Scribing: Lecture 9: Classification

Created	@October 9, 2025 1:31 AM
Class	BMEG3105
Instructor	Ng Sui Ip 1155213651

Hierarchical clustering:

Hierarchical clustering helps producing a set of nested clusters organised as a hierarchical tree. It can be visualised as a dendrogram which is a tree like diagram that records the sequences of merges. Clusters can correspond to meaningful taxonomies like gene clusters, phylogeny reconstruction, etc.

Steps:

1. Compute the similarity or distance matrix
2. merge the two closest clusters/ value
3. update the similarity or distance matrix
4. repeat the process of merging and update until same value/ single cluster is left

Note: Similarity or distance matrix can be calculated by methods such as linear correlation.

Example:

Assume we use linear correlation to compute the similarity:

A running example

Compute the Similarity or Distance matrix
Let each data point be a cluster
Merge the two closest clusters
Update the similarity or distance matrix
Merge the two closest clusters
Update the similarity or distance matrix
Merge the two closest clusters
Update the similarity or distance matrix
...
Until only a single cluster remains

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

Visualization after normalization

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

1. Each gene be a cluster

We merge At1g30720 and At2g34930 (with the most similar values) and update the minimum value (largest correlation)

A running example

Compute the Similarity or Distance matrix
Let each data point be a cluster
Merge the two closest clusters
Update the similarity or distance matrix
Merge the two closest clusters
Update the similarity or distance matrix
Merge the two closest clusters
Update the similarity or distance matrix
...
Until only a single cluster remains

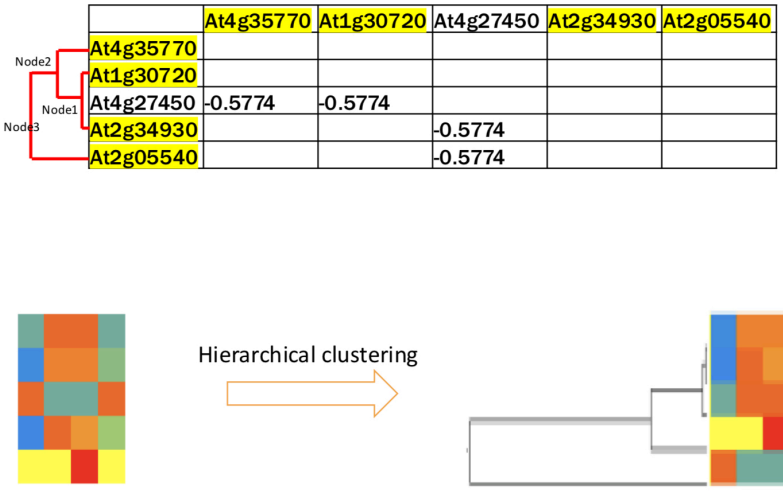
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

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

1. Each gene be a cluster
2. Merge At2g34930 and At1g30720
3. Update with minimum distance (largest correlation)

Repeat the process until one single cluster is left

A running example



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

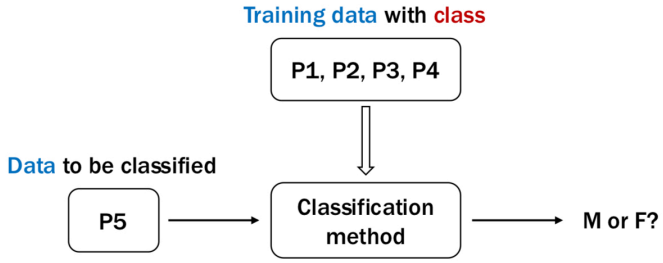
頁， 共53頁
ification Yu Li Lecture 9-14

Classification:

Classifying items is essential in organisation and putting new items into correct place. With classification, you can understand whether the person or item is within the targeting group. In biology, it is common to classify different kinds of items, ranging from new gene, new cell, new species, etc.

Given a collection of records (training set), including a set of attributes and class. Classification aims to find a method to assign the class of previously unseen records based on other attributes and the records collection as accurate as possible.

What are needed to do classification?



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	??

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

lassification Yu Li Lecture 9-22

Classification Method: K-nearest Neighbour Classification:

K-nearest Neighbour (KNN) classification is a simple, instance-based learning algorithm that classifies new samples based on similarity measures. It works by finding the K closest training examples among their attributes and using the most frequent class (mode) among these neighbors to determine the class of the new point.

Imagine you are having a encyclopaedia of animals as training set, you are then given an unknown animal with sufficient attributes including appearance, walking style, sound, etc. Reading through the encyclopaedia, you have found out that the quack sound and walking style is very frequently found in duck. Maybe there are other animals which also quack and walk like duck, but the animal is the most likely a duck.

How KNN works:

1. Store the available training instances
2. Calculate the distance between the query instance and all training samples

3. Sort the distances and identify the K (usually 5-10) training instances that are closest to the query instance
4. Gather the category labels of these K-nearest neighbors
5. Return the most frequent class label among those K instances

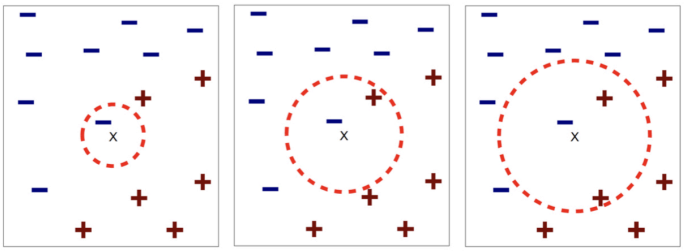
Notes: Data should be normalised!

What should we determine when using KNN?

❖A distance metric

❖How many neighbors to look at
➢K

❖A weighing function (optional)



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

29頁， 共53頁
lassification Yu Li Lecture 9-2

Example:

We first normalise the data

A running example of KNN

❖Suppose we have chosen the Euclidean distance and K=2

- ➢Normalization
- ➢Compute distances
- ➢Identify the K most similar data
- ➢Take their class out and find the mode class

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	??

4頁，共53頁 Yu Li Lecture 9

Then we compute distances (or similarities) using different method, in this example we use Euclidean distance

A running example of KNN

❖Suppose we have chosen the Euclidean distance and K=2

- ➢Normalization
- ➢Compute distances
- ➢Identify the K most similar data
- ➢Take their class out and find the mode class

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	??

第36頁，共53頁 Yu Li Lecture 9-35

Suppose K=2, we try to identify the K most similar data

A running example of KNN

❖Suppose we have chosen the Euclidean distance and K=2

➢Normalization

➢Compute distances

➢Identify the K most similar data

➢Take their class out and find the mode class

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	??

第38頁， 共53頁

We take their class out and find the mode class. In this example, the mode of gender in K most similar data is obviously M. Hence, P5 is probably a male

Clustering VS Classification:

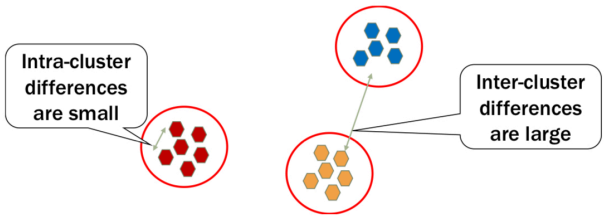
Considered both are unsupervised learning methods, clustering and classification serve different purposes:

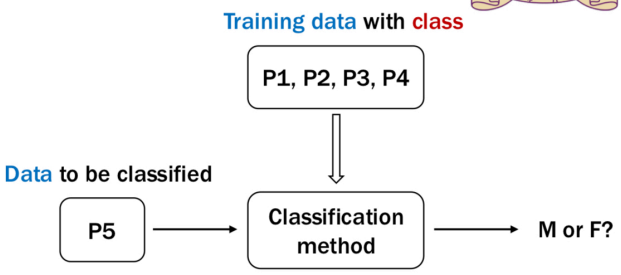
Clustering groups similar items together without predefined categories, discovering natural groupings in data.

While classification assigns new data to predefined categories based on a training set.

In short, clustering finds out "what groups exist in this data?" while classification finds out "which known group does this new data belong to?"

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)

Classification Yu Li Lecture 9-41

Machine learning