Zhou Jieya 1155173741
October 2, 2024

BMEG3105 Lecture 8: Classification

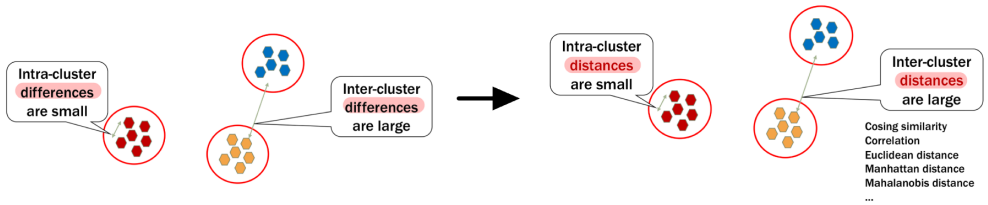
Clustering

Clustering analysis

- 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.
- Intra-cluster differences are small, while inter-cluster differences are large.
- Problem: how we define "differences" and its size?

Clustering

- We use similarity or dissimilarity to measure the differences.



Minkowski distance (A generalization of Euclidean Distance)

- r = 1. City block (Manhattan, taxicab, L1 norm) distance.
  A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors.
- r = 2. Euclidean distance.
- r→∞. "supremum" ($L_{max}norm, L_{\infty}norm$) distance.
  This is the maximum differences between any component of the vectors.

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

Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram:
  A tree like diagram that records the sequences of merges
- They may correspond to meaningful taxonomies:
  Gene clusters, phylogeny reconstruction, animal kingdom…

Steps of hierarchical clustering

1. Compute the Similarity or Distance matrix ← Can use only after we define distance between 2 points.
2. Let each data point be a cluster.
3. Merge the 2 closest clusters
4. Update the similarity or distance matrix
5. Repeat the steps 3 and 4 until only a single cluster remains.

Distance matrix

Methods:

- Min
- Max
- Group average
- Distance between centroids

A 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	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

Highest similarity

2. Merge the 2 most similarity clusters (At2g34930 and At1g30720).
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 2 most similarity clusters (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	

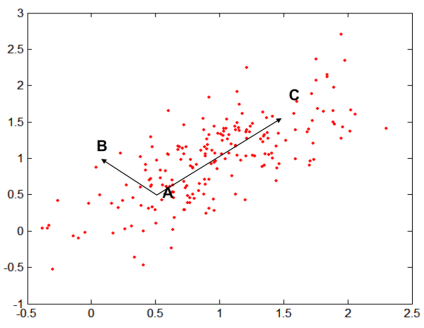
6. Merge the 2 most similarity clusters (At2g34930, At1g30720, At4g35770 and At2g05540).
7. Update the Similarity or Distance Matrix with minimum distance (largest correlation).

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

Node2
Node1
Node3

8. Merge the last two clusters.

Mahalanobis distance



- Calculating distance considering the data distribution

Function

$$mahalanobis(p,q) = (p-q)^T \sum^{-1} (p-q)$$

$\sum$ is the covariance matrix (Physical significance: whether 2 arrivals will change at the same time).

How to calculate the inverse of the covariance matrix?

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

determinant

$$\begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}^{-1} = \frac{1}{4\times6-7\times2} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

Why need consider data distribution?

Example:

Suppose we have two quizzes.

Quiz 1: std = 10

Student A – Student B = 1

Quiz 2: std = 1

Student A – Student B = 1

To compare the 2 differences fairly: Compute how many deviations away between 2 points.

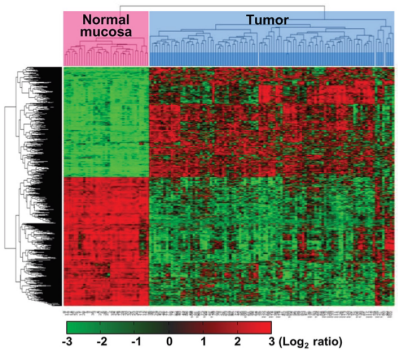
Classification

Why classification

- Characteristic of each class.
- Classify items
  - Better organization.
  - Find position to put new items.
- Classify people
  - Patients: different treatment for different groups (e.g. children, elder)
  - Customers: whether a person within the targeting group.
- In biology
  - Given a new gene expression profile, predict normal or tumor.

What is classification?

- Given a collection of records (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

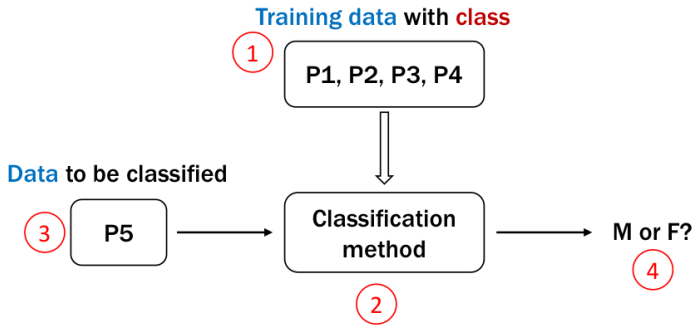


Person	Height(m)	Weight(kg)	Gender
P1	1.79	75	M
P2	1.64	54	F
P3	1.70	63	M
P4	1.88	78	M
P5	1.75	70	??

Find some rules in it

Predict gender based on height and weight.

How to do classification?



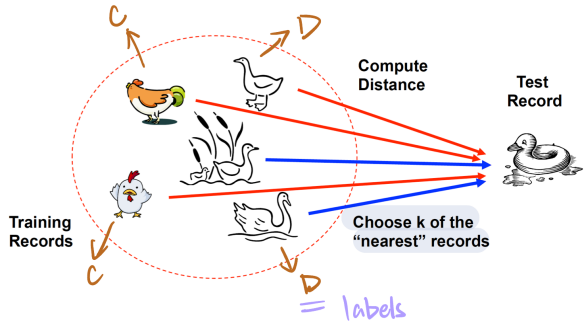
Necessary things for classification:

- Training data with class.
- Classification method / algorithm
- Data to be classified

K-nearest neighbors

Basic idea:

If it walks like a duck, quacks like a duck, then it's probably a duck.

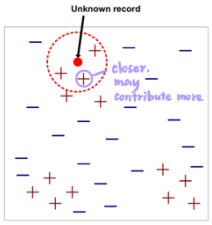


KNN:

A simple algorithm that stores all available instances and classifies new instances based on a distance metric to the available ones.

Training process:

Store the available training instances.



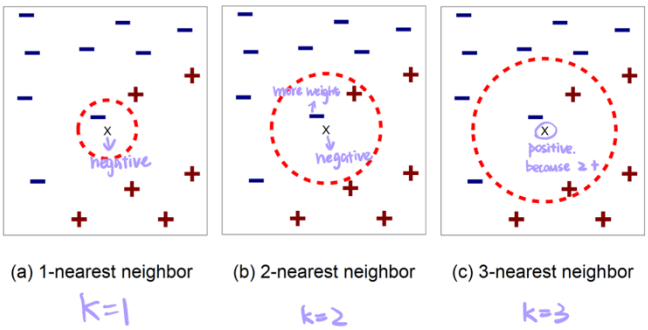
Predicting process:

- Find the K training instances that are closest to the query instance.
- Return the most frequent class label among those K instances.

Data should be normalized!

Factors needed to determine when using KNN:

- A distance matric
  - Cosing similarity
  - Correlation
  - Euclidean distance
  - Manhattan distance
  - Mahalanobis distance
- How many neighbors to look at (K)
- A weighing function (optional)



Be careful K>N!

How should we choose K?

- In practice, using a value of K somewhere between 5 and 10 gives good results for most low-dimensional data sets.

- A good K can also be chosen by using cross-validation.

The standard procedure of KNN

Suppose we have chosen the distance matric and K.

1. Normalization
2. Compute distance
3. Identify the K most similar data
4. Take their class out and find the mode class

A running example of KNN

Suppose we have chosen the Euclidean distance matric and K = 2.

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 Euclidean distance

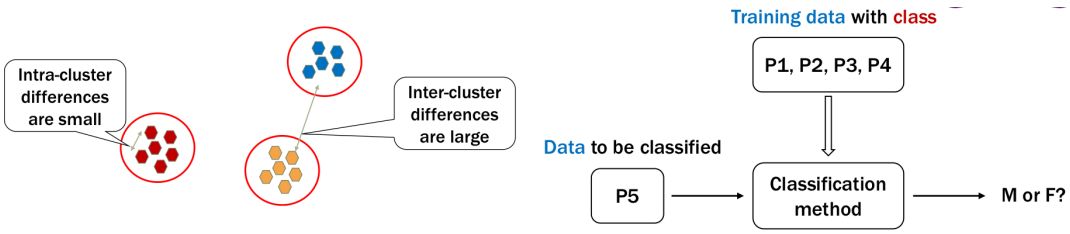
Person	P5	Gender
P1	0.267	M
P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

3. Identify the K (= 2) 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.
   - M.

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 and supervised learning

- 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