NAYANA DAVIS

INTRO TO CLUSTERING

WHAT IS SUPERVISED MACHINE LEARNING?

SUPERVISED VS UNSUPERVISED MACHINE LEARNING

- Supervised: train an algorithm to predict an unknown variable from known variables
- Unsupervised: Finding patterns in data. Not trying to predict anything

WHAT TYPES OF SUPERVISED MACHINE LEARNING HAVE YOU LEARNED?

SUPERVISED LEARNING TYPES

- Classification: focuses on estimating the relationship between the independent variables and the dependent, categorical variable
- Regression: used to predict continuous values with numerical values as the independent variable and dependent variable

UNSUPERVISED LEARNING TYPE: CLUSTERING

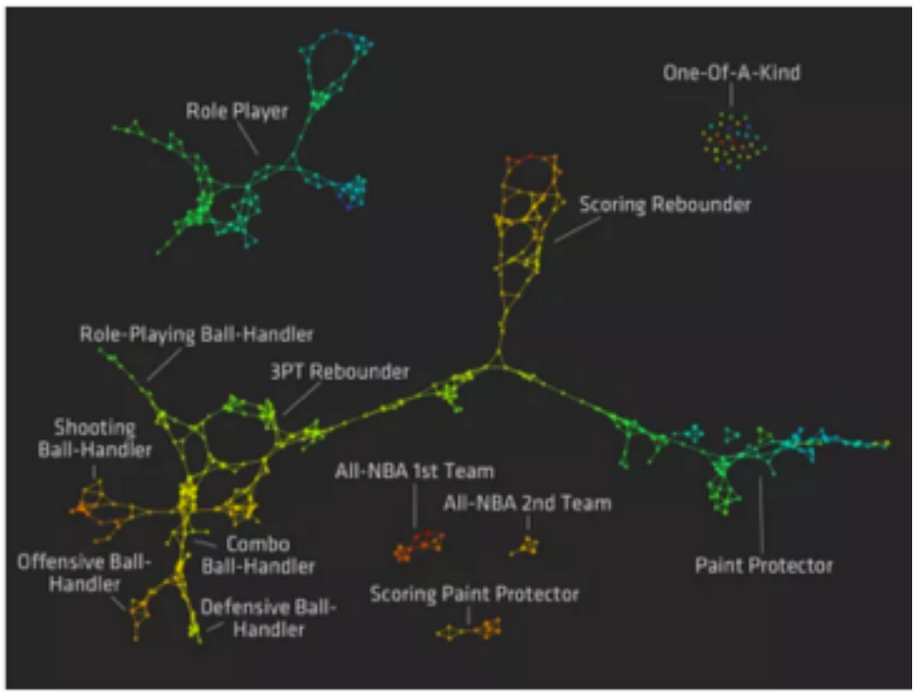
- We use clustering when we're trying to explore a dataset, and understand the connections between the various rows and columns.

- Clustering algorithms group similar rows together. There can be one or more groups in the data, and these groups form the clusters.

- We set out to figure out if the points in our dataset have relationships with each other. That is, discover the classes themselves

CLUSTERING EXAMPLES

- Putting US senators into groups based on how they voted. Did they vote along party lines?
- Clustering NBA players based on stats: http://offthedribble.blogs.nytimes.com/2012/03/06/trading-small-forward-for-scoring-rebounder/?_r=0#

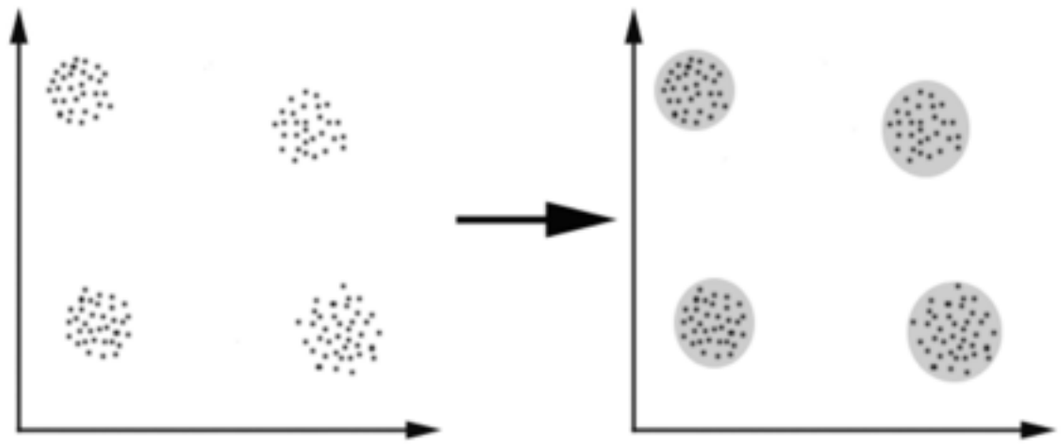


K-MEANS CLUSTERING

- There are numerous algorithms for clustering. K-means is one of the most common and is centroid based
- Centroid based clustering works well when the clusters resemble circles with centers (or centroids). The centroid represent the arithmetic mean of all of the data points in that cluster
- K in K-Means refers to the number of clusters we want to segment our data into. We have to specify what k is.
- These centroids should be placed in a cunning way because of different location causes different result.

K-MEANS CLUSTERING

- A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.



K-MEANS CLUSTERING

- Euclidean distance is the most common technique used in data science for measuring distance between vectors and works extremely well in 2 and 3 dimensions

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

EXAMPLE: SENATORS VOTING RECORDS

```
name,party,state,00001,00004,00005,00006,00007,00008,00009,00010,00020,00026,00032,00038,00039,00044,00047
Alexander,R,TN,0,1,1,1,1,0,0,1,1,1,0,0,0,0,0
Ayotte,R,NH,0,1,1,1,1,0,0,1,0,1,0,1,0,1,0
```

Let's calculate Euclidean Distance between these two using only numeric columns

$$d = \sqrt{(0-0)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2 \cdots + (0-0)^2}$$

K-MEANS ALGORITHM

- Setup K-Means is an iterative algorithm that switches between recalculating the centroid of each cluster and the players that belong to that cluster

EXAMPLE: BASKETBALL

- To start, select 5 players at random and assign their coordinates as the initial centroids of the just created clusters
- For each player, calculate the Euclidean distance between that player's coordinates and each of the centroids' coordinates. Assign the player to the cluster whose centroid is the closest to, or has the lowest Euclidean distance to, the player's values
- For each cluster, compute the new centroid by calculating the arithmetic mean of all of the points (players) in that cluster. We calculate the arithmetic mean by taking the average of all of the X values and the average of all of the Y values of the points in that cluster.

GUIDED PRACTICE

INDEPENDENT PRACTICE