Outline of lecture

- Classification
- K-nearest neighbor (KNN) classification
- Clustering VS classification

0 Recap from last lecture

➢ Clustering

✓ Purpose: Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

✓ Correlation: measures the linear relationship between objects

Formula:

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

✓ Distance: measures the difference between objects

Formula:

$$dist(\boldsymbol{p},\boldsymbol{q}) = \left(\sum_{k=1}^{m} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

Manhattan distance (r = 1): Number of bits different between two binary vectors

Euclidean distance (r = 2):

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2}$$

Supremum distance ($r \rightarrow \infty$): The maximum difference between any component of vectors

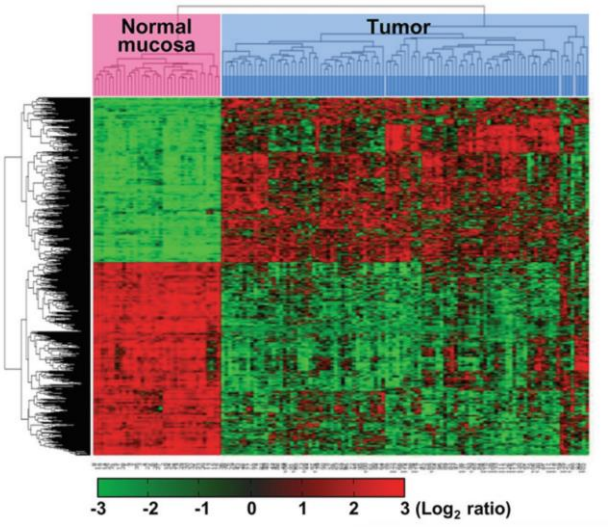
- ✓ Hierarchical clustering: Data matrix --- Distance matrix --- cluster updates
  - ✧ Procedure: Calculate the distance matrix between clusters, then merge the closest clusters to one cluster. Repeat the steps and finally only one cluster left
  - ✧ Tip: When updating the distance matrix, we normally do not need the original data matrix.

1 Classification

- ➢ Why classification?

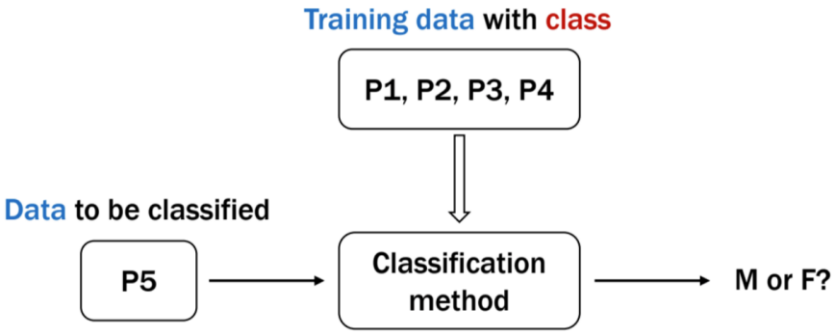
  Extract the features of each class, then classify the items for better organization and possible further prediction of new items.

- ➢ Classification in biology



- Label: Normal mucosa and tumor
- Features: Gene expression level
- By classification, we can investigate the gene expression difference between the classes and thus classify new samples precisely and accurately

- How to do classification?
  - Procedure



1. Training data with class
2. Classification method is defined and developed by the training data set
3. Testing set is classified (predicted) by the classification method

2 K-nearest neighbor (KNN) classification

- KNN is one of the simplest machine learning algorithms based on supervised learning technique
- Training: Store the available training instances
  Predicting:
  - Find the K training instances that are closest to the query instance
  - Return the most frequent class label among those K instances

  Data normalization is required
- Required parameters:
  - Distance metric: How to find the distance between neighbor and data point
  - Value of K:
    - ✧ Depend on many factors (distance matrix, scale of sets, algorithm...)
    - ✧ Normally between 5 and 10 (Less than 5 is noisy and lead to effect of outliers in the model; More than 10 is time and space consuming)
    - ✧ It can be chosen by using cross-validation
  - Weighing function: different distance of a neighbor from the data point should have different weight (the further, the less) to assist the voting procedure
- The standard procedure of KNN:
  1. **Normalization**

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	??

Person	Height	Weight	Gender
P1	0.625	0.875	M
P2	0	0	F
P3	0.25	0.375	M
P4	1	1	M
P5	0.4583	0.6667	??

2. Compute distances

Person	Height	Weight	Gender
P1	0.625	0.875	M
P2	0	0	F
P3	0.25	0.375	M
P4	1	1	M
P5	0.4583	0.6667	??

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

P5 has gender of Male

- Advantage of KNN:
  1. It is simple to implement.
  2. It is robust to the noisy training data
  3. It can be more effective if the training data is large.

  Disadvantage of KNN:
  1. Need to store all the data
  2. Need to calculate the distance matrix

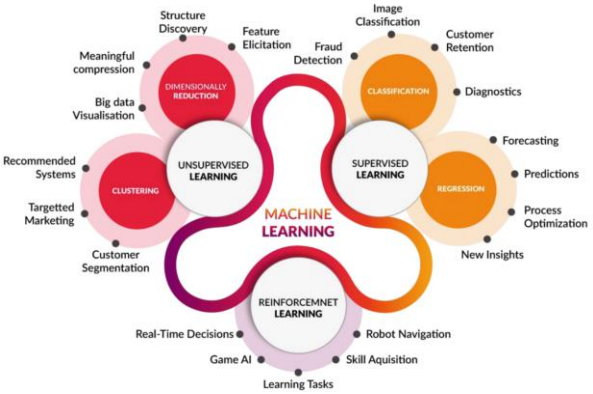
Therefore, KNN has a very fast training procedure (just store the data) but a slow predicting procedure (Need to calculate the distance to all the training data points), which is not useful in applications required rapid prediction, but it is still a very basic and powerful classification algorithm

3 Clustering VS classification

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

- Unsupervised learning
  - Machine learning algorithms to analyze and cluster unlabeled data
  - Example: clustering and dimension reduction
- Supervised learning
  - Machine learning algorithms to classify and predict outcomes, trained on labelled data
  - Example: classification and regression
- Machine learning:

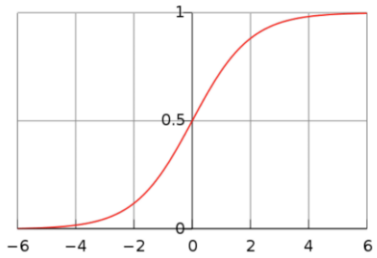
Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.



Logistic function

A formula that we can getting the results with simple arithmetic calculation instead of calculate distance to all the neighbors in KNN
As the different attributes may not be equally important, we add weights (w_h, w_w) and bias (w_0) to adjust the formula

- $\frac{1}{1+e^{-(w_h H + w_w W + w_0)}} \geq 0.5$
- Training: fit the training data
  - To get w_h and w_w, and w_0
- Testing: run the formula



$\frac{1}{1+e^{-t}} \geq 0.5$

Potential project-3

Data preprocessing for the gene expression matrix

- Data collecting and merging
- Exploration
- Visualization
- Data cleaning
- Perform classification

Useful links

KNN in python:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Logistic regression in Python:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Clustering and classification:

https://www.youtube.com/watch?v=B0TI2q7wgIQ