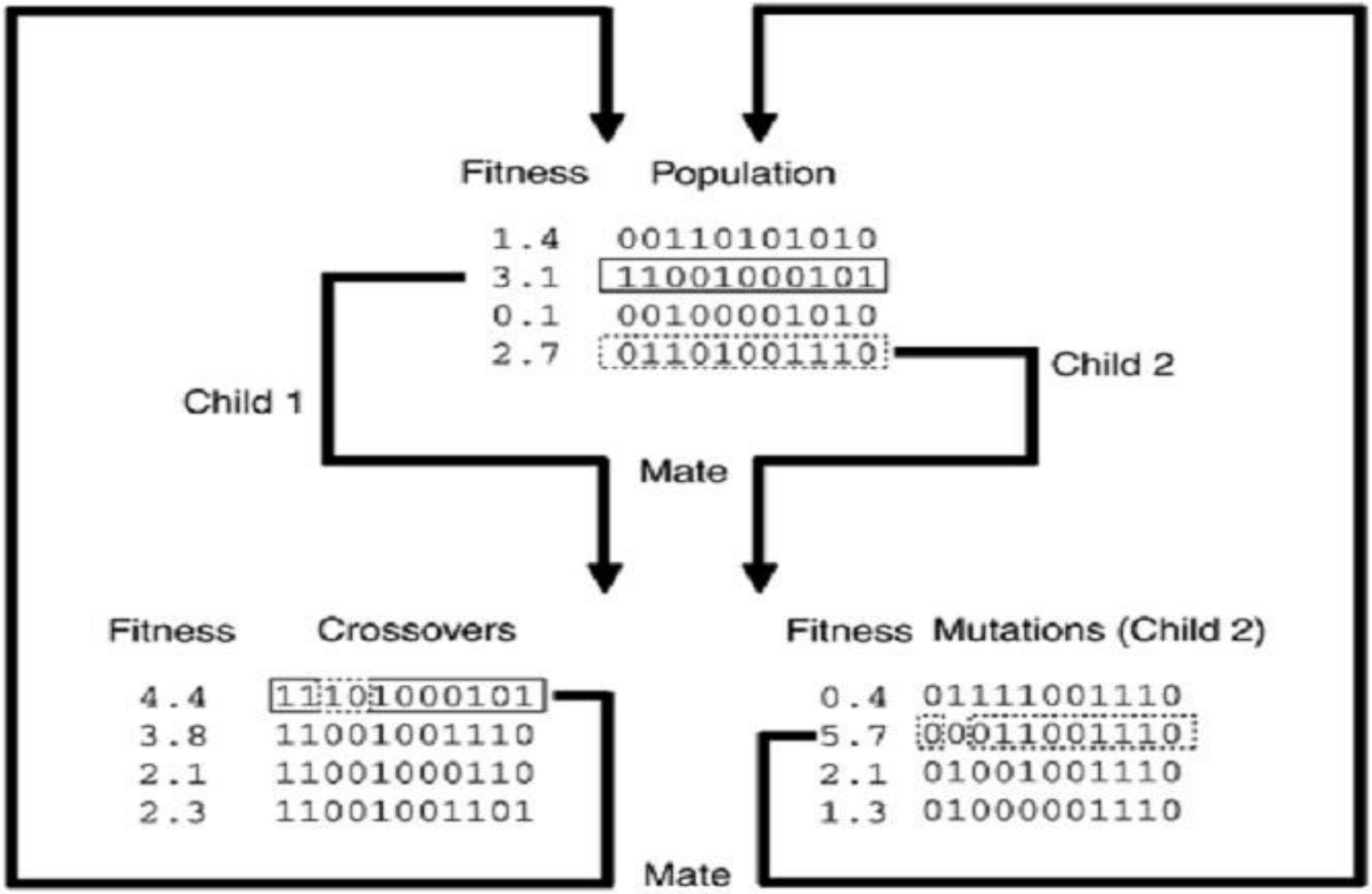
Datamining Models & Algorithms

MBI304- Data Mining & Data Analytics

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

Figure 7-8. Genetic Algorithm Operation.



Genetic algorithm

- Genetic algorithm guide the learning process of data mining algorithm
- Genetic algorithm act as a method for performing a guided search or good models in the solution space

- They are called genetic algorithm because the loosely follow the pattern of biological evolution in which the member of one generation (of models) compete to pass on their characteristics to the next generation (of models, until the best (model) is found
- The information to be passed on is contained in chromosomes, which contain the parameter for building the model

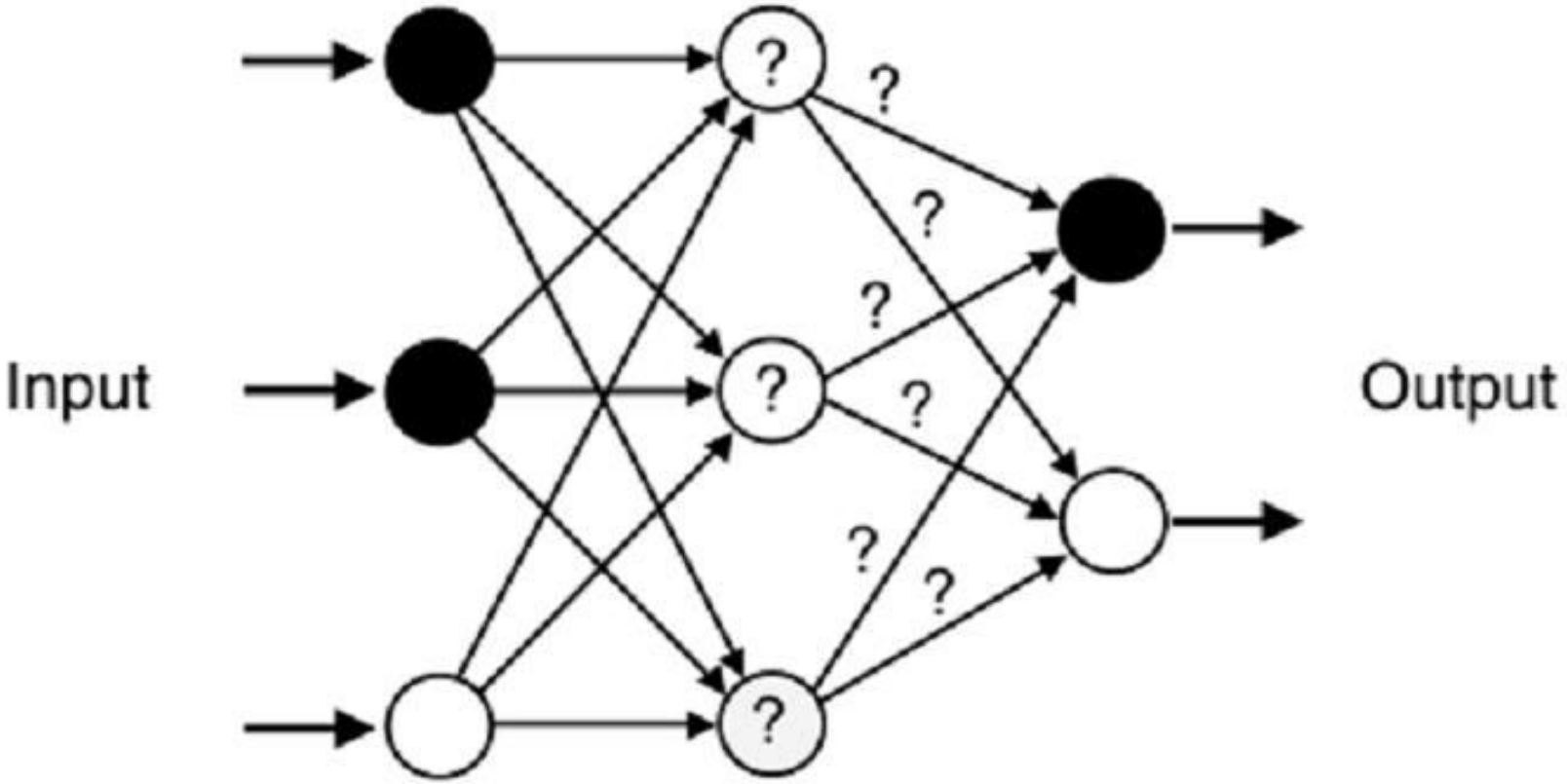
Neural network

- Neural networks are particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.
- A neural network start with an input layer, where each node correspond to a predictor variable connected to hidden layer that is further connected with output layer

Response variable, dependent or target variable

- The output layer consist of one or more response variable
- In predictive models the value or classes are we are predicting are called the response variable, dependent or target variable.
- The value used to make the prediction are called the predictor or dependent variable

Figure 7-9. Neural Network. One of the limitations of a neural network is that the significance of the strength of the internal interconnections is unknown. As a result, as a pattern recognizer or categorizer, the neural network can be treated as a black box.



Connection weight

- After the input layer, each node takes in a set of inputs, multiplies them by connection weight Wxy(e.g., the weight from node 1 to 3 is W_{13} see Figure 5), add them together, applies a function to them13(called the activation or squashing function) and passes the output to the nodes in the next layer.
- Connection weight are unknown parameters which are estimated by training method

- For example, the value passed from node 4 to node 6 is:
- Activation function applied to ([W * value of node 1] + [W * value of node 2])

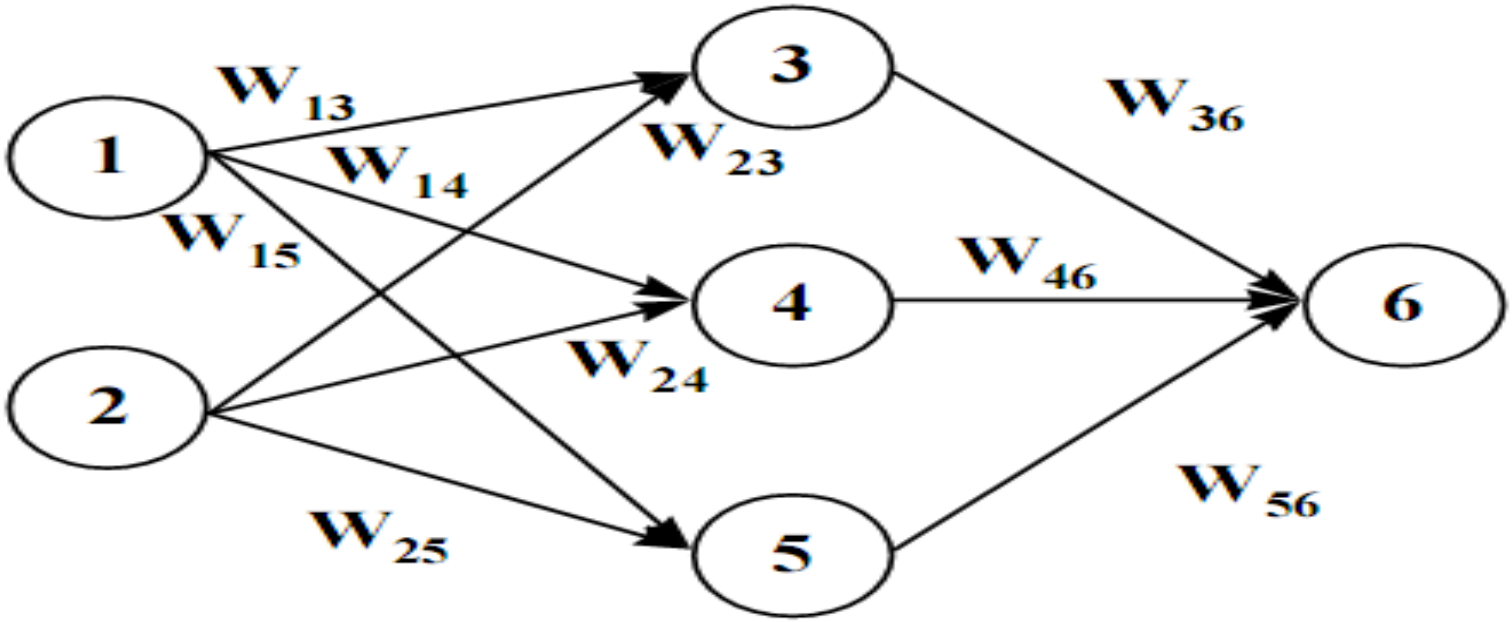


Figure 5. W_{xy} is the weight from node x to node y.

- **Feed forward**: The value of the output node is calculated based on the input node values and a set of initial weights.
- **Backpropagation:** The error in the output is computed by finding the difference between the calculated output and the desired output .

This process is repeated for each row in the training set.
Each pass through all rows in the training set
is called an epoch

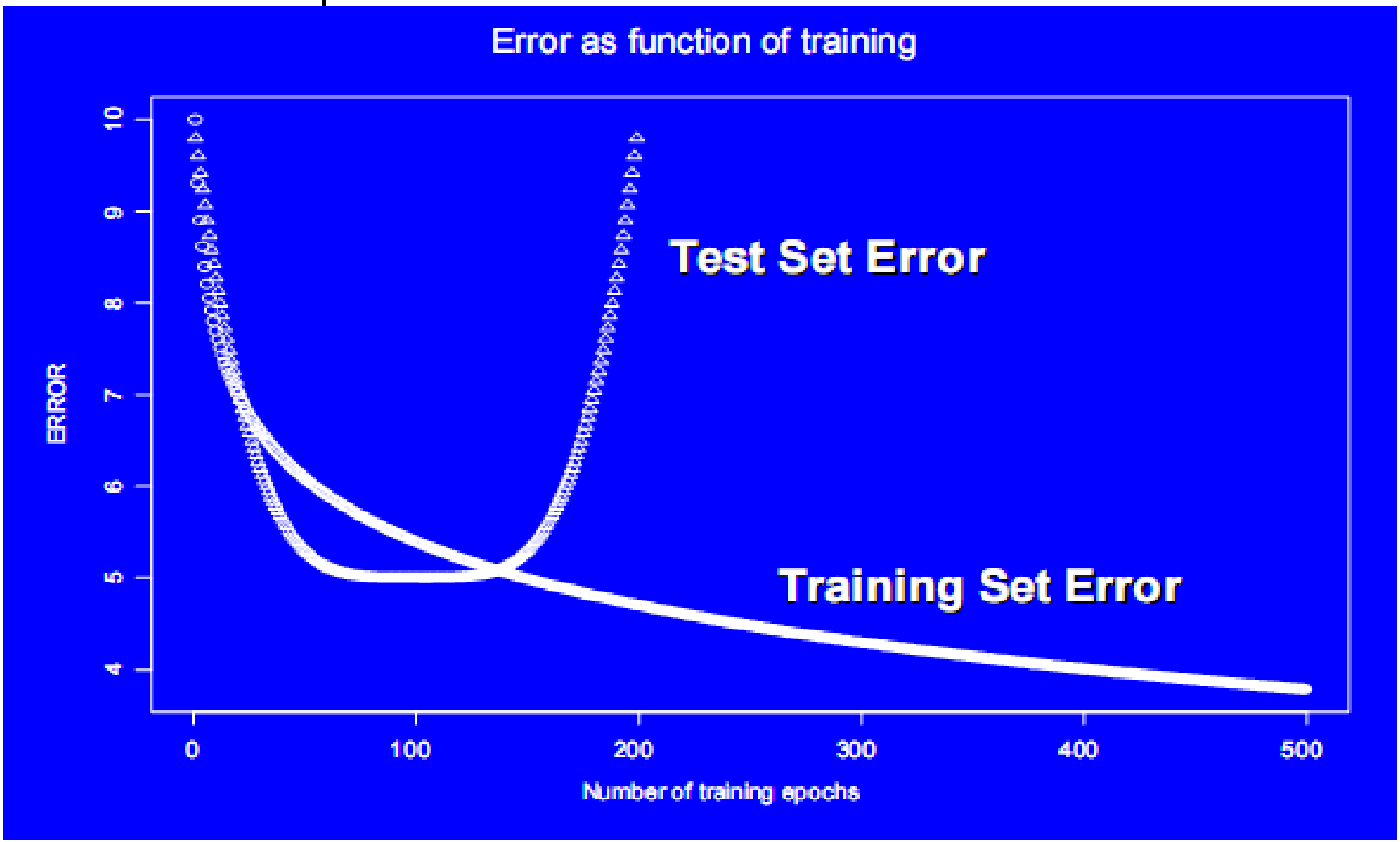
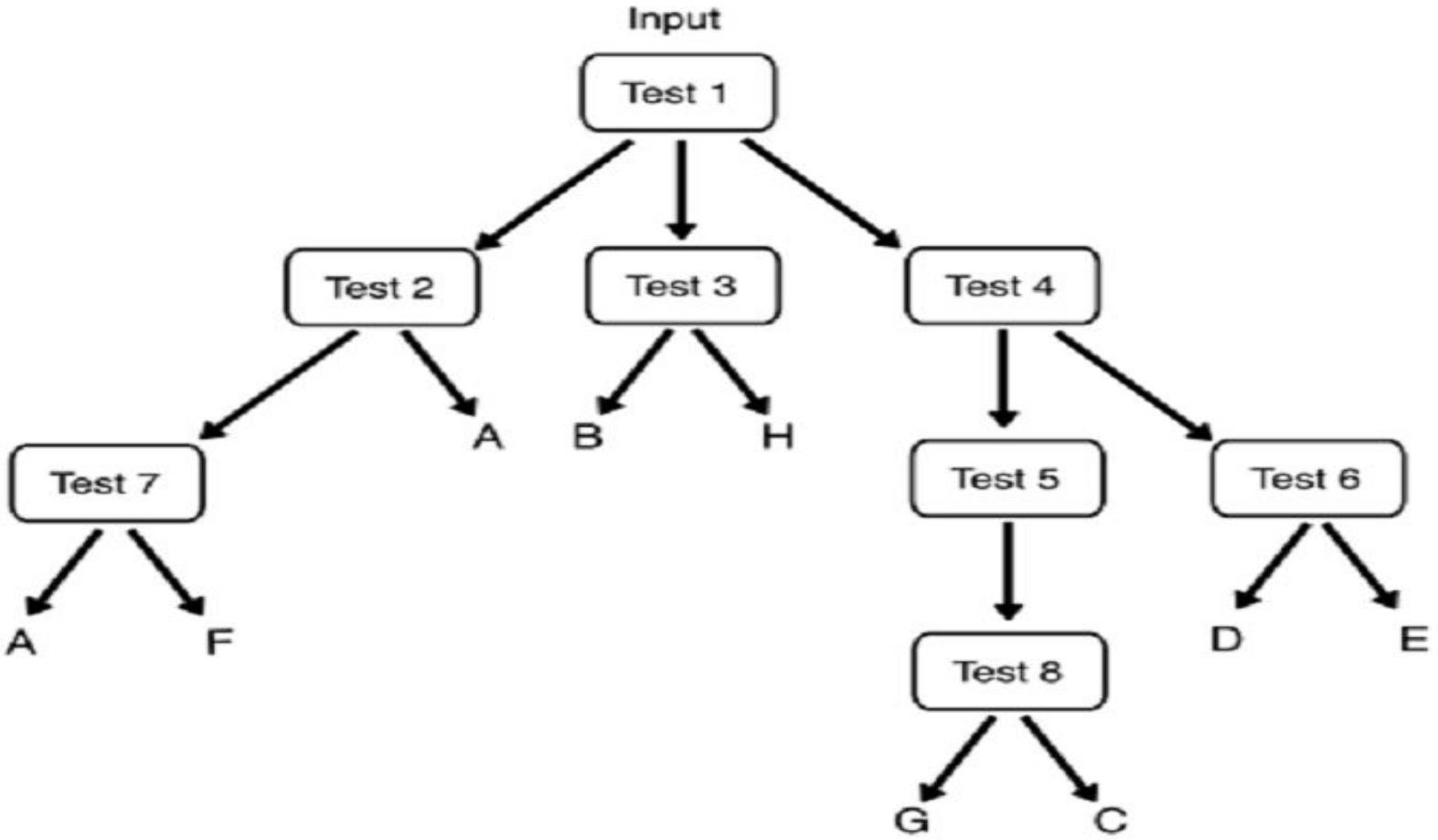


Figure 6. Error rate as a function of the number of epochs in a neural net.

Figure 7-12. Decision Trees. A decision tree categorizes a pattern by filtering it down through the tests in a tree.

Decision trees are a way of representing a series of rules that lead to a class or value.



- Depending on the algorithm, each node may have two or more branches.
- For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multiway tree.
- Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node.

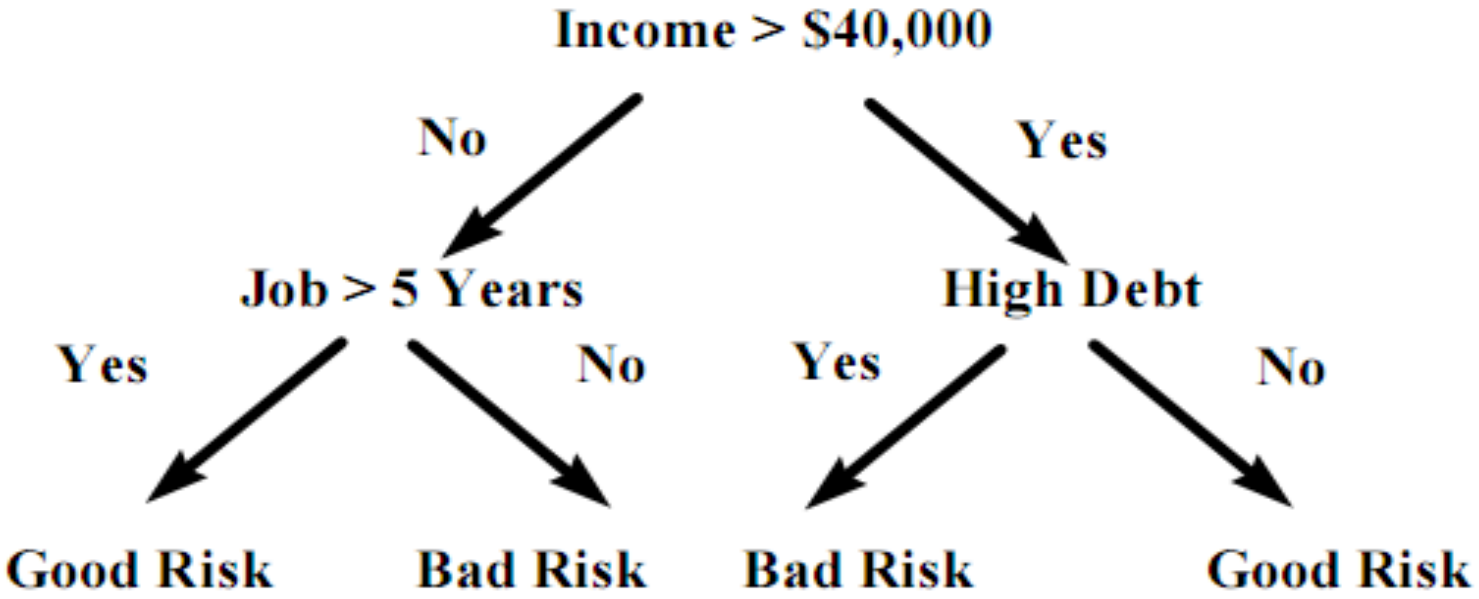


Figure 7. A simple classification tree.

- Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions.
- A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART(Classification And Regression Trees), Quest, and C5.0.
- Decision trees are grown through an iterative splitting of data into discrete groups, where the goal is to maximize the “distance” between groups at each split.

- By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached.
- Each node uses the data from the case to choose the appropriate branch.

- Decision trees which are used to predict categorical variables are called classification trees because they place instances in categories or classes.
- Decision trees used to predict continuous variables are called regression trees.
- Trees left to grow without bound take longer to build and become unintelligible,

Stopping rules

- Tree size can be controlled via stopping rules that limit growth
- An alternative to stopping rules is to prune the tree. The tree is allowed to grow to its full size and
- then, using either built-in heuristics or user intervention, the tree is pruned back to the smallest size that does not compromise accuracy.

Multivariate Adaptive Regression Splines (MARS)

In the mid-1980s one of the inventors of CART, Jerome H. Friedman, developed a method designed to address its shortcomings.

The main disadvantages he wanted to eliminate were:

- Discontinuous predictions (hard splits).
- Dependence of all splits on previous ones.
- Reduced interpretability due to interactions, especially high-order interactions.

To this end he developed the MARS algorithm. the CART disadvantages are taken care of by:

- Replacing the discontinuous branching at a node with a continuous transition modeled by a pair of straight lines.
- At the end of the model-building process, the straight lines at each node are replaced with a very smooth function called a spline.
- Not requiring that new splits be dependent on previous splits.

Rule induction

- Rule induction is a method for deriving a set of rules to classify cases.
- Although decision trees can produce a set of rules, rule induction methods generate a set of independent rules which do not
- necessarily (and are unlikely to) form a tree.

K-nearest neighbor and memory-based reasoning (MBR)

- When trying to solve new problems, people often look at solutions to similar problems that they have previously solved.
- K-nearest neighbor (k-NN) is a classification technique that uses a version of this same method.
- It decides in which class to place a new case by examining some number — the “k” in k-nearest neighbor — of the most similar cases or neighbors (Figure 8). It counts the number of cases for each class, and assigns the new case to the same class to which most of its neighbors belong.

The first thing you must do to apply k-NN is to find a measure of the distance between attributes in the data and then calculate it. While this is easy for numeric data, categorical variables need special handling. For example, what is the distance between blue and green? You must then have a way of summing the distance measures for the attributes.

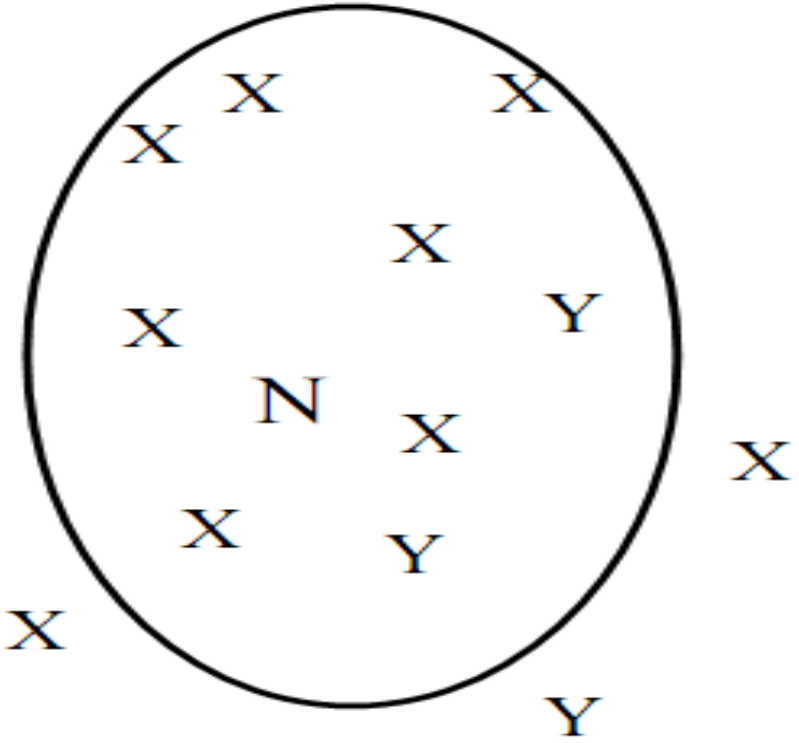


Figure 8. K-nearest neighbor. N is a new case. It would be assigned to the class X because the seven X's within the ellipse outnumber the two Y's.

Logistic regression

- Logistic regression is a generalization of linear regression. It is used primarily for predicting binary variables (with values such as yes/no or 0/1) and occasionally multi-class variables.
- Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predict whether the event itself (the response variable) will occur, we build the model to predict the logarithm of the odds of its occurrence.
- This logarithm is called the log odds or the logit transformation.
- The odds ratio:

$$\frac{\text{probability of an event occurring}}{\text{probability of the event not occurring}}$$

- Having predicted the log odds, you then take the anti-log of this number to find the odds.
- Odds of 62% would mean that the case is assigned to the class designated “1” or “yes,” for example.

Discriminant analysis

- Discriminant analysis is the oldest mathematical classification technique, having been first published by R. A. Fisher in 1936 to classify the famous Iris botanical data into three species.
- It finds hyper-planes (e.g., lines in two dimensions, planes in three, etc.) that separate the classes.
- The resultant model is very easy to interpret because all the user has to do is determine on which side of the line (or hyper-plane) a point falls. Training is simple and scalable. The technique is very sensitive to patterns in the data.

Discriminant analysis

- Discriminant analysis is not popular in data mining, however, for three main reasons. First, it assumes
- that all of the predictor variables are normally distributed (i.e., their histograms look like bell-shaped curves), which may not be the case.
- Second, unordered categorical predictor variables (e.g., red/blue/ green) cannot be used at all. Third, the boundaries that separate the classes are all linear forms (such as lines or planes), but sometimes the data just can’t be separated that way.

Generalized Additive Models (GAM)

- There is a class of models extending both linear and logistic regression, known as generalized additive models or GAM.
- They are called additive because we assume that the model can be written as the sum of possibly non-linear functions, one for each predictor.
- GAM can be used either for regression or for classification of a binary response. The response variable can be virtually any function of the predictors as long as there are not discontinuous steps

backfitting

- The most common estimation procedure is backfitting. Instead of estimating large numbers of parameters as neural nets do, GAM goes one step further and estimates a value of the output for each value of the input — one point, one estimate.
- As with the neural net, GAM generates a curve automatically, choosing the amount of complexity based on the data.

Hidden Markov Model

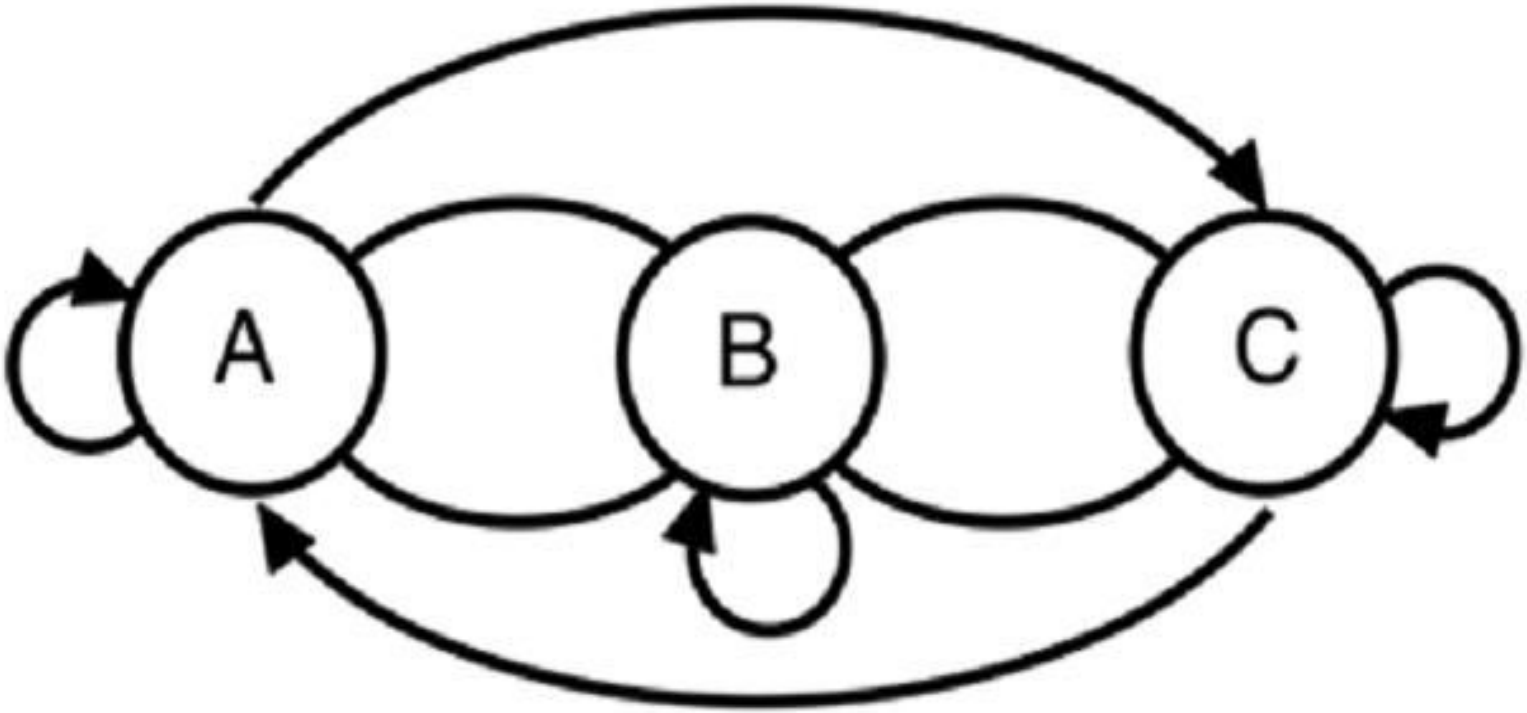
A powerful statistical approach to constructing classifiers that deserves a separate discussion is the use of Hidden Markov Modeling.

A Hidden Markov Model (HMM) is a statistical model for an ordered sequence of symbols, acting as a stochastic state machine that generates a symbol each time a transition is made from one state to the next.

Transitions between states are specified by transition probabilities.

Hidden Markov Model

Figure 7-13. Markov Chain. (A), (B), and (C) represent states, and the arrows connecting the states represent transitions.



A Markov process

- A Markov process is a process that moves from state to state depending on the previous n states.
- The process is called an order n model where n is the number of states affecting the choice of the next state.
- The Markov process considered here is a first order, in that the probability of a state is dependent only on the directly preceding state.

- When presented with data in the database, the HMM provides a measure of how close the data patterns—sequence data, for example—resemble the data used to train the model.
- HMM-based classifiers are considered approximations because of the often unrealistic assumptions that a state is dependent only on predecessors and that this dependence is time-independent.

Question?

What is Logistic regression ?

References

- Introduction to Data Mining and Knowledge Discovery Third Edition by Two Crows Corporation