Scribing 1 for lecture 8

Chan Chun Ming Tony
1155156523

Scribing agenda

Classification

- Why classification?
- What is classification?
- How to do classification?

K-nearest neighbour classification

Clustering VS classification

Why classification?

Characteristics of each class

- Classify items

- Better organization
- Where to put the new items?

- Classify people

- Patients: different treatment for different groups

E.G. Children, elder

- Customers:

E.G. Is the person within the targeting group?

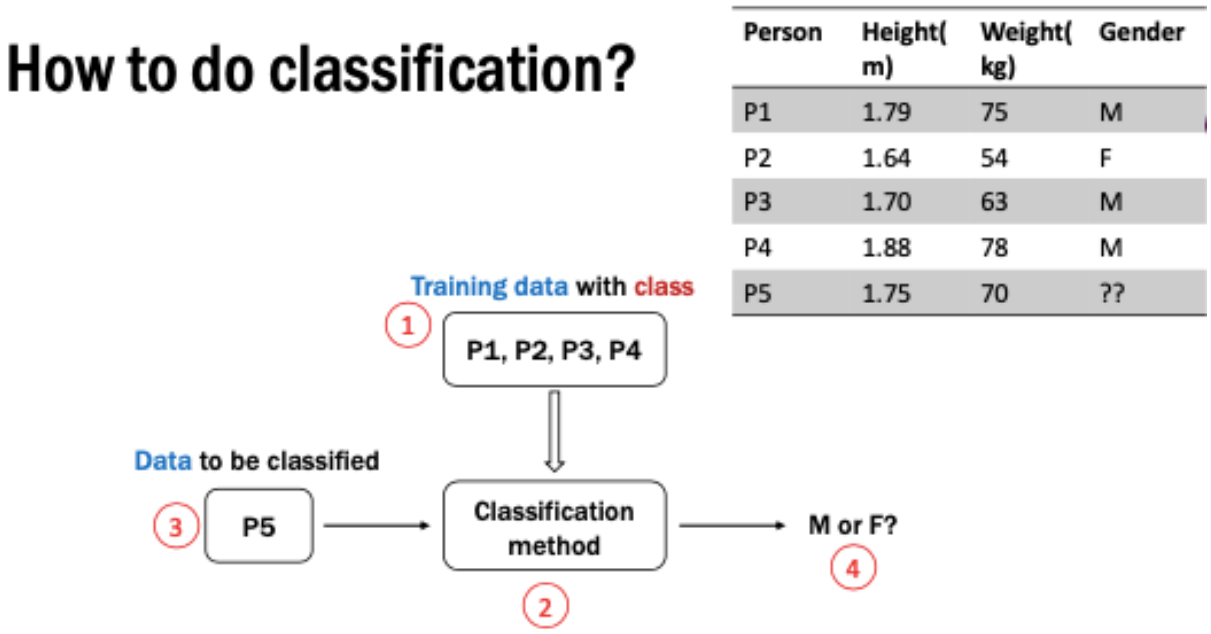
What is classification?

- Given a collection of records (training set)

  - Each record contains a set of attributes, one of the attributes is the class

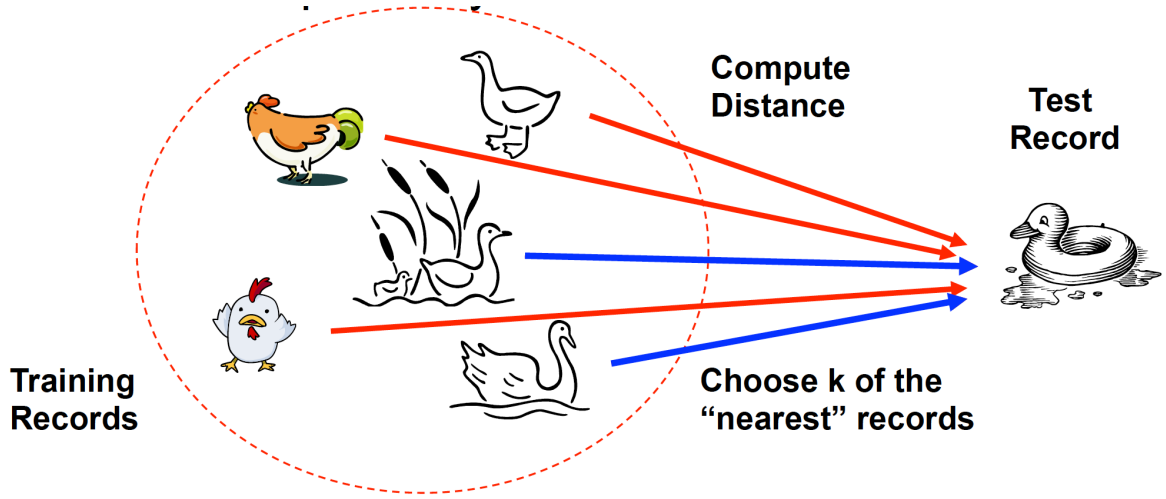
- Find a method to assign the class of previously unseen records based on their other attributes and the training set as accurately as possible

How to do classification?



K-nearest neighbours

KNN is 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

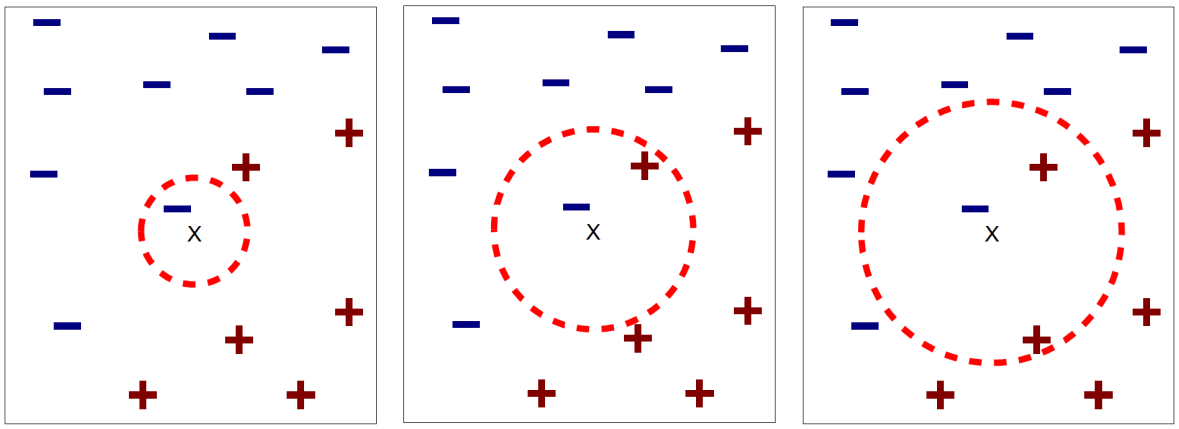
Data should be normalized

What should we determine when using KNN?

A distance metric

How many neighbours to look at (K)

A weighing function (optional)



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

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

The standard procedure of KNN

Suppose we have chosen the distance metric and K

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

Here's an example:

Suppose we have chosen the Euclidean distance and 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	??

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 (m)	Weight (kg)	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

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

Identify the K most similar data

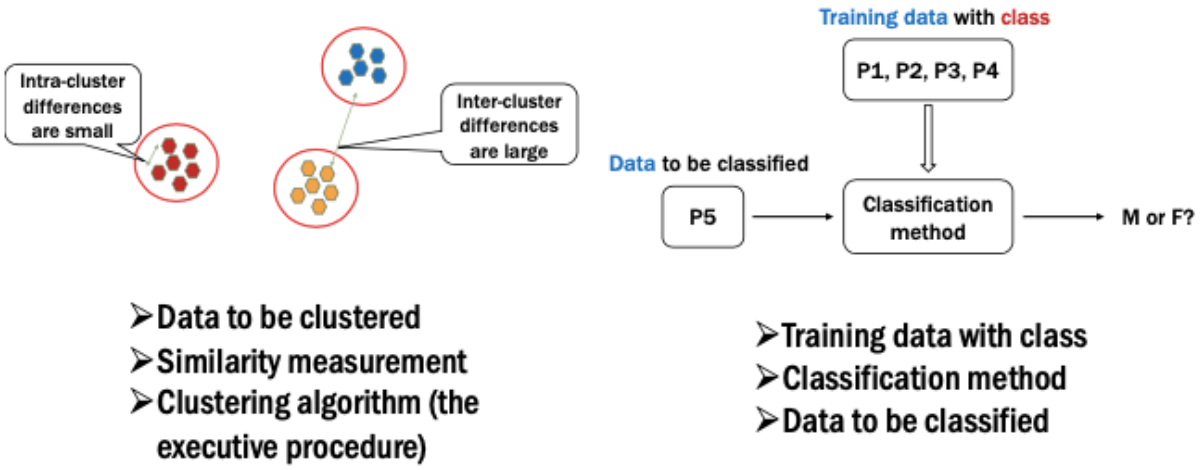
Person	P5	Gender
P1	0.267	M

P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

Take their class out and find the mode class

M

Clustering VS Classification



- Data to be clustered
- Similarity measurement
- Clustering algorithm (the executive procedure)

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

	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 analyse and cluster unlabelled data

Example: clustering and dimension reduction

Supervised learning
Machine learning algorithms to classify and predict outcomes, trained on labelled data
Example: classification and regression