

Scribing Lecture 9 (classification)

① Classification

→ classify items, people into groups

→ for better organization, prediction of new data

eg classify a new gene expression profile into normal and tumor

Definition:

- given a collection of records
training set
- each records contains a set of attributes, one of the attributes is the class (what we want to predict)

eg

Person	weight	height	Gender
P1	75	1.79	M
P2	54	1.64	F
P3	63	1.70	M
P4	78	1.88	M
P5	70	1.75	??

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

→ method can be different

eg k-nearest neighbor classification (KNN),

Decision tree, neural networks

Steps :

① training data with class

Classification
method (KNN)

Person	weight	height	Gender
P1	75	1.79	M
P2	54	1.64	F
P3	63	1.70	M
P4	78	1.88	M
P5	70	1.75	??

② classify the unseen data

P5 → Classification
method (KNN) → result (M/F)

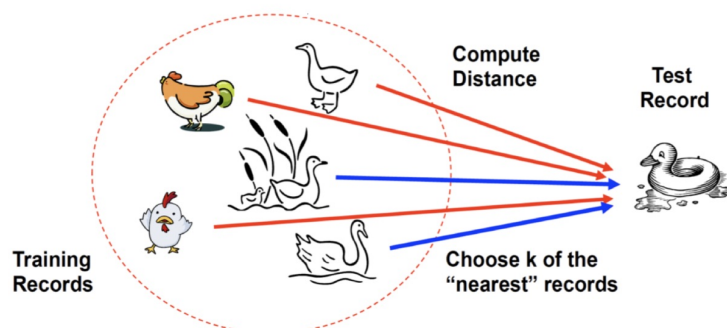
3 key components :

- ① training data with class
- ② classification method
- ③ Data to be classify

② K-nearest neighbors

basic idea : compare the similarity of the new record with training records.

in detail : it stores all available instances and classifies new instances based on a distance matrix to the available ones



steps :

① training

→ store the available training instances

② Predicting

→ find the K training instances that are closest to the query instance

→ return the most frequent class label among those K instances

③ Data should be normalized

We should determine :

① a distance matrix

② How many neighbor to look at → K

③ A weighing function

~~A~~ K

5-10 for most low-dimensional data sets
can be chosen by using cross-validation

Standard procedure

→ chosen distance metric and K

① normalization

② Compute distance

③ Identify the K most similar data

④ Take their class out and find the mode class

Clustering vs classification

Clustering

Goal find similarity

Data Data without class

Classes Unknown number of classes

Output The cluster index for each point

Algorithm One phase

Classification

Assign class to new data

Training data with class and testing data without class

known number of classes

the class assignment of the testing data

two phases

unsupervised learning

analyse and cluster unlabelled data

supervised learning

classify and predict outcomes, trained on labeled data