Classification & Clustering

Dr Pradipta Biswas, PhD (Cantab)
Assistant Professor
Indian Institute of Science
http://cpdm.iisc.ernet.in/PBiswas.htm

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

2

Classification and Prediction

- What is classification? What is regression?
- Issues regarding classification and prediction
- Classification by decision tree induction

3

Classification vs. Prediction

- Classification:
  - predicts categorical class labels
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Regression:
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

4

Why Classification? A motivating application

- Credit approval
  - A bank wants to classify its customers based on whether they are expected to pay back their approved loans
  - The history of past customers is used to train the classifier
  - The classifier provides rules, which identify potentially reliable future customers
  - Classification rule:
    - If age = "31…40" and income = high then credit_rating = excellent
  - Future customers
    - Paul: age = 35, income = high ⇒ excellent credit rating
    - John: age = 20, income = medium ⇒ fair credit rating

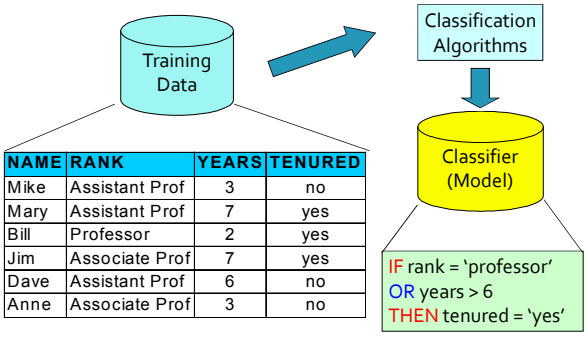
5

Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction: training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test samples is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur

6

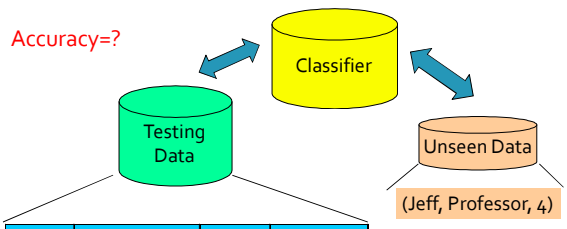
Classification Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

7

Classification Process (2): Use the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Mellisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

8

Classification by Decision Tree Induction

- Decision tree
  - A flow-chart-like tree structure
  - Internal node denotes a test on an attribute
  - Branch represents an outcome of the test
  - Leaf nodes represent class labels or class distribution
- Decision tree generation consists of two phases
  - Tree construction
    - At start, all the training examples are at the root
    - Partition examples recursively based on selected attributes
  - Tree pruning
    - Identify and remove branches that reflect noise or outliers
- Use of decision tree: Classifying an unknown sample
  - Test the attribute values of the sample against the decision tree

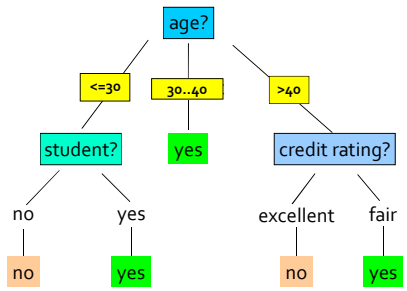
9

Training Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

10

Output: A Decision Tree for "*buys_computer*"



11

Neural Network

- Here x1 and x2 are normalized attribute value of data.
- y is the output of the neuron , i.e the class label.
- x1 and x2 values multiplied by weight values w1 and w2 are input to the neuron x.
- Value of x1 is multiplied by a weight w1 and values of x2 is multiplied by a weight w2.
- Given that
  - w1 = 0.5 and w2 = 0.5
  - Say value of x1 is 0.3 and value of x2 is 0.8,
  - So, weighted sum is :
  - sum= w1 x x1 + w2 x x2 = 0.5 x 0.3 + 0.5 x 0.8 = 0.55

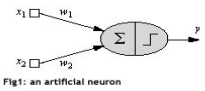


Fig1: an artificial neuron

One Neuron as a Network

- The neuron receives the weighted sum as input and calculates the output as a function of input as follows :
- y = f(x) , where f(x) is defined as
- f(x) = 0 [when x< 0.5]
- f(x) = 1 [when x >= 0.5]
- For our example, x (weighted sum) is 0.55, so y = 1 ,
- That means corresponding input attribute values are classified in class 1.
- If for another input values , x = 0.45 , then f(x) = 0,
- so we could conclude that input values are classified to class 0.

Neuron with Activation

- The neuron is the basic information processing unit of a NN. It consists of:

1 A set of links, describing the neuron inputs, with weights W_1, W_2, …, W_m

2. An adder function (linear combiner) for computing the weighted sum of the inputs (real numbers):

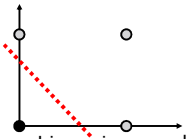
$$u = \sum_{j=1}^{m} w_j x_j$$

3 Activation function : for limiting the amplitude of the neuron output.

$$y = \varphi(u + b)$$

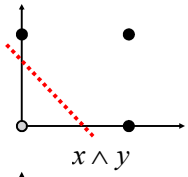
Why We Need Multi Layer ?

- Linear Separable:



- Linear inseparable:

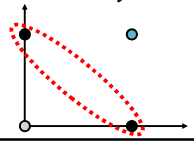
$x \vee y$



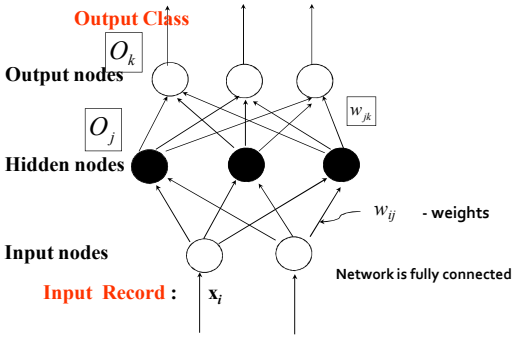
$x \wedge y$

- Solution?

$x \veebar y$



A Multilayer Feed-Forward Neural Network



Neural Network Learning

- The inputs are fed simultaneously into the input layer.
- The weighted outputs of these units are fed into hidden layer.
- The weighted outputs of the last hidden layer are inputs to units making up the output layer.

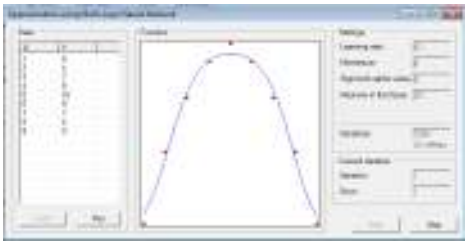
A Multilayer Feed Forward Network

- The units in the hidden layers and output layer are sometimes referred to as neurodes, due to their symbolic biological basis, or as output units.
- A network containing two hidden layers is called a three-layer neural network, and so on.
- The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

Classification by Back propagation

- *Back Propagation learns by iteratively processing a set of training data (samples).*
- For each sample, weights are modified to minimize the error between network's classification and actual classification.

Demonstration



20

Validation

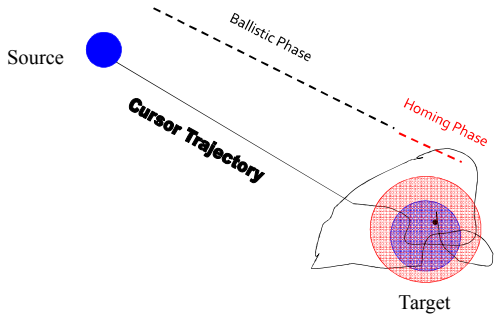
- Cross Validation (10-fold)
  - Randomly divide training data set in 10 segments
  - Train with 9 and test on remaining 1
  - Repeat the procedure 10 times
  - Training sample should be balanced
    - Nearly equal number of all possible classes
- Leave-1-out Validation: same as above, we take one sample as test set and train with the rest

21

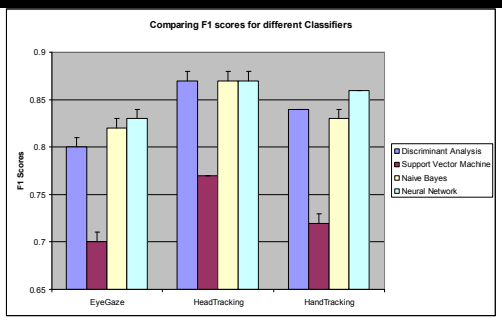
Case Study – Classification for Target Prediction

22

Analysing Trajectory

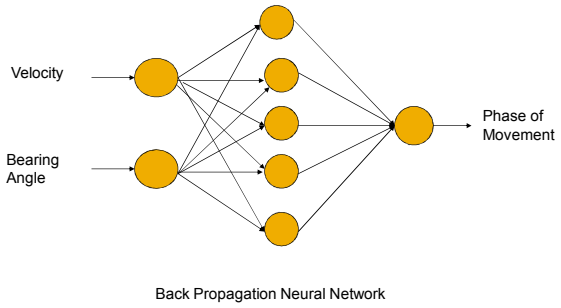


Classifier Result



Engineering Design Centre

Neural Network



Back Propagation Neural Network

25

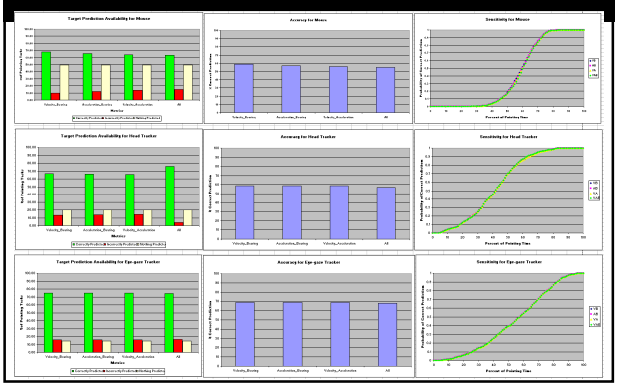
Algorithm – Neural Network

- For every change in position of pointer in screen
  - Calculate angle of movement
  - Calculate velocity of movement
  - Calculate acceleration of movement
- Run Neural Network with Angle, Velocity and Acceleration
- Check output
- If output predicts homing phase
  - Find direction of movement
  - Find nearest target from current location towards direction of movement

Evaluation Criteria

- **Availability**: In how many pointing tasks the algorithm makes a successful prediction.
- **Accuracy**: Percentage of correct prediction among all predictions
- **Sensitivity**: How quickly an algorithm can detect intended target

Results

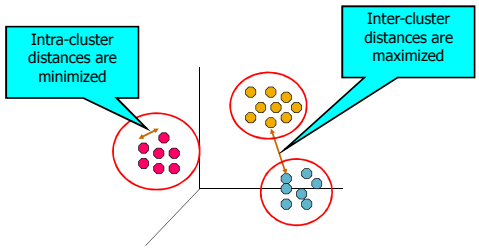


Demonstration

Cluster Analysis

What is Cluster 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



31

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Clustering is used:
  - As a stand-alone tool to get insight into data distribution
    - Visualization of clusters may unveil important information
  - As a preprocessing step for other algorithms
    - Efficient indexing or compression often relies on clustering

32

Applications of Clustering

- Pattern Recognition
- Image Processing
  - cluster images based on their visual content
- Bio-informatics
- WWW and IR
  - document classification
  - cluster Weblog data to discover groups of similar access patterns

33

Similarity and Dissimilarity Between Objects

- Distance metrics are normally used to measure the similarity or dissimilarity between two data objects
- The most popular conform to *Minkowski distance*:

$$L_p(i,j)=\left(|x_{i1}-x_{j1}|^p+|x_{i2}-x_{j2}|^p+...+|x_{in}-x_{jn}|^p\right)^{1/p}$$

where $i = (x_{i1}, x_{i2}, ..., x_{in})$ and $j = (x_{j1}, x_{j2}, ..., x_{jn})$ are two n-dimensional data objects, and p is a positive integer

- If $p = 1$, L_1 is the Manhattan (or city block) distance:

$$L_1(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+..+|x_{in}-x_{jn}|$$

34

Similarity and Dissimilarity Between Objects (Cont.)

- *If* $p = 2$, L_2 is the Euclidean distance:

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{in}-x_{jn}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also one can use weighted distance:

$$d(i,j)=\sqrt{(w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+...+w_n|x_{in}-x_{jn}|^2)}$$

35

Major Clustering Approaches

- Partitioning algorithms: Construct random partitions and then iteratively refine them by some criterion
- Hierarchical algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

36

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

37

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

38

Limitations of K-means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-spherical shapes
- K-means has problems when the data contains outliers. Why?

39

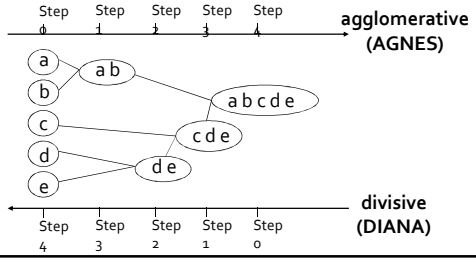
The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling

40

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



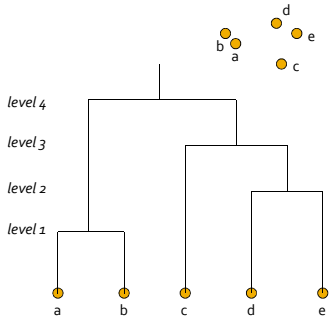
41

A *Dendrogram* Shows How the Clusters are Merged Hierarchically

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

E.g., level 1 gives 4 clusters: {a,b},{c},{d},{e},
level 2 gives 3 clusters: {a,b},{c},{d,e}
level 3 gives 2 clusters: {a,b},{c,d,e}, etc.



42

Soft Clustering

- What happens when we can not specify the optimum number of clusters beforehand
- Can we find the optimum number of clusters?
- Two methods can return overlapping clusters
  - Fuzzy c-means
  - EM Clustering algorithm

43

Fuzzy c-means

- Place a set of cluster centres
- Assign a fuzzy membership to each data point depending on distance
- Compute the new centre of each class
- Termination is based on an objective function
- Returns cluster centres and membership values of each data point to each cluster

44

EM Algorithm

- Assume data came from a set of Gaussian Distribution
- Assign data points to distributions and find Expected probability
- Update mean and std dev of distributions to Maximize probabilities

45

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

46

Summary

- Classification and Clustering
  - Decision tree and neural network for classification
  - Cross validation
  - Hierarchical & K-means clustering
  - Soft Clustering
  - Cluster Validation Index
  - Case studies on IUI

47