BMEG3105 - Data analytics for personalized genomics & precision medicine
Professor: Yu Li
Scribing L8 – Clustering & Classification
Wed 02-10-2024

Clustering & Classification

1 Clustering

1.1 why clustering

- goal
  - better organise things
  - get faster searching
- very common & almost everywhere
- clustering in biology
  - cluster genes
    - identify co-expressed genes that are involved in same pathway
    - identify differentially expressed genes related to diseases
  - cluster samples & cells
    - identify new disease sub-types
      - very useful to develop more effective/efficient medicine
      - part of developing precision medicine
    - identify new cell types
- e.g. shopping sites: shop by category
- e.g. cluster people
  - patients
    - different treatment for different groups
    - children vs. elderly → hospital specialisations
  - customers
    - different groups with different needs
    - not necessarily grouping people by age or gender
    - optimise product based on need of targeting group

1.2 clustering analysis

- definition: 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 (within a group)
  - inter cluster differences are large (between 2 groups)
- it's everywhere
  - used for understanding
    - as stand-alone tool to get insight into data distribution
    - as pre-processing step for other algorithms
    - e.g. group related documents for browsing
      - group genes & proteins that have similar functionality
      - group stocks with similar price fluctuations
      - discover new groups → e.g. cell types
  - for summarisation

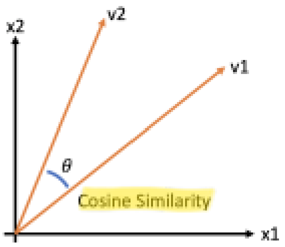
        - to reduce size of large data sets
        - to preserve privacy → e.g. in medical data
- principle
    - needed:
        - data to be clustered
        - similarity measurement
        - clustering algorithm = executive procedure
    - give pictures to computer
    - change pictures to data → pixels with values = matrix
    - cluster methods
    - output: clustering indicator

1.3 similarity & dissimilarity

- Similarity
    - Numerical measure of **how alike** two data objects are
    - Higher when objects are more alike
    - Often falls in the range [0,1]
- Dissimilarity (distance)
    - Numerical measure of **how different** are two data objects
    - Lower when objects are more alike
    - **Minimum** dissimilarity is often **0**
    - **Upper limit varies**

1.3.1 Cosine similarity

- If d1 & d2 are 2 vectors then: cos(d1,d2) = $\frac{d1*d2}{(|d1|*|d2|)}$
    - With * indicating vector dot product
    - With |d| the length of vector d



    - E.g.

$$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$
$$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$$\cos(d_1, d_2) = 0.3150$$

1.3.2 Correlation

- Measures linear relationship between objects
- Shows whether 2 properties/variables are changing together or not
- Formula: $\rho X, Y = corr(X, Y) = \frac{cov(X,Y)}{\sigma x \sigma y} = \frac{E[(X-\mu x)*(Y-\mu y)}{\sigma x \sigma y}$
    - With cov = covariance
    - With sigma = standard deviation of x and y
    - With E(…) = expectation of each point minus the variation
- Principle:
    - Calculate means: μx and μy
    - Subtract means: x – μx = a and y – μy = b
    - Calculate ab, a^2 and b^2
    - Sum up
    - Calculate correlation

1.3.3 Euclidian distance

- Formula: $Ed(p, q) = \sqrt{\sum_{k=1}^{m}(pk - qk)^2}$
    - With m = number of dimensions
    - With pk and qk = k-th attributes/components/data objects p and q
- Normalisation necessary because scales of different dimensions differ
- Results in distance matrix with each cell containing a data point

1.3.4 Minkowski distance

- Generalisation of Euclidian distance
- Formula: $dist(p, q) = (\sum_{k=1}^{m} |pk - qk|^r)^{\frac{1}{r}}$
    - With r = parameter
    - With m = number of dimensions/attributes
    - With pk & qk = k-th attributes/components/data objects p & q
- Special cases
    - r = 1
        - Manhattan distance/city block distance/taxicab/L1 norm
        - Common example: Hamming distance = number of bits that are different between 2 binary vectors
    - r = 2
        - Euclidian distance
    - r = ∞
        - supremum
        - L max norm
        - Maximum difference between any component of the vectors
        - Lim(r → ∞) = |pk – qk|

1.4 Hierarchical clustering

- Rearrangement of entire columns to organise data
- Goal is to identify groups/clusters withing the gene expression matrix

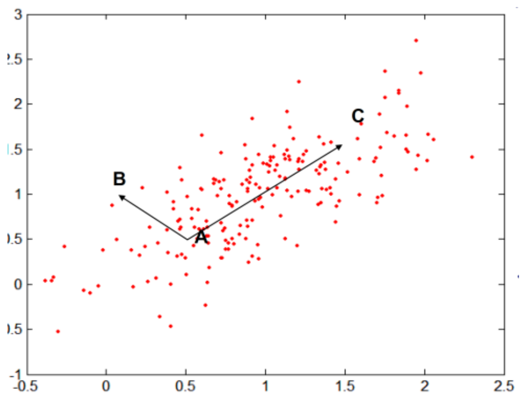
- Produces set of nested clusters organised as hierarchical tree
- Can be visualized as dendrogram = tree like diagram that records sequences of merges
- May correspond to meaningful taxonomies
    - Gene clusters
    - Phylogeny reconstruction
    - Animal kingdom
- Steps
    - Compute the Similarity or Distance matrix
    - Let each data point be a cluster
    - Merge the two closest clusters
    - Update the similarity or distance matrix
        - Need for original data matrix depends on how the distance is defined
            - For min/max: no need for original data matrix
            - For centroids: requires original data matrix
    - Repeat until only one single cluster remains
- How to update distance matrix after merging
    - 4 possibilities
        - As long as you can make your assumption clear it's fine
        - Minimum
        - Maximum
        - Group average
        - Distance between centroids = between middle point of 2 clusters

1.5 Mahalanobis distance

- Calculates the distance considering the data distribution
- Formula: $mahalanobis\ (p, q) = (p - q)^T * \sum^{-1}$ (p – q)
    - With $\sum$ = covariance matrix
    - With -1 showing that the matrix is inverted:

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

- Useful when scales of 2 tests are not the same
    - Alternative: min-max normalisation of the tests to get them on the same scale
        - This will give the same std for both tests
    - E.g. comparing the results of 2 students on 2 quizzes with different standard deviation
        - On both quizzes person A has 1 point more than student B
        - However the difference for them between both quizzes is not the same as on quiz 1 the std = 10 and on quiz the std = 1
        - So for quiz 2 student A can be the best of the class while student B is the worst of the class
- E.g. AB seems intuitively smaller than AC when just looking at the combining line, however it can be bigger than AC when considering distribution of the data set



Covariance matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)
B: (0, 1)
C: (1.5, 1.5)

Mahal(A,B) = 5
Mahal(A,C) = 4

1.6 Programming:

- Scikit-learn: https://scikit-learn.org/stable/

2 Classification

2.1 Why classification

- Goals:
  - To determine characteristics of each class
    - E.g. when learning words as a kid you do this by classification
  - To classify items
    - In order to get a better organisation
    - In order to know where to put new items
  - To classify people
    - Patients → different treatment for different groups: elderly vs. kids
    - Customers → is the person within the targeting group or not
- Why classification in biology
  - To determine whether a new gene expression profile is normal or a tumour

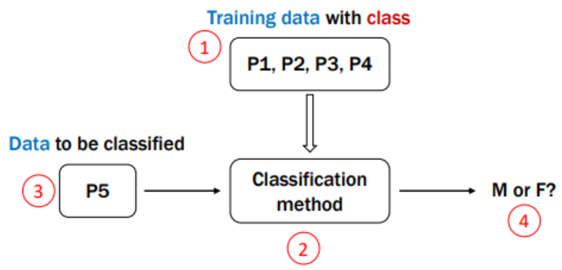
2.2 What is classification

- Find a method to assign the class of previously unseen records based on their other attributes + training set as accurately as possible
  - Start from a training set/collection of records
  - Each record contains a set of features/attributes
  - One of the attributes is the class
- E.g. try to predict someone's gender, of which only height and weight are known, based on a training set with people with known height, weight & gender
  - The gender of the people in the training set is essential to allow determination of the unknown gender

2.3 How to do classification

- Needed:

    - Training set with class
    - Classification method
    - Data to be classified
- Principle:
    - Start from training set with class
    - Develop classification method based on the training set
    - Collect data to be classified
    - Classify the data using the developed classification method



2.4 K-nearest neighbours (KNN)

- Simple algorithm that stores all available instances & classifies new instances based on distance metric to available ones
- Data should be normalised!
- 2 phases
    - Training process = storing available training instances
    - Predicting process
        - Finding K training instances that are closest to query instance
        - Return most frequent class label among those K instances
- What to determine when using KNN
    - Distance metric
        - Cosine similarity
        - Correlation
        - Euclidian distance
        - Manhattan distance
        - Mahalanobis distance
    - How many neighbours to look at = K → number depends on definition
    - Weighing function
        - Optional
        - Different properties should be considered
        - Distance between data points can count, e.g. nearest neighbour will be more important than the 4^{th} or 10^{th} neighbour
- How to choose K
    - In practice:
        - Use value of K somewhere between 5 & 10 to get good result for most low-dimensional data sets
        - Good K can also be chosen by using cross validation (see later)
- Principle
    - Choose a distance metric & K

- Normalise data
- Compute distances
- Identify K most similar data
- Take class out & find mode class
- E.g.
  - Euclidian 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	??

  - Min-max normalisation of data

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	??

  - Compute distances of the normalised data with unknown person

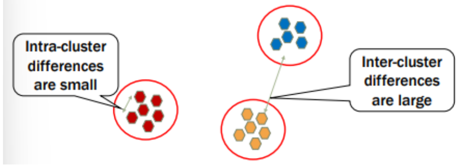
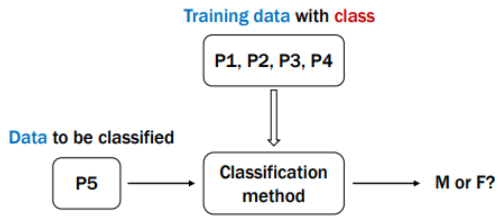
Person	P5	Gender
P1	0.267	M
P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

  - Identify 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	??

  - Take class out: male

2.5 Clustering vs. classification

	Clustering	Classification
		
What needed	-data to be clustered -similarity measurement -clustering algorithm = executive procedure	-training data with class -classification method -data to be classified
Goal	Find similarity/clusters in data	Assign class to new data point
Data	Without class →since we try to identify the category	1)Training data with class =annotated data 2)Testing data without class
Classes	Unknown number	Known number
Output	Cluster index for each point	Class assignment of testing data
Algorithm	One phase →put clustering index	Two phases →training: get classification method →application: get result

2.6 Unsupervised vs. supervised learning

- Unsupervised learning
    - Machine learning algorithms to analyse and cluster unlabelled data
    - E.g. clustering and dimension reduction
- Supervised learning
    - Machine learning algorithms to classify and predict outcomes, trained on labelled data
    - Annotation of data to guid us through the problem
    - E.g. classification and regression
- Biggest difference is the data

2.6.1 Machine learning