Date : / /
Page :

Module 3
CLASSIFICATION & PREDICTION

* There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends.
* These are the two forms: classification prediction
* classification models predict categorical class labels, and prediction models predict continuous valued function

• eg: we can build a classification model to categorize bank loan applications as either safe/risky. or a prediction model to predict the expenditures in dollers

classification vs prediction

classification

* predicts categorical class labels[discrete or nominal]
• 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.

prediction

models continuous -valued functions, ie predicts unknown or missing value.

Typical Application

• credit Approval

- Target marketing
- medical diagonosis
- Fraud detection

what is classification

Following are the examples of cases where the Data analysis task is classification :-

- A bank loan officer wants to analyze the data in order to know which customer (loan applicants) are risky or which are safe
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new Computer

In both of the above examples, a model or classifier is Constructed to predict the categorical labels. These labels are risky or safe for loan application Data and yes or no form marketing data.

→ How does classification works ?

The data classification process Includes 2 types

i] Building the classifier or model

ii] using classifier For classification.

i, Building the classifier or model

The step is the learning step or the learning phase

→ In this step the classification algorithms build the classifier.

→ The classifier is built from the training set ~~model~~ made up of database tuples and their -

associated class labels.
-> Each tuple that constitutes the training set is reffered to as a category or class. These tuples can also be referred to as sample, object or data points

Supervised learning Vs unsupervised learning

supervised learning [classification]

- ~~Supervised~~ supervision:- The training data [observations, measurements etc] are accompanied by labels indicating the class of the ~~obser~~ observations.

- New data is classified based on the training set

Unsupervised learning [clustering]

- These class labels of training data [observations is unknown
a set of ~~measurement~~ measurement, observations ects with the aim of establishing the existance of classes or clusters in the data.

Issues :- Data preparation.

-> Preparing the data for classification & prediction
Data cleaning
- preprocess the data in order to reduce noise & handle missing values
- Relevance analysis [Feature selection]

Date: / /
Page:

- ~~Remove analysis~~
    - Remove the irrelevant or redundant attrib
- → Data transformation & reduction
    - Generalize and /or normalize data.

Issues: Evaluating classification methods.

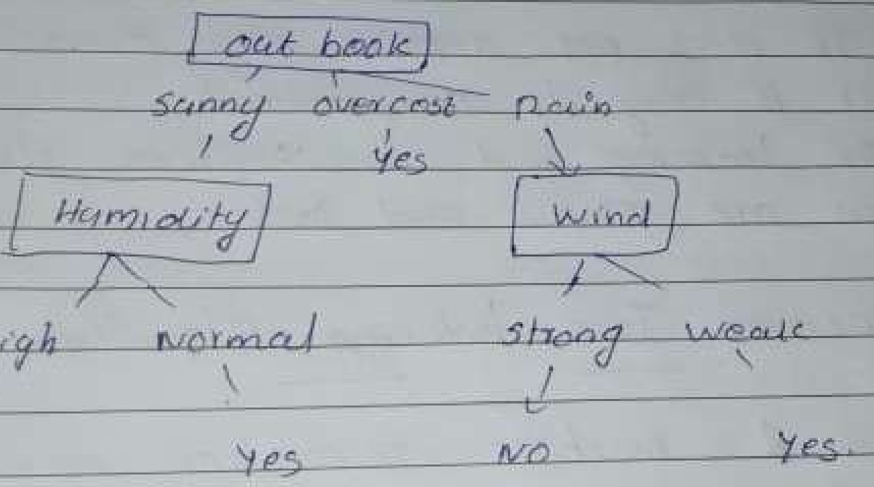
- Accuracy:
    - classifier accuracy: predicting class label
    - Predictor accuracy: guessing value of predicted attributes.
- speed:
    - time to construct the model [training time]
    - time to use the model [classification /prediction time]
- Robustenss : handling noise & missing values
- scalability : efficiency in disk ~~resi~~ resident data bases
- ~~Interpre~~ Interpretability
    - Understanding & insight provided by The model
- other measure

  eg: goodness of rules. such as decision tree size or compactness of classification rules.

Decision Trees Induction.

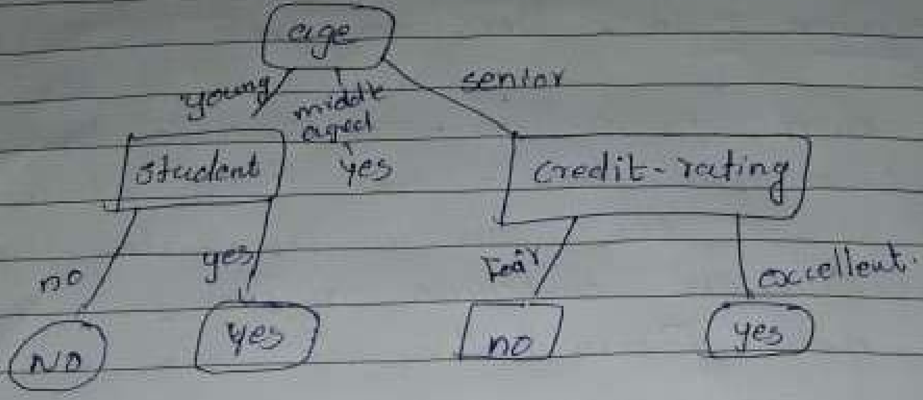
Decision tree:- Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents the outcome of the test and each leaf node [terminal node] holds a class labels.

Decision Tree for [play tennis]



A decision tree is a structure that Includes root node branches, and leaf nodes. Each internal node denotes a test on attribute, each branch denotes the outcome of a test, and each leaf node holds a class labels. The top most node in the tree is the root node.

The following decision tree is for the concept buy computer that indicates whether a customer or not Each internal node represents a test on an attribute. Each leaf node represents a class



The benefits of having a decision tree are as follows:-

- It does not require any domain knowledge
- It is easy to comprehened
- The learning and classification steps of a decision tree are simple and fast

Decision Tree Inducing Algorithm.

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 [Iterative Dichotomiser]. Later he presented C4.5 adopt a greedy ~~approach~~ approach. In this algorithm, there is no back tracking, the trees are constructed in a top-down recursive divide-and-conquer manner

Generating a decision tree from training tuples of data partition D

Algorithm: General - decision tree

Input:

Data partition, D, which is a set of training tuples and their associated class labels.
Attribute - list, the set of candidate attributes.
Attribute selection method, a procedure to determine the splitting criterion that best partitions that the data tuples into individual classes. This criterion includes a splitting - attribute and either a splitting point or splitting subset

output:

A decision tree

Method

Create a node N;

IF tuples in D are all of the same class, C then return N as leaf node labeled with class. C,

IF attribute list is empty then return N as leaf with labeled with majority class in D, // ~~major~~ majority voting.

apply attribute selection - method [D, attribute - list] to find the best splitting criterion,
label node N with splitting - criterion.

If splitting - attribute is discrete valued and multiway spliting allowed then// no restricted to binary trees.

Attribute list - spliting attributes. // remove splitting

attribute for each outcome j of splitting criterion

// partition the types and grow subtrees for each partition let Dj be the set of data tuples in D satisfying outcome j. // a partition

if Dj is empty then
attach the node a leaf labeled with the majority class in D to node N;
else
attach the node returned by Generate ~~dessi~~ Decision tree [Dj. attribute list] to Node Nj
end for
return N;

Tree ~~Pran~~ Pruning

Tree pruning is performed in ~~odder~~ order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches:-

- pre - pruning - The tree is pruned by halting its

Date : / /
Page :

Constrcuction early

- Post - pruning - This approach removes a sub-Tree from a fully grown tree.

Cost complexity

Cost Complexity is measured by the following two parameters:-

i) Number of leaves in the tree
ii) Error rate of the tree

strengths and Weakness of ~~Bee~~ Decision Tree approach

strength:-

- Decision trees are able to generate understandable ~~set~~ rules
- Decision trees perform classification without requiring much computation
- Decision trees are able to handle both continuous and Categorical variabels.
- Decision trees provide a clear indication of which fields are important for prediction or classification

Weakness:-

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a

Page :

continous attribute

- Decision trees are prone to errors in classification problems with many class & relatively small number of training examples

- Decision tree can be computationally expensive to train. The process of growing a decision tree is computationaly expensive. Pruning algorithm can be also expensive since many Candidata Sub-trees must be formed and compared

→ classifier accuracy:-
The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of a new or previously unseen data

→

The accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new previous unseen data

Attributs selection measure

- Attribute selection measure is a heuristic for selecting the splitting criterion that "best" seperates a given data partition, D, of a class-labeled training tuples into individual classes. The information gain is used to select the splitting attribute in each node in the tree.

Date : / /
Page :

- select the attribute with the highest information gain
  - Expected Information
  - Information needed
  - Information gained

computing Information Gain for Continuous value attributes.

- Let attribute A be a continuous-valued attribute
- must determine the best split point for A
- Sort the value A in increasing order
- Typically, the midpoint between each pair of adjacent value is considered as a possible split point.
  $[a_i + a_{i+1}]/2$ is the midpoint b/w values of a_i & a_{i+1}
- The point with the minimum expected information - requirement for A is selected as the split-point for A.

overfitting and Tree pruning

overfitting : An induced tree may overfit the Training data.
- Too many branches, some may reflect anomalies due to noise or outliers.
- poor accuracy for unseen samples.

→ Two Approaches for avoid overfitting :-

- Prepruning : Halt tree construction early - do not split a node if this would result

Date : / /
Page :

In the goodness measure falling below a threshold
- Difficult to choose appropriate threshold

- Postpruning :- Remove branches from a "fully grown" tree - get a sequence of progressively pruned trees.

- use a set of data different from training data to decide which is the "best pruned-tree"

Scalable Decision Tree Induction methods.

SLIQ

Builds an index for each attribute and only class list and the current attribute list reside in memory.

SPRINT

Constructs an attribute list data structure

PUBLIC

Integrates tree splitting and tree pruning, stop growing the tree earlier

- Rain Forest

Builds an AVC - list (Attribute, value, class label)

- BOAT

uses bootstrapping to create several small samples.

Date : / /
Page :

Bayesian classification

Bayesial classification is Based on Bayes Theorom Bayesian classifiers are the statistical classifiers Bayesian classfiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorom

Baye's, Theorom is named after Thomas Bayes. There are two types of probabilities-

- Posterior probability [P(H/X)]
- Prior Probability [P(H)]

Where X is data tuple and H is some hypothesis According to Baye's Theorom,

$$P(H/X) = P(X/H)\,P(H)\,/\,P(X)$$

Bayesian classification uses Bayes Theorom to predict the occurrence of any event. Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings. The theory expresses how a level of belief, expressed as a probability.

Bayes theorem come into existence after Thomas Bayes who first utilized conditional probability to provide an algorithm that uses evidence to calculate limits on an unknown parameter.

Baye's theorem is expressed mathematically by the following equation i.e given below

$$P(x|y) = \frac{P(y/x)\ P(x)}{P(y)}$$

where x and y are the events and P(y) ≠ 0

P(x/y) is a conditional probability that describes the occurence of event x is given that y ~~is~~ is true

P(y/x) is a conditional probability that describes the occurence of event y is given that x is true

P(x) and P(y) are the probabilities of observing x and y independently of each other. This is known as the marginal probability.

Bayesian interpretation

In the bayesain interpretation, probability determines a "degree of belief." Bayes Theorem connects the degree of belief in a hypothesis before and after accounting for evidence. eg: let us consider an example of the coin, if we toss a coin, then we get either heads or tails and the percent of occurence of either heads and tail is 50%. If the coin is flipped numbers of times and the outcomes are observed, the degree of belief may rise, fall or ~~...~~ remain the same depending on the outcomes.

For proposition X and evidence Y,

- P(x), the prior, is the primary degree of belief in x.

- P(x/y), the posterior is the degree of belief having accounted for Y

- The quotient $\frac{P(Y/x)}{P(Y)}$ represents the supports Y provides for X.

Bayes theorem can be derived from the conditional probability:-

$$P(X/Y) = \frac{P(X \wedge Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

where $P(X \cap Y)$ is the joint probability of both X and Y being true because

$$P(Y \cap X) = P(X \cap Y)$$

$$\text{or } P(X \cap Y) = P(X/Y)\,P(Y) = P(Y/X)\,P(X)$$

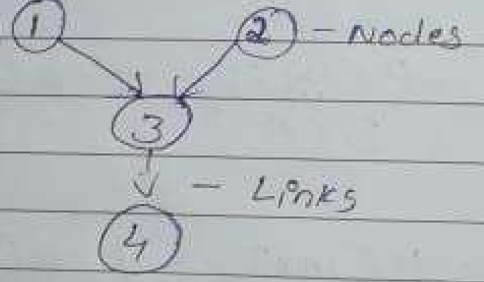
Date : / /
Page :

or $P(X/Y) = \frac{P(Y/X)\ P(X)}{P(Y)} = P(Y) \neq 0$

Bayesian n/w :-

A Bayesian n/w falls under the classification of probabilistic Graphical modelling [PGM] procedure is utilized to compute uncertainties by utilizing the probability concept. Generally known as Belif n/ws, Bayesian n/ws are used to show uncertainities using directed A cyclic Graph (DAG)

A Directed A cyclic Graph is used to show a Bayesian n/w, and like some other statistical graph. DAG consists of a set of nodes and links, where the links signify the connection between the nodes.



The nodes here represent random variables and the edges define the relationship b/w these variables.

A DAG models the uncertainty of an event Taking place based on the Conditional

Date: / /
Page:

probability Distribution [CDP] of each random variable. A conditional probability table [CPT] is used to represent the CPD of each variable in a n/w.

Bayesian classification: way ?

A statistical Classifier: performs probablistic Prediction ie predicts class membership probabilities.

Foundation :- Based on Bayes Theorem

Performance : A simple Bayesian classifier, native Bayesian classifier, has comparable Performance with decision tree and selected neural network classifiers.

Incremental : Each training example can incrementally increase / decrease the Probability that a hypothesis is correct - prior knowledge can be combined with observed data

standard : Even when Bayesian methods are computationally interactable, they can provide a standard of optimal decision making against which other methods can be measured.