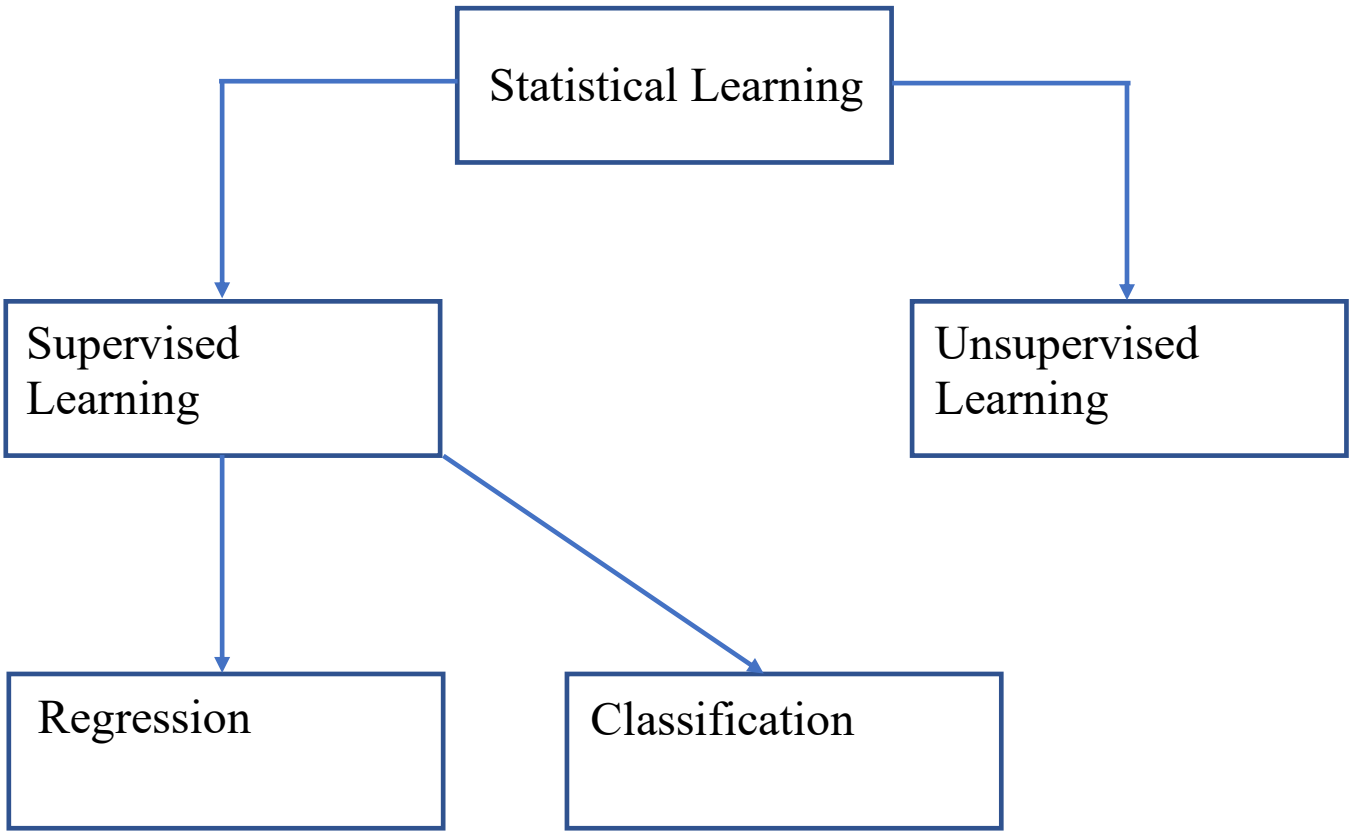
STAT 380 – Clustering Part 1 (Lecture 18)

RECALL: Consider the following:



Regression Techniques:

- Multiple Linear Regression
- k-nearest Neighbor
- Decision/Regression Tree
- Regression Random Forest

Classification Techniques:

- Logistic Regression
- k-nearest Neighbor
- Decision/Classification Tree
- Classification Random Forest

Cross Validation is used for assessing
quality of prediction in supervised learning

Definition: Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set. The major idea is that we want to partition observations into distinct groups so that the observations within the same group are quite "similar" to each other, while observations in different groups are quite "different" from each other.

NOTE: Clustering differs from classification in that you use the known labels (subgroup/cluster membership) to train the model in a classification problem.

EXAMPLE 1: (Refers to Powerpoint slides.) How many slides do you see?

IDEA: As clusters become less separated or when we cannot visualize the data, clustering becomes much more challenging. So, we turn to mathematical procedures.

NOTE: Given two points in space, we must decide whether the points are similar or different. This is often done by calculating the distance between the points. Suppose we have the following points in p-dimensional space $\boldsymbol{x_1} = \left(x_{11}, x_{12}, \ldots x_{1p}\right)^T$ and $\boldsymbol{x_2} = \left(x_{21}, x_{22}, \ldots x_{2p}\right)^T$ The following are popular distance metrics:

- Euclidean Distance (Ruler Distance, Generalization of Pythagorean Theorem)

$$\boldsymbol{d}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{\sum_{j=1}^{p}\left(x_{1j} - x_{2j}\right)^2}$$

- Manhattan Distance (City Block Distance, Taxi Distance)

$$\boldsymbol{d}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sum_{j=1}^{p}\left|x_{1j} - x_{2j}\right|$$

- Minkowski Distance

$$\boldsymbol{d}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \left(\sum_{j=1}^{p}\left(x_{1j} - x_{2j}\right)^m\right)^{\frac{1}{m}}$$

- Hamming Distance - between two equal-length strings of symbols is the number of positions at which the corresponding symbols are different

- Cosine "Distance"

$$\boldsymbol{d}(\boldsymbol{x_1}, \boldsymbol{x_2}) = 1 - \frac{\boldsymbol{x_1} \cdot \boldsymbol{x_2}}{|| \boldsymbol{x_1} ||| \boldsymbol{x_2} ||}$$

<u>K-means Clustering</u>

NOTE: One of the most simplistic, yet most commonly employed clustering methods is called K-means clustering. The procedure for K-means is as follows:

1. Specify the number of clusters, K, that you wish to use.
2. Randomly assign a number, from 1 to K, to each of the observations. These serve as the initial cluster assignments for the observations.
3. Iterate the following steps until the cluster assignments stop changing:
    a. For each of the K clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the p feature means for the observations in the k^{th} cluster.
    b. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

EXAMPLE 2: Suppose you are given the following data:

Name	X_1	X_2
A	7	9
B	3	3
C	4	1
D	3	8

Suppose we decide to partition the points into 2 clusters. After randomly assigning each observation to a cluster, Cluster 1 contains points A and C, while Cluster 2 contains B and D. Find the K-means clustering.

Step 1: Pick k. $k = 2$

Step 2: Randomly Assign points to clusters: Cluster 1: $\{A, C\}$
Cluster 2: $\{B, D\}$

Step 3: Iteration 1

a. Update centroids (average X_1 for points in cluster, average X_2 in cluster)

$\underline{M}_1$ = average of A and C $= \begin{bmatrix} \frac{7+4}{2} \\ \frac{9+1}{2} \end{bmatrix} = \begin{bmatrix} 5.5 \\ 5 \end{bmatrix}$

$\underline{M}_2 = \begin{bmatrix} \frac{3+3}{2} \\ \frac{3+8}{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 5.5 \end{bmatrix}$

b. Update cluster memberships. Find distance between each point and each centroid. Assign point to closest centroid

One distance calculation as example

$$d(\underline{A}, \underline{M}_1) = \sqrt{(x_{A1} - M_{11})^2 + (x_{A2} - M_{12})^2} =$$

$$= \sqrt{(7 - 5.5)^2 + (9 - 5)^2} = \sqrt{18.25}$$

Repeating for all pairs, we get the distances

	A	B	C	D
$\underline{M}_1$	$\sqrt{18.25}$	$\sqrt{10.25}$	$\sqrt{18.25}$	$\sqrt{15.25}$
$\underline{M}_2$	$\sqrt{28.25}$	$\sqrt{6.25}$	$\sqrt{21.25}$	$\sqrt{6.25}$

EXAMPLE 3: Suppose you are given the following data:

Name	X_1	X_2
A	7	9
B	3	3
C	4	1
D	3	8

Suppose we decide to partition the points into 2 clusters. After randomly assigning each observation to a cluster, Cluster 1 contains points A and B, while Cluster 2 contains C and D. Find the K-means clustering.

Step 1: Pick k. $k = 2$

Step 2: Randomly Assign points to clusters: Cluster 1: {A, D}
Cluster 2: {C, D}

Step 3: Iteration 1

a. Update centroids (average X_1 for points in cluster, average X_2...)

$\underline{M}_1$ = average of A and B = $\begin{bmatrix} 5 \\ 6 \end{bmatrix}$

$\underline{M}_2 = \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix}$

b. Update cluster memberships. find distance between each point and each centroid. Assign point to closest centroid

	A	B	C	D
$\underline{M}_1$	$\sqrt{13}$	$\sqrt{13}$	$\sqrt{26}$	$\sqrt{8}$
$\underline{M}_2$	$\sqrt{32.5}$	$\sqrt{2.5}$	$\sqrt{12.5}$	$\sqrt{12.5}$

EXAMPLE 3: (More Space)

Step 3: Iteration 2 (New Cluster 1: {A, D}, Cluster 2 {B, C})

a. Update centroids (average x_1 for points in cluster, average x_2...)

$\underline{\mu}_1$ = average of A and D = $\begin{bmatrix} 5 \\ 8.5 \end{bmatrix}$

$\underline{\mu}_2 = \begin{bmatrix} 3.5 \\ 2 \end{bmatrix}$

b. Update cluster memberships. find distance between each point and each centroid. Assign point to closest centroid

centroid	A	B	C	D
$\underline{\mu}_1$	$\sqrt{4.25}$	$\sqrt{34.25}$	$\sqrt{57.25}$	$\sqrt{4.25}$
$\underline{\mu}_2$	$\sqrt{61.25}$	$\sqrt{1.25}$	$\sqrt{1.25}$	$\sqrt{36.25}$

EXAMPLE 4: Obviously the results are sensitive to the initialization. Which clustering should we use?

IDEA: We want to minimize:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Where $W(C_k)$ is the amount by which observations within component k differ from each other. IN OTHER WORDS, this formula states that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

NOTE: The within-cluster variation is found by squaring the Euclidean distance between each point and its centroid and summing over these values.

EXAMPLE 5: Using the objective to minimize the within-cluster variation, which clustering is preferred?

NOTE: Total within cluster variation (i.e., final value of sum above) is denoted as WSS.

NOTE: In practice, several initializations are used. The final clustering is based on the clustering that minimizes the within-cluster variation.