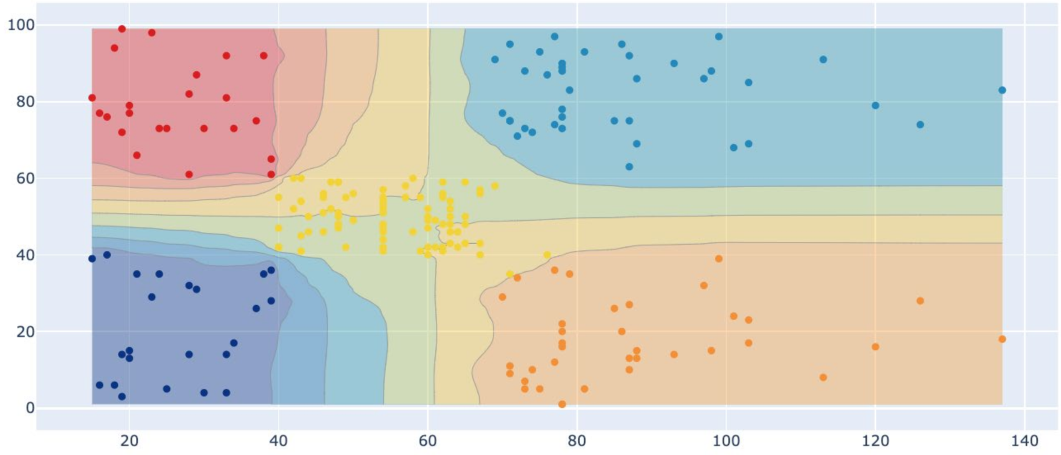

2.1776056
5.5684457
3.5096521
9.435856
4.470732
5.5538325
2.4625342
1.5441579
1.4556874
5.784018
6.3510284
1.3594581
6.5601497
1.5797284
2.011162
4.3223295
11.870548
1.9279203
3.2405043
2.202127
3.301215
4.034736
1.5722786
2.4371912
3.4089184
1.8581811

With clustering, the Machine Learning algorithm predicts which class or group each element or individual belongs to without knowing how many or which classes or groupings exist.

PLATINUM
VIP
1234 5678 9012 3456
VALID FROM 01/16
VALID THRU 01/21

K-Means

K Means



Non-supervised algorithm

Groups data in “k” types

- Possible to define how many groups

Example

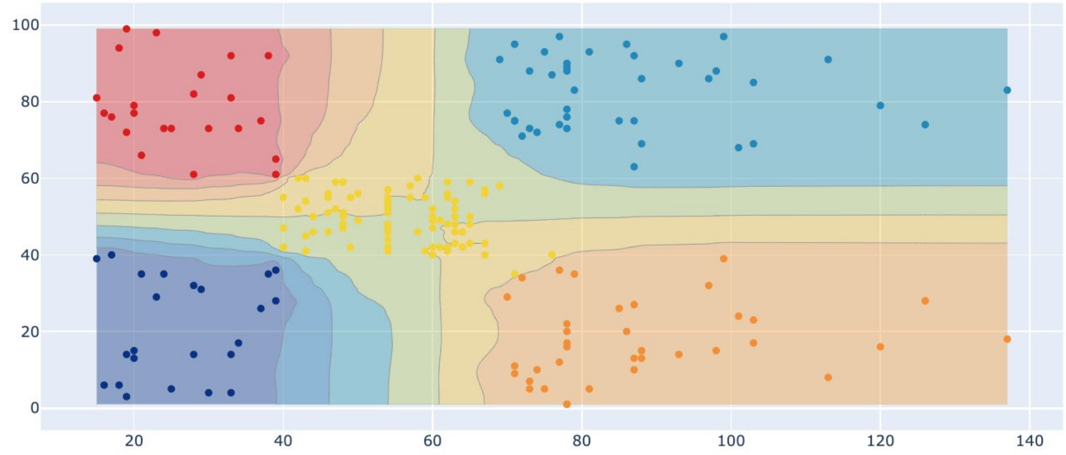
- Data with two points defined (weight, height)

Define centroids and determine distance to each one

- Iterate and recalculate distances until change between iterations is close to zero

K-means is a distance-based algorithm that clusters data based on k-random centroids that will move toward the mean of the labeled neighbors

Keep in Mind for K Means



K is the number of classes

- Selected randomly; can vary on each iteration

Sensible to outliers

Euclidean distance works best if data is standardized

Does not work with categorical data

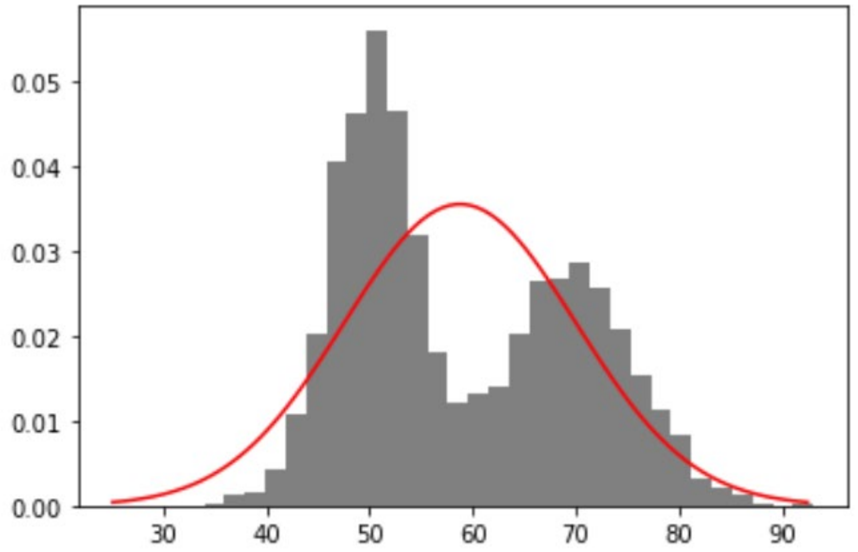
- Use dummy variables or one-hot-encoding

Demo

K-Means

Gaussian Mixtures

The Problem At Hand

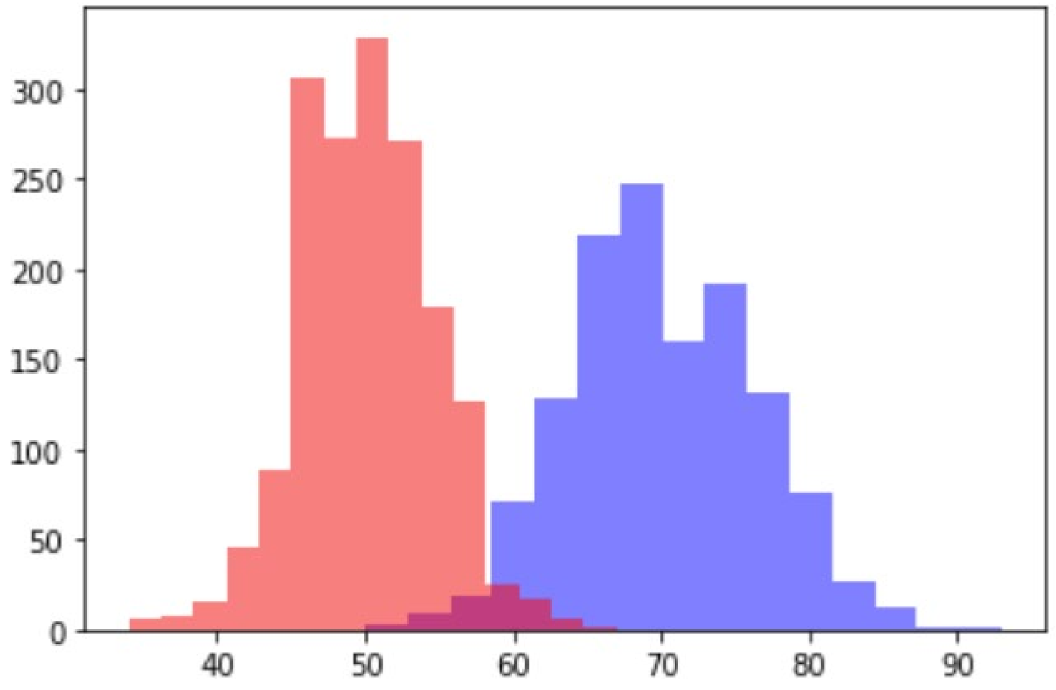


(weight, height)

Male?

Female?

Gaussian Mixtures



Data composed of k-normal (Gaussian) distributions

- In a male/female example then two distributions are most likely present (k=2)
- You get also 2 means and 2 standard deviations

$$p(x) = \pi_1 \mathcal{N}_1 \left(x | \mu_1, \sigma_1\right) + \pi_2 \mathcal{N}_2 \left(x | \mu_2, \sigma_2\right)$$

- Formula for the probability from two distributions mixed together

$$\pi_{i,1} = \frac{\hat{\pi}_{i,1} \mathcal{N}(x_i | \mu_1, \sigma_1)}{\hat{\pi}_{i,1} \mathcal{N}(x_i | \mu_1, \sigma_1) + \hat{\pi}_{i,2} \mathcal{N}(x_i | \mu_2, \sigma_2)}$$

$$\mu_1 = \frac{\sum_{i=1}^{N} x_i \pi_{i,1}}{\sum_{i=1}^{N} \pi_{i,1}}$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{N} \pi_{i,1} (x_i - \mu_1)^2}{\sum_{i=1}^{N} \pi_{i,1}}$$

- Expectation-maximization algorithm gathers the required parameters

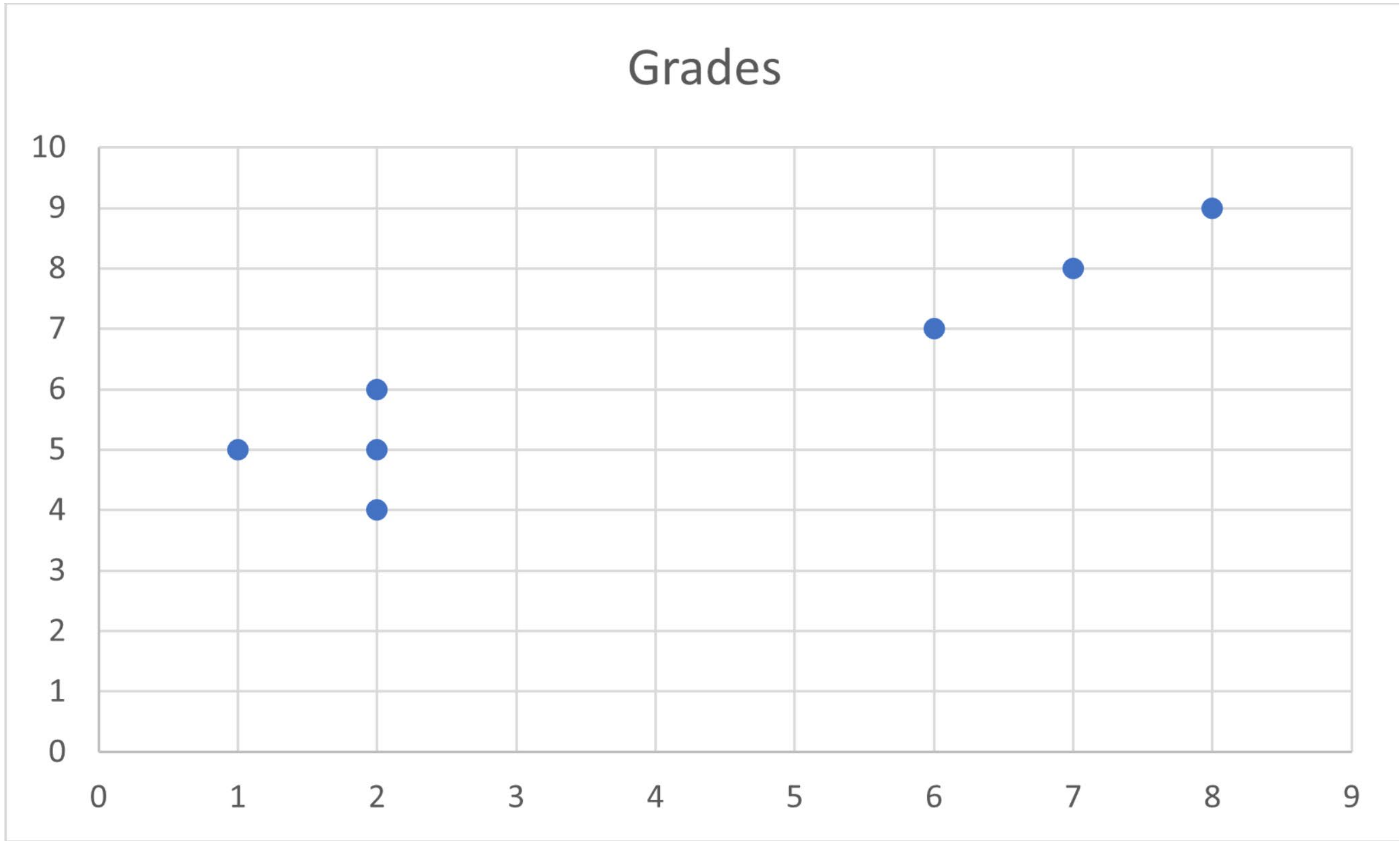
Gaussian Mixtures models (GMMs) are beneficial when data is mixed and can be explained in terms of normal distributions.

Demo

Gaussian Mixtures

Hierarchical Clustering

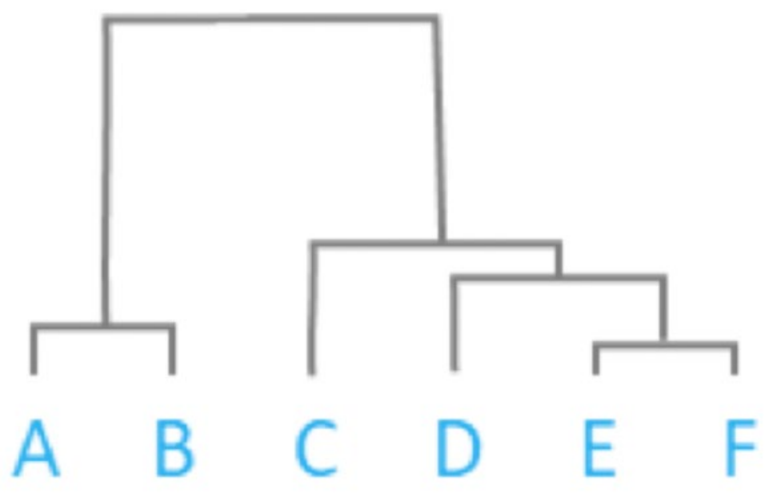
Dataset of Students Grades



(2,6), (5,7), (6,7), (8,9), (1,5), (2,6), and (1,9)

Hierarchical Clustering Algorithm

Dendrogram



Select any random point

Calculate Euclidean distance to all other points

Select nearest neighbor and make a new cluster

Return to first step and repeat until you have k clusters

Hierarchical Clustering

Sensible to outliers

Consider using standardization to scale values

Distance methods must be chosen based on the situation

- Euclidean or Manhattan are two possibilities

Demo

Linear Regression

Affinity Propagation

Unsupervised clustering algorithm
that does not require the number of clusters to be determined

The Four Matrixes

Similarity matrix, where all data rows are compared against each other

Responsibility matrix quantifies how well-suited each element is against the other elements

Availability matrix compares to determine how "appropriate" they are

Criterion matrix shows the exemplar or highest values for each row

Demo

Affinity Propagation

Takeaway

Clustering is an essential application of unsupervised learning

Multiple different clustering algorithms

K-means groups data into a number of types

- Creates centroids and calculates distances between all points
- Classifies data points and iterates until all points belong to their corresponding cluster

Takeaway

Gaussian mixture models

- Useful when data is mixed but can be explained in terms of normal distributions

Hierarchical clustering

- Uses distance methods to create clusters
- Tree-like structure

Affinity propagation

- No need to define number of clusters
- Finds data points that are representative of the clusters