BMEG 3105 Lecture 9

Classification

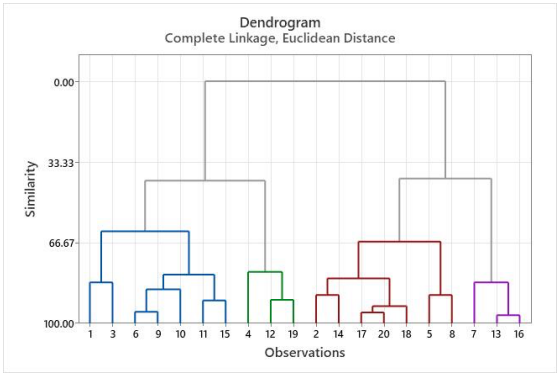
October 3, 2025 (Friday)

Clustering:

- Finding groups of objects that is like one another and different from others.
- Use similarity or dissimilarity to measure the differences.
- Intra-cluster distances are small while inter-clustering distances are large.

Hierarchical clustering:

- Produces a set of nested clusters (hierarchical tree)
- Can be visualized as a dendrogram
- They may correspond to meaningful taxonomies
  - E.g. Gene cluster, phylogeny reconstruction, animal kingdom…



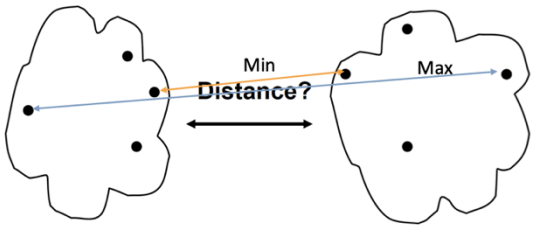
dendrogram

Steps of hierarchical clustering:

1. Compute the similarity or distance matrix
2. Let each data point be a cluster
3. Merge the two closest clusters
4. Update the similarity or distance matrix
5. Repeat the steps 3 and 4 until only a single cluster remains.

Method of updating distance matrix after merging:

- Min
- Max
- Group average
- Distance between centroids



Running Example:

Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

We use correlation (linear correlation) as distance:

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

1. Let each gene be a cluster. Compute the Similarity or Distance Matrix with linear correlation.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

2. Merge the At2g34930 and At1g30720. (The most similarity clusters)
3. Update the Similarity or Distance Matrix with minimum distance (largest correlation).

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

4. Merge the At2g34930, At1g30720 and At4g35770.
5. Update the Similarity or Distance Matrix with minimum distance (largest correlation).

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.5774	-0.5774	0.5774	

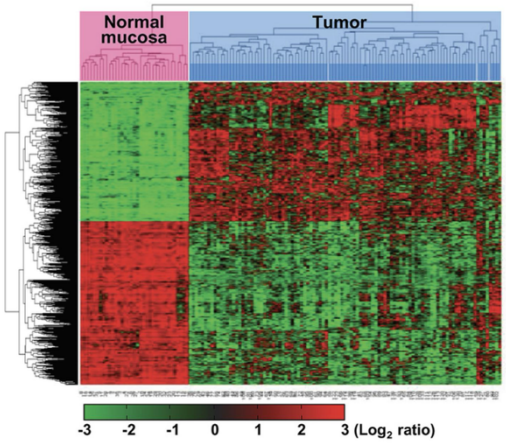
6. Merge the At2g34930, At1g30720, At4g35770 and At2g05540.
7. Update the Similarity or Distance Matrix with minimum distance (largest correlation)

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		

Node2
Node1
Node3

Classification:

Why?



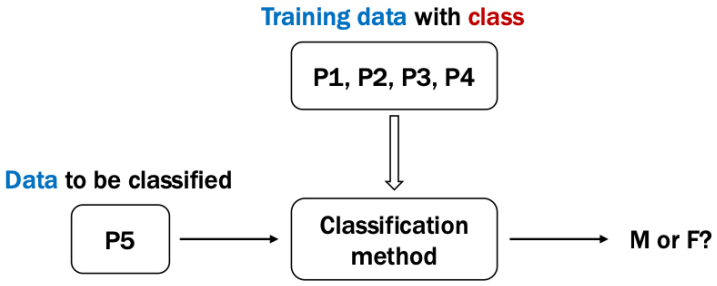
- Characteristics of each class
- Classify items and people
- In Biology: Predict a new gene expression profile is normal or tumor

What is Classification?

- Given a training set
- Each record contains a set of attributes, one of the attributes is the class
- Find a method to assign the class of previously unseen records based on their other attributes and the training set as accurately as possible

How? What is required?

- Training data with class
- Classification method
- Data to be classified



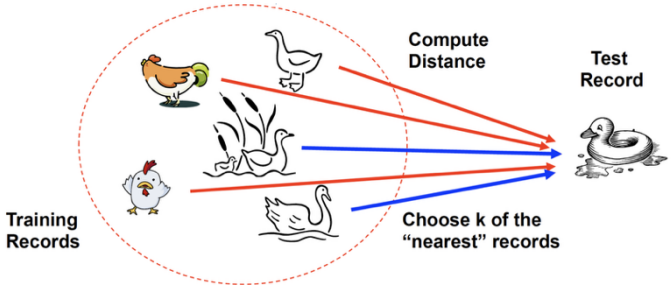
K-nearest neighbors:

- Definition: a simple algorithm that stores all available instances -> classifies new instances based on distance metric to available ones.

How?

Training process:

- Store the available training instances



Predicting process:

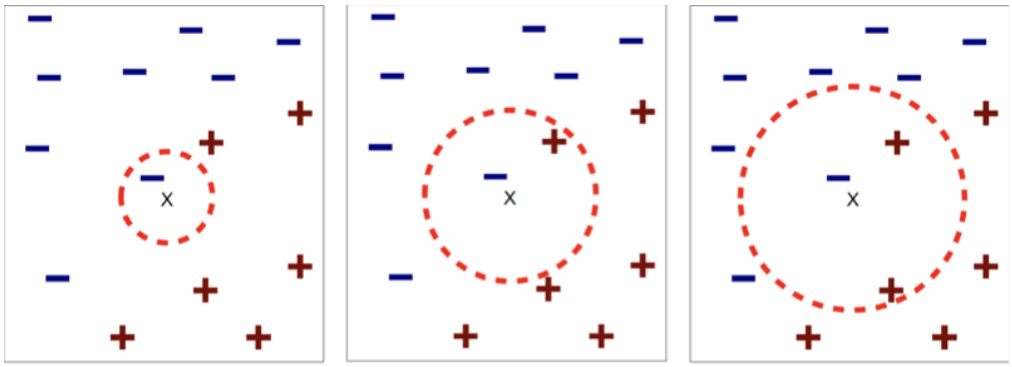
- Find the K training instances that are closest to the query instance
- Return the most frequency class label among these K instances

**Data should be normalized*

What factor will affect the result?

1. Choosing of K (No. of neighbors to look at)

E.g. (K=1) (K=2) (K=3)



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

2. Distance metric to use
   - E.g. Euclidean distance, Manhattan Distance, etc.
3. A weighing function (optional, not major factor)

E.g. We have three data: 0.6(-), 0.3(+) and 0.1(+). As a result, we can predict the result is (-).

How to choose K?

- Between 5-10 gives good result for most low-dimensional data sets
- Chosen by using cross-validation

Example of KNN:

Distance metric: Euclidean distance; K=2

Person	Height(m)	Weight(kg)	Gender
P1	1.79	75	M
P2	1.64	54	F
P3	1.70	63	M
P4	1.88	78	M
P5	1.75	70	??

1. Normalization
   - Normalize the data set with min-max normalization

Person	Height(m)	Weight(kg)	Gender
P1	0.625	75	M
P2	0	0	F
P3	0.25	0.375	M
P4	1	1	M
P5	0.4583	0.6667	??

2. Compute distances
   - Compare their distance with Person 5 respectively

Person	P5	Gender
P1	0.267	M
P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

3. Identify the K most similar data

Person	P5	Gender
P1	0.267	M
P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

4. Take their class out and find the mode class

   Gender: M

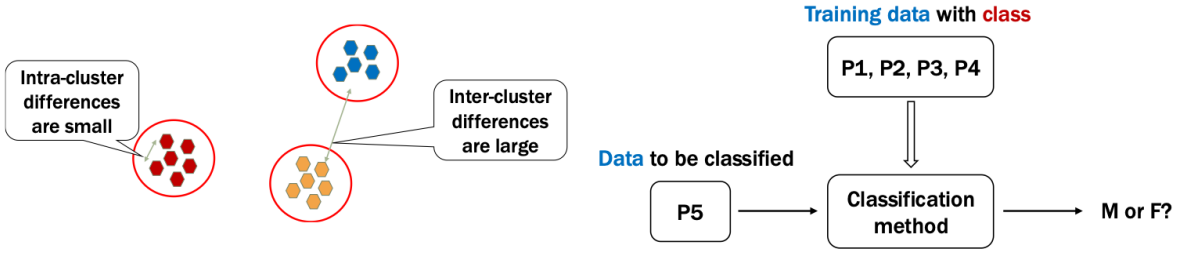
Clustering and Classification

Definition:

Clustering	**Classification**
Data to be clustered	Training data with class
Similarity measurement	Classification method
Clustering algorithm	Data to be classified

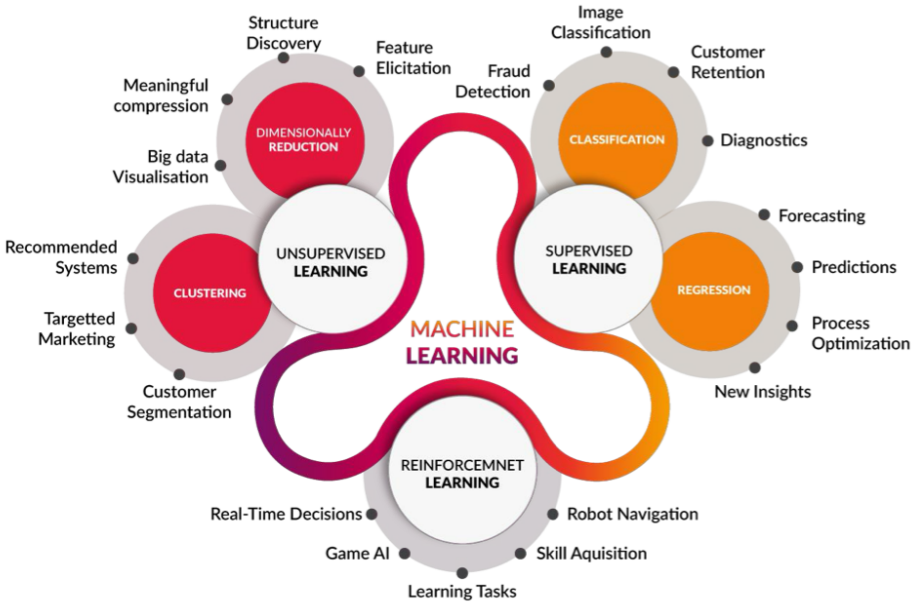
Differences:

	Clustering	**Classification**
Goal	Find similarity (clusters) in the data	Assign class to the new data
Data	Without class	Training data with class and testing data without class
Classes	Unknown number	Known number
Output	The cluster index for each point	The class assignment of the testing data
Algorithm	One phase	Two phases (training and application)



Machine learning

1. **Unsupervised learning**
   - Analyze and cluster unlabeled data
   - E.g. clustering and dimension reduction
2. **Supervised learning**
   - Classify and predict outcomes, trained on labelled data
   - E.g. classification and regression
3. *Reinforcement learning (will not talk about in the course)*



Disclaimer: All figures are adapted from Prof. Li BMEG3105 Lecture Notes and from the internet