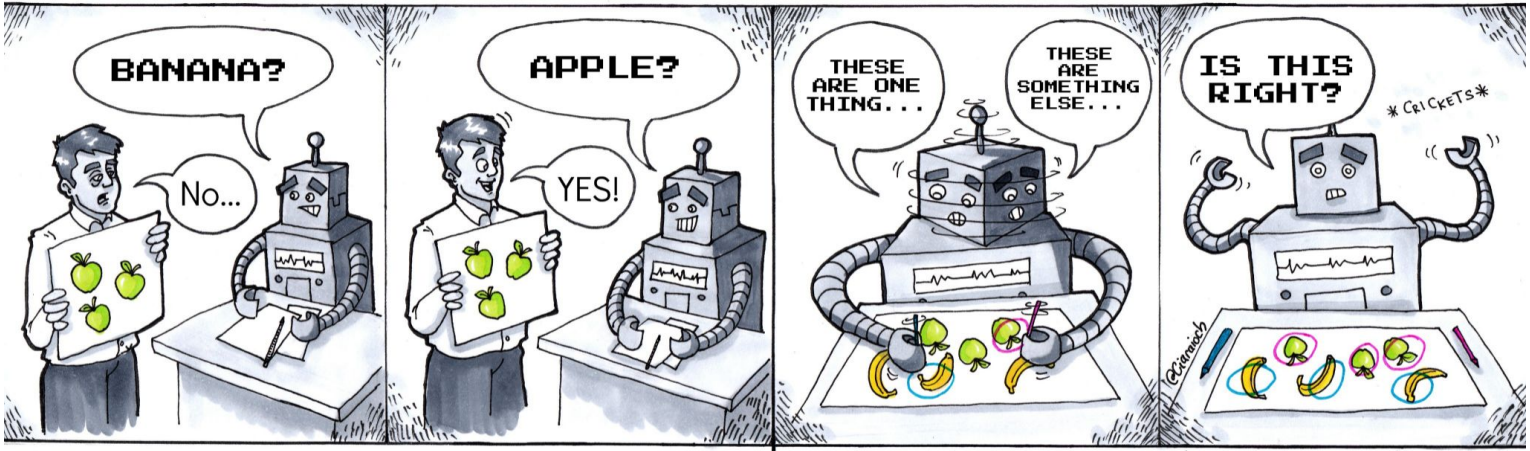


Supervised Learning

Unsupervised Learning

Source

Welcome to Week 9 Lecture 1!

Data Science in Python & Machine Learning

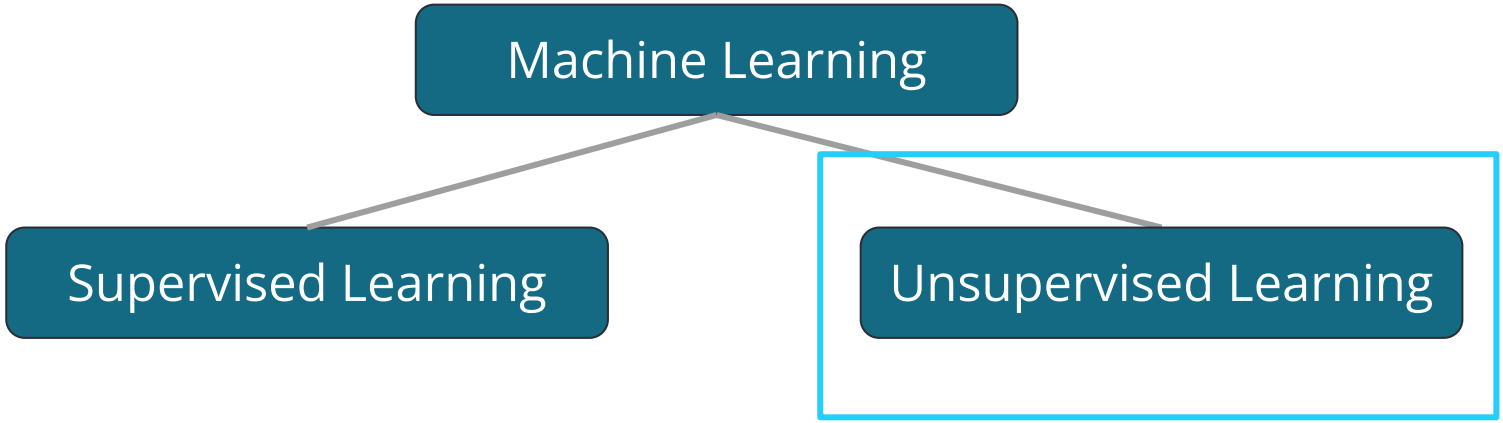
Announcements:

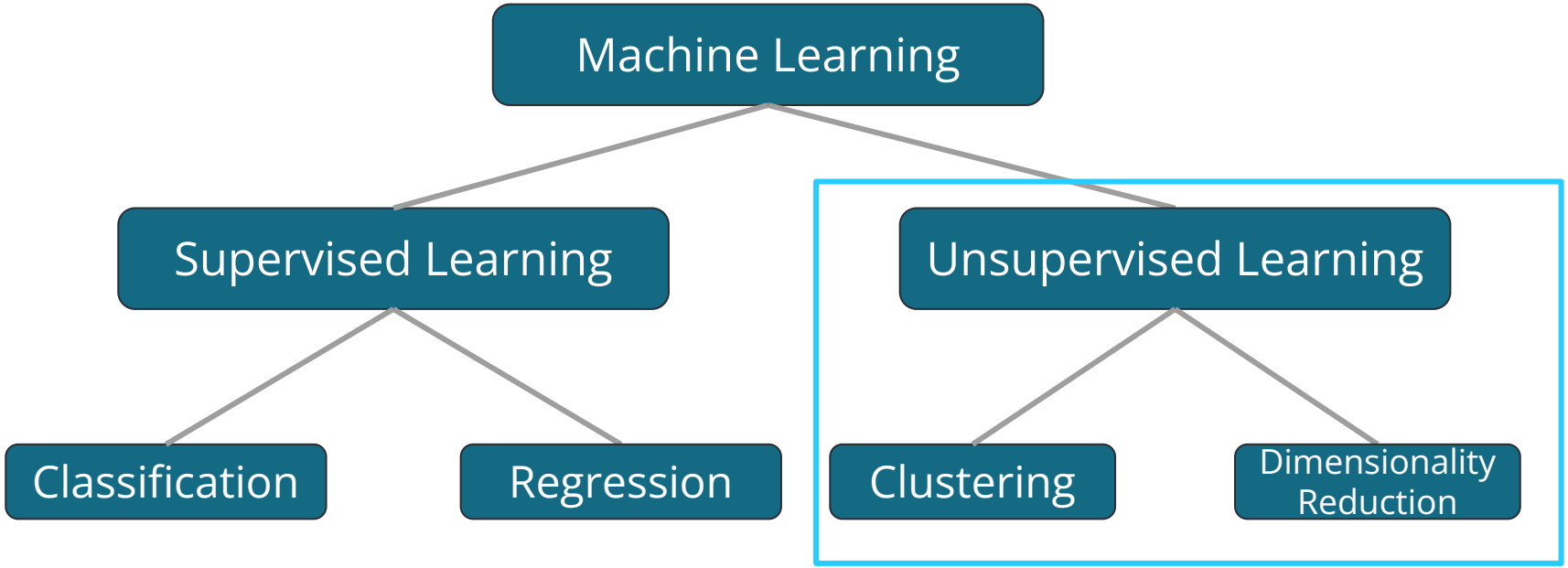
- First week assignments are due until Sunday night.
- All the following week assignments are due Friday morning 9AM (PST)
- Belt Exams are in third week of stack
- Belt Exam eligibility
    - Completion of week 9 assignments including resubmits , and week 10 assignments

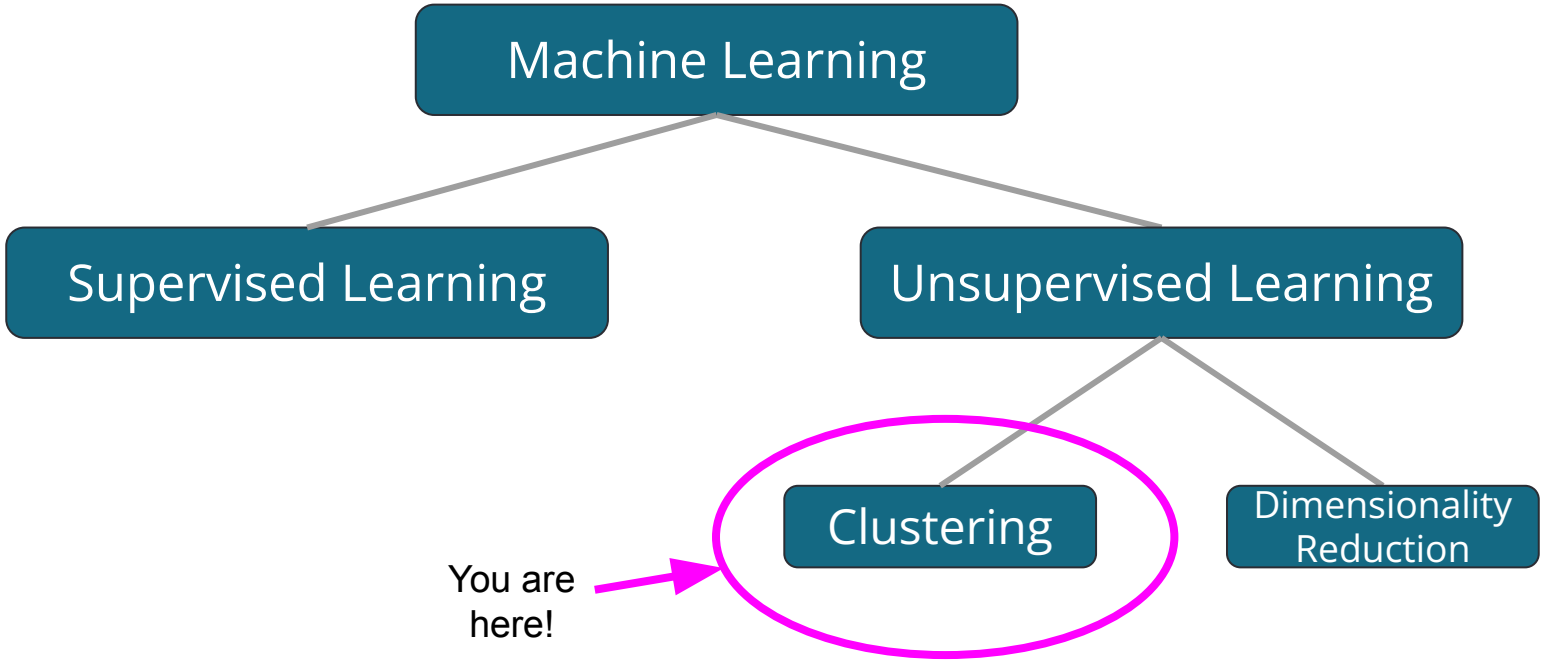
Learning Goals

After this lesson you will be able to:

1. Explain how supervised and unsupervised learning is different.
2. Cluster data using the K Means clustering algorithm.
3. Evaluate the clusters predicted by the K Means clustering algorithm using two different metrics.
4. Analyze clusters to discover insights about data.

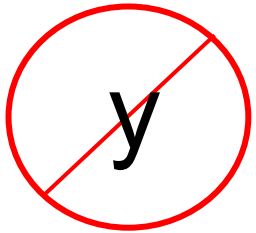

Machine Learning
Supervised Learning
Unsupervised Learning


Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering
Dimensionality Reduction

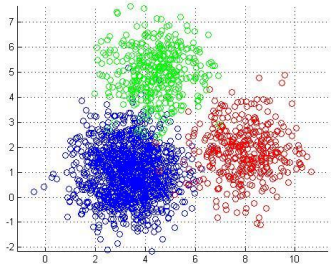

Machine Learning
Supervised Learning
Unsupervised Learning
Clustering
Dimensionality Reduction
You are here!

Supervised or Unsupervised?

- Unsupervised vs. supervised learning methods is based on the *type of problem* you are trying to solve.
- One is not better than the other, they just address *different* questions.
- With unsupervised learning, you do not have a target variable you are trying to predict.
- Unsupervised learning finds patterns in data.



Unsupervised Learning- Clustering



Clustering is a way of grouping your data

Consider this example:

- You have data set that has information about your customers.
- Clustering would allow you to group your customers.
- This helps the sales team better address the needs of each of these groups.
- As a new customer is added, they can be assigned the appropriate group (and immediately receive the marketing email designed for their group).

Understanding Your Data with Clusters

- Today, we are going to take time to explore a data set with the assistance of Kmeans clustering
- Clustering and data exploration go together and can be an iterative process.
- One approach is to use subject matter expertise to explore combinations of features that you think will provide relevant clusters.
- Clustering can be done on many features.
- However, as we learn and visualize the process, it is often easiest to explore clusters in just two dimensions (two features).

Preprocessing for a clustering model

- There is no target variable, so no defining of X and y
- We do NOT do a train test split
- We DO need to scale our data using standardscaler() before we fit our model

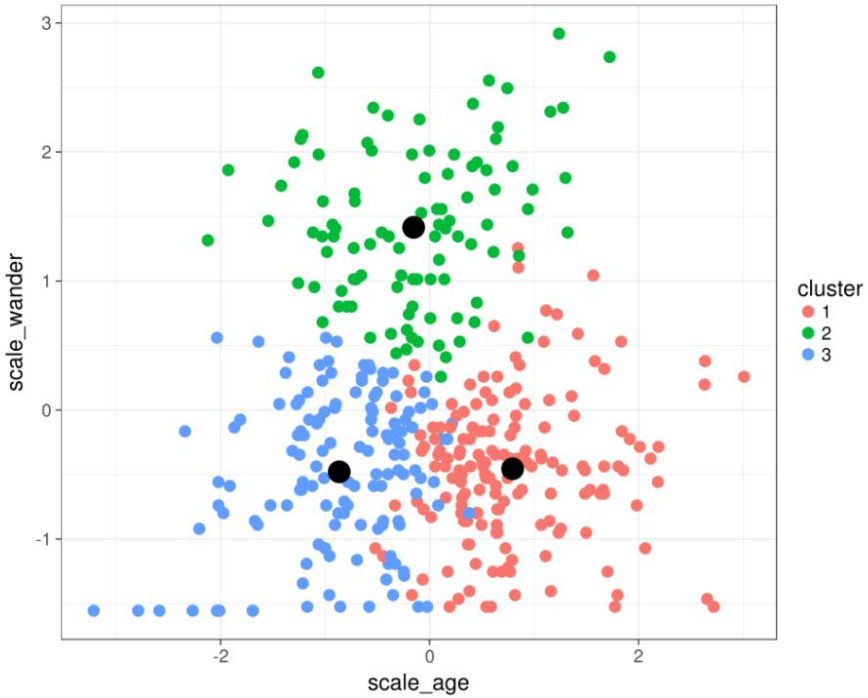
The “job” of the model is to find patterns and group the data into distinct “clusters”

Two aspects of clustering are:

- Points in same cluster should be similar
- Points in different groups should be different

For our mall data, we are looking to divide our customers into target groups that can be the focus of specialized marketing efforts

KMeans Clustering



- k is the number of clusters
- You will have to determine how many clusters as part of your tuning
- The center of each cluster is called the centroid
- Check the learn platform for an explanation of how these centroids are determined.

This visual shows a sample scatterplot based on 2 features and the clusters are color coded. This graph has a k of 3.

Source

Challenge: Cluster Mall Customers

1. Cluster mall customers into similar groups
2. Explore those groups to understand what they are like

	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

Colab Notebook