

Skill Based Subject – III: DATA MINING AND WAREHOUSING - 18BIT55S

UNIT II: Classification: Introduction – Decision Tree – Over fitting and Pruning – Decision Tree Rules – Naïve Bayes Method – Estimating Predictive Accuracy of Classification Methods – Improving Accuracy of Classification Methods – Other Evaluation Criteria for Classification Methods – Classification Software.

TEXT BOOK


G.K Gupta, “Introduction to Data Mining with Case Studies”, Prentice Hall of India(Pvt) Ltd, India, 2008.

Prepared by : Mrs. G. Shashikala, Assistant Professor, PG Department of Information Technology

Classification

- Classification is the separation or ordering of objects(or things) into classes. If the classes are created without looking at the data (non-empirically), the classification is called **apriori classification**.
- If the classes are created empirically, by looking at the data, the classification is called posteriori classification.
- The classification consists of training the system, so that when a new object is presented to the trained system, it is able to assign the object to one of the existing classes. This approach is also called supervised learning.

- Some techniques are available for posteriori or unsupervised classification, in which classes are determined based on the given data. The algorithm used for this is clustering
- Each object has a number of attributes, and one of them tells us , which class the object belongs to. This attribute is known for the training data and it has to be determined for the test data, by the classification method.
- This attribute is the output of all attributes and is referred to as the output attribute or the dependent attribute. Other attributes are called the input attributes or the independent attributes.

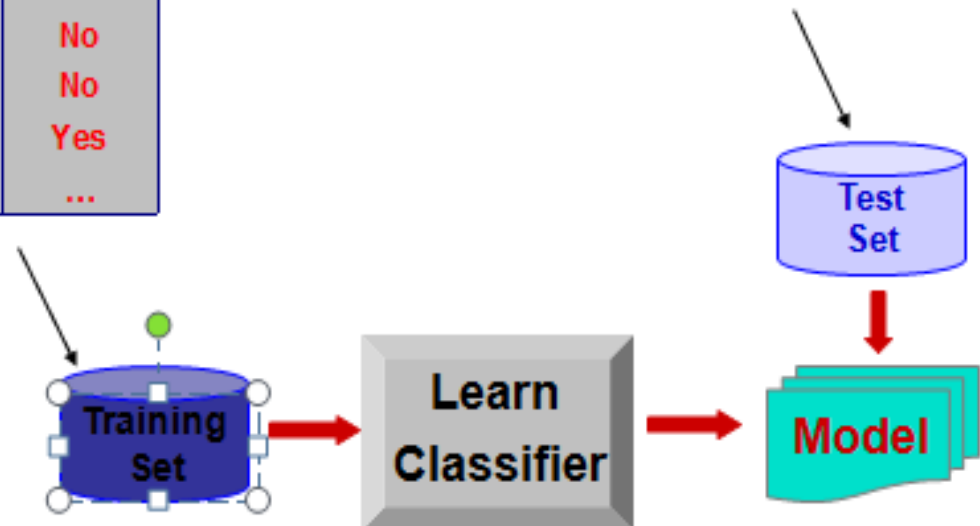
- 
- The attributes may be of different types
 - Attributes whose domain are numerical are called numerical attributes
 - Non-numerical attributes are called categorical attributes. Categorical attributes may be ordered or unordered.

General Approach for Building a Classification Model

categorical *categorical* *quantitative* *class*

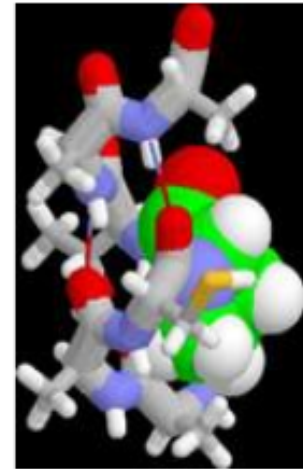
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc



Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Decision Tree

- This is a classification method that results in a flow-chart like tree structure, where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the classes.
- Ex: to classify Australian animals

Table 3.1 Training data for a classification problem

<i>Name</i>	<i>Eggs</i>	<i>Pouch</i>	<i>Flies</i>	<i>Feathers</i>	<i>Class</i>
Cockatoo	Yes	No	Yes	Yes	Bird
Dugong	No	No	No	No	Mammal
Echidna	Yes	Yes	No	No	Marsupial
Emu	Yes	No	No	Yes	Bird
Kangaroo	No	Yes	No	No	Marsupial
Koala	No	Yes	No	No	Marsupial
Kookaburra	Yes	No	Yes	Yes	Bird
Owl	Yes	No	Yes	Yes	Bird
Penguin	Yes	No	No	Yes	Bird
Platypus	Yes	No	No	No	Mammal
Possum	No	Yes	No	No	Marsupial
Wombat	No	Yes	No	No	Marsupial

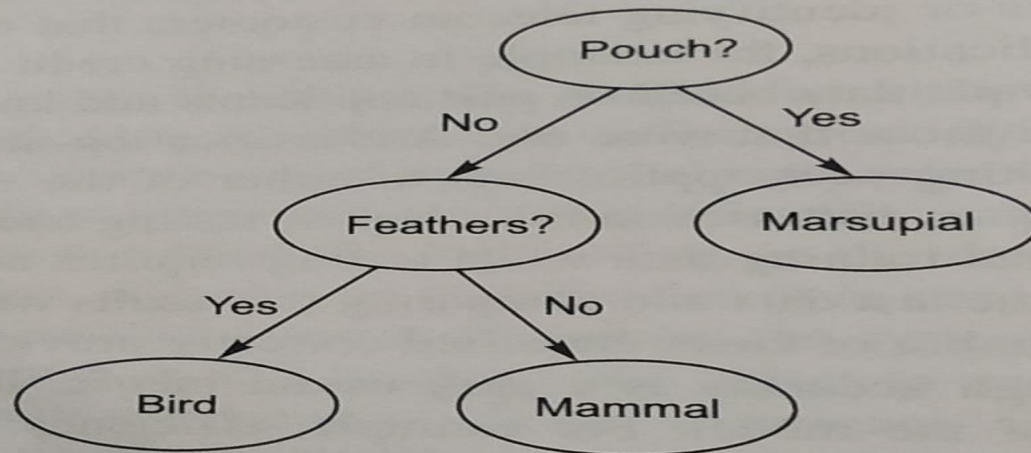


Figure 3.1 A decision tree for the data in Table 3.1.

Decision tree is a model that is both predictive and descriptive

Displays relationships found in the training data

The tree contains zero or more internal nodes and one or more leaf nodes with each internal node being a decision node having two or more child nodes.

Using the training data, the decision tree method generates a tree that consists of nodes that are rules, to determine the class of an object after the training is completed. (similar to 20 questions method)

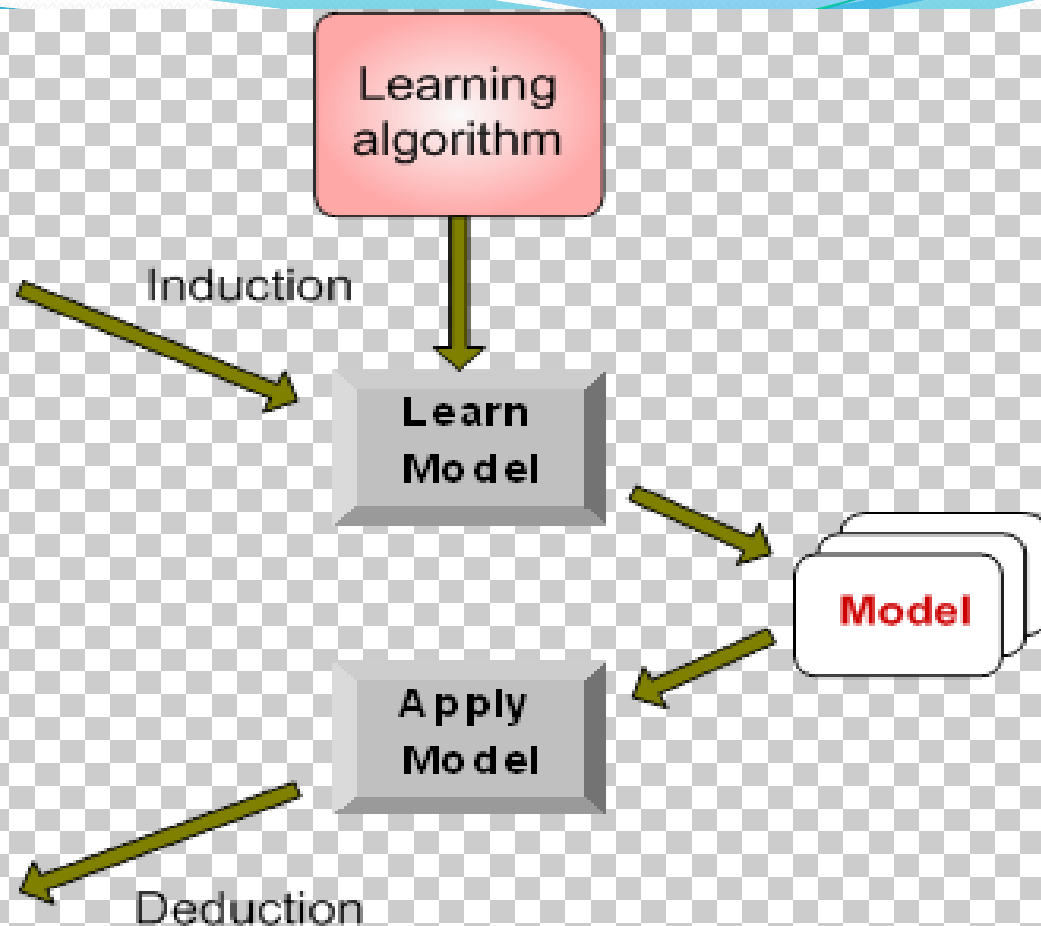
- Each node represents a choice between a number of alternatives and each leaf node represents a classification or decision.
- The training process that generates the tree is called induction.
- The no. of training samples required is small if the no. of independent attributes is small and the no. of training samples is large when the no. of attributes is large.
- Normally, the complexity of a decision tree increases as the no. of attributes increases
- The quality of training data plays an important role in determining the quality of the decision tree
- If there a no. of classes, then there should be sufficient training data available that belongs to each of the classes.
- Measuring the quality of a decision tree is by
 - Classification accuracy determined using test data
 - Average cost and worst case cost of classifying an object

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



- The decision tree algorithm is a
 - Simple top-down greedy algorithm
 - Builds a tree that has leaves that are as homogeneous as possible

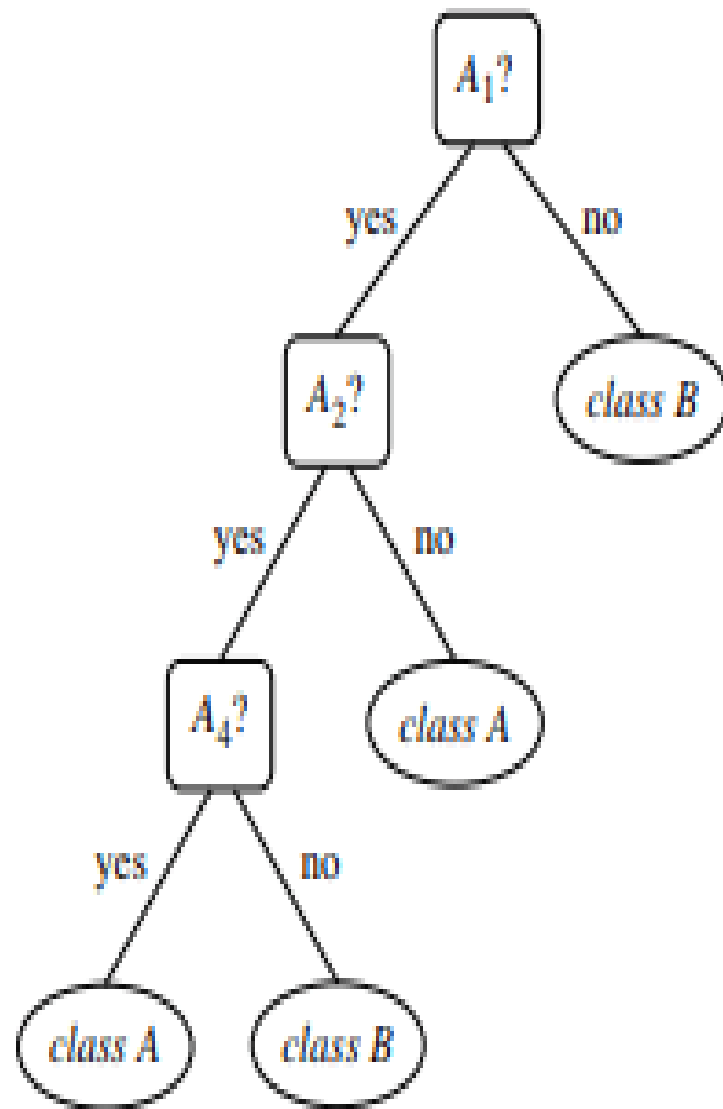
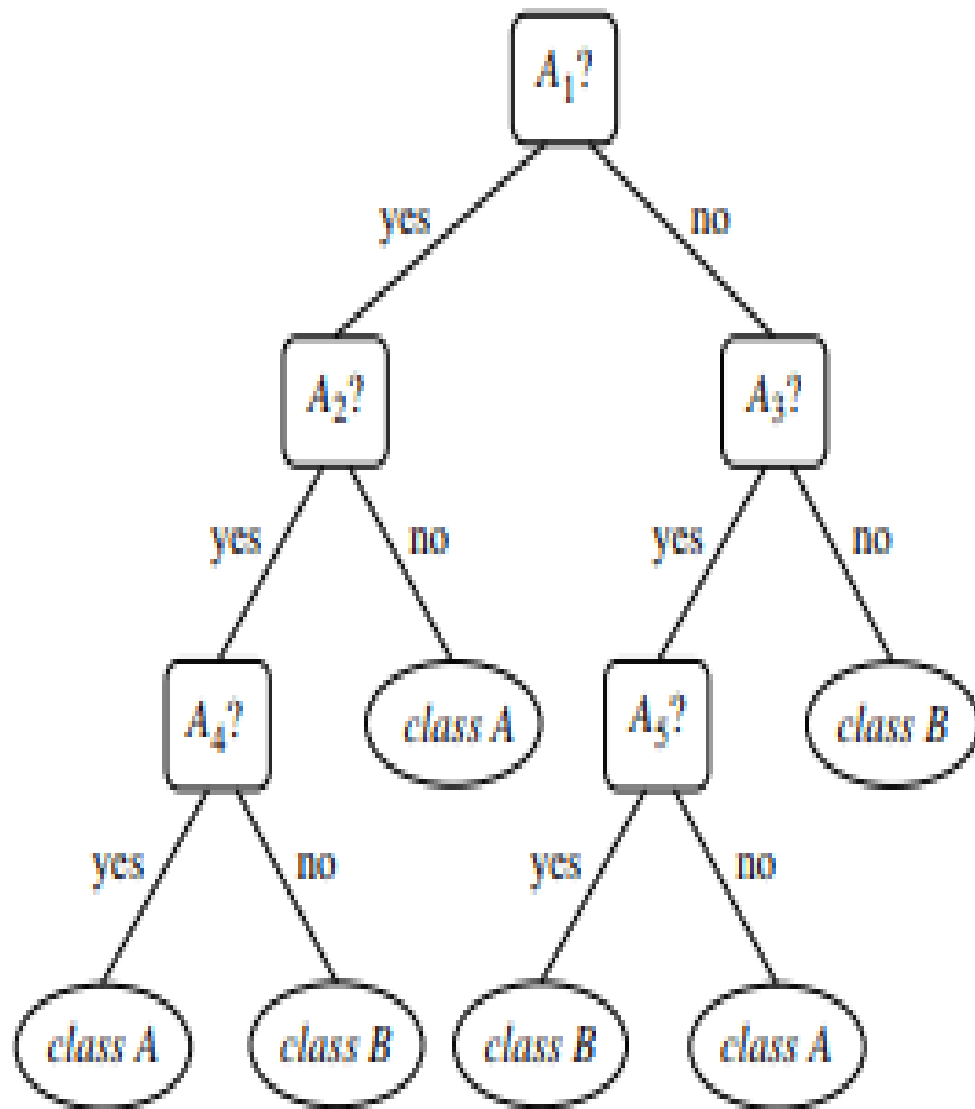
Building a decision tree – the tree induction algorithm

- Let the set of training data be S . If some of the attributes are continuous-valued, they should be discretized
- If all instances in S are in the same class, then stop
- Split the next node by selecting an attribute A from amongst the independent attributes that best divides or splits the in the node into subsets and create a decision tree node
- Split the node according to the values of A

- Stop, if either of the following conditions is met, otherwise continue with step 3
 - If this partition divides the data into subsets that belong to a single class and no other node needs splitting
 - If there are no remaining attributes on which the sample may be further divided

Overfitting and Pruning

- Overfitting results in decision trees that are more complex than necessary. Training error no longer provides a good estimate of how well the tree will perform on previously unseen records
- A tree that fits the training data too well may not be a good classifier for new examples. Overfitting results in decision trees more complex than necessary.
- Typically, $2/3$ of the data set is reserved to model building and $1/3$ for error estimation.
- Disadvantage: less data is available for training- Overfitted trees may have a low re-substitution error but a high generalization error.



Decision tree rules

- Each node of the tree specifies a test of some attribute and each branch from the node corresponds to one of the values of the attribute
- each path from the root to a leaf of the decision tree consists of attribute tests, finally reaching a leaf that describes the class
- Advantages in converting a decision tree to rules are :
 - Decision rules make it easier to make pruning decisions since it is easier to see the context of each rule
 - Removes distinction between attribute tests that occur near the root of the tree and those that occur near the leaves

- IF-THEN rules may be derived based on the various paths from the root to the leaf nodes.
- Rules may be combined to produce a smaller set of rules.
- For ex.,

If Gender = “Male” then Class = B

If Gender = “Female” and Married = “Yes” then Class = C,
else Class = A

- Rules with only one antecedent (If Gender = “Male” then Class = B) cannot be further simplified
- A number of rules that lead to the same Class may be combined

Naïve Bayes Method

- This method is based on the work of Thomas Bayes
- In Bayesian classification, we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true
- Some notations
- The expression $P(A)$ refers to \rightarrow the probability that event A will occur
- $P(A|B)$ refers to the probability that event A will happen given that event B has already happened
- $P(A|B)$ is the conditional probability of A based on the condition that B has already happened.
- Bayes' theorem :

$$P(A|B) = p(B|A)*P(A)/P(B)$$

• It is known that

$$P(A|B) = P(A \& B) / P(B)$$

and $P(B|A) = P(A \& B) / P(A)$

Dividing the first equation by the second gives us the Bayes' theorem

$P(A \& B)$ refers to the prob. Of having both A and B

- Let X be an object to be classified
- We have to find the prob. Of X belonging to one of the classes C_1 , C_2 , C_3 , etc by calculating $P(C_i|X)$
- Then we assign X to the class that has the highest conditional probability.
- We have

$$P(C_i|X) = [P(X|C_i) P(C_i)] / P(X)$$

- $P(C_i|X)$ is the prob. Of the object X belonging to class C_i
- $P(X|C_i)$ is the prob. Of obtaining attribute values X if we know that it belongs to class C_i
- $P(C_i)$ is the prob. Of any object X belonging to class C_i , without any other information
- $P(X)$ is the prob. Of obtaining attribute values X whatever class the object belongs to

- $P(X)$ is independent of C_i
- So we have to compute $P(X|C_i)$ and $P(C_i)$ for each class
- To compute $P(X|C_i)$ we use a naïve approach, assuming that all attributes are independent
- Now the prob. Of each of the attribute values is identified by counting the frequency of those values for class C_i
- Then we determine the class allocation of X by computing $[P(X|C_i) P(C_i)]$ for each of the classes and allocating X to the class with the largest value
- In Bayesian approach, the prob. Of the dependent attribute can be estimated by computing the probabilities of independent attributes

Example 3.3—Naïve Bayes Method

Once again we go back to the example in Table 3.2 that we have used before.

<i>Owns home?</i>	<i>Married</i>	<i>Gender</i>	<i>Employed</i>	<i>Credit rating</i>	<i>Risk class</i>
Yes	Yes	Male	Yes	A	B
No	No	Female	Yes	A	A
Yes	Yes	Female	Yes	B	C
Yes	No	Male	No	B	B
No	Yes	Female	Yes	B	C
No	No	Female	Yes	B	A
No	No	Male	No	B	B
Yes	No	Female	Yes	A	A
No	Yes	Female	Yes	A	C
Yes	Yes	Female	Yes	A	C

There are 10 ($s = 10$) samples and three classes.

Credit risk Class A = 3

Credit risk Class B = 3

Credit risk Class C = 4

... these frequencies by the total number in the training

- The prior prob. Are obtained by dividing these frequencies by the total no. in the training data
- $P(A) = 0.3$
- $P(B) = 0.3$
- $P(C) = 0.4$

Table 3.10 Probability of events in the Naïve Bayes method

<i>Owens home</i>	<i>Married</i>	<i>Gender</i>	<i>Employed</i>	<i>Credit rating</i>	<i>Class</i>
No	No	Female	Yes	A	A
No	No	Female	Yes	B	A
Yes	No	Female	Yes	A	A
1/3	1	1	1	2/3	Probability of having {yes, no, female, yes, A} attribute values given the risk Class A
Yes	Yes	Male	Yes	A	B
Yes	No	Male	No	B	B
No	No	Male	No	B	B
2/3	2/3	0	1/3	1/3	Probability of having {yes, no, female, yes, A} attribute values given the risk Class B
Yes	Yes	Female	Yes	B	C
No	Yes	Female	Yes	B	C
No	Yes	Female	Yes	A	C
Yes	Yes	Female	Yes	A	C
0.5	0	1	1.0	0.5	Probability of having {yes, no, female, yes, A} attribute values given the risk Class C

Given the estimates of the probabilities in Table 3.10, we can compute the posterior probabilities as

$$P(X|A) = 2/9$$

$$P(X|B) = 0$$

$$P(X|C) = 0$$

Therefore the values of $P(X|C_i)P(C_i)$ are zero for Classes B and C. If X is assigned to Class A, it is

- If the data presented to us is { yes, no, female, yes, A } for the five attributes, we can compute the posterior prob. For each class
- $P(X|C_i) = P(\{\text{yes, no, female, yes, A}\} | C_i) = P(\text{Owns home} = \text{yes} | C_i) \times P(\text{Married} = \text{no} | C_i) \times P(\text{Gender} = \text{female} | C_i) \times P(\text{Employed} = \text{yes} | C_i) \times P(\text{Credit rating} = A | C_i)$
- Using the above expression, we compute the three posterior prob., for the three classes, that the person with attribute values X has credit risk class A, or class B or class C

Estimating predictive accuracy of classification methods

- The accuracy of a classification method is the ability of the method to correctly determine the class of a randomly selected data instance
- It is estimated as the probability of correctly classifying unseen data
- Different sets of training data may lead to different models and so testing is very important in determining how accurate each model is .
- Then the most accurate model is selected
- The metrics used for measuring the efficiency of a model are sensitivity, specificity, precision and accuracy

- The methods for estimating errors include holdout, random sub-sampling, k-fold cross-validation, and leave-one-out
- Let us assume that the test data has T objects. When testing a method, we find that C of the T objects are correctly classified. The error rate then may be defined as

$$\text{Error rate} = (T-C)/T$$
- A “confusion matrix” is used to represent the result of testing in more detail as shown below:

Table 3.11 A confusion matrix for three classes

<i>Predicted class</i>	<i>True class</i>		
	1	2	3
1	8	1	1
2	2	9	2
3	0	0	7

- The confusion matrix tells us how many objects got misclassified and also what misclassifications occurred
- In the above ex., we can see that 1 out of 10 objects that belonged to Class 3 got misclassified to Class 1
- There are two terms “false positive” (FP) and “false negative” (FN)
- False positive cases are those that did not belong to a class but were allocated to it .

In the above ex., there were two false positives for class 1 and 4 for class 2

- False negative are cases that belong to a class but were not allocated to it.

In the above ex., there were 2 false negatives for class 1, one for class 2, and three for class 3

Therefore, there were a total of 6 false positives and 6 false negatives for the above ex.

- Sensitivity = $TP / (TP + FN)$
- Specificity = $TN / (TN + FP)$

Where TP (total positives) is the total correctly classified objects and TN (total negatives) is the total number of objects that did not get classified to a class they did not belong to.

- Consider Class 1 in the above table. There are 10 objects that belong to this class and 20 do not. Of this 10, only 8 are classified correctly. In total, 24 objects are classified correctly.

Out of the 20 that did not belong to class 1, 2 objects are classified wrongly to belong to it.

so, we have, for Class 1, $TP = 8$, $TN = 18$, $FN = 2$, $FP = 2$.

For Class 2, $TP = 9$, $TN = 16$, $FN = 1$, $FP = 4$


For Class 3, $TP = 7$, $TN = 20$, $FN = 3$, $FP = 0$

• Now, we can compute :

- Sensitivity = $TP / (TP + FN) = 24 / 30 = 80\%$

- Specificity = $TN / (TN + FP) = 54 / 60 = 90\%$

- The methods used for estimating the accuracy of a classification method are :
- 1. Holdout Method
 - This is also called as test sample method
 - It requires a training set and a test set and they are mutually exclusive
 - The dataset available is divided into two subsets ($2/3$) and ($1/3$)
 - Once the classification method produces the model using the training set, the test set can be used to estimate the accuracy.

- 
- Random sub-sampling method
 - This method is like holdout method except that it does not rely on a single test set.
 - The holdout estimation is repeated several times and the accuracy estimate is obtained by computing the mean of the several trials.
 - This method produces better error estimates

- K-fold cross validation method
 - Here the available data is randomly divided into k disjoint subsets of approximately equal size.
 - One of the subsets is then used as the test set and the remaining $k-1$ sets are used for building the classifier
 - The test set is then used to estimate the accuracy.
 - This is done repeatedly k times so that each subset is used as a test subset once
 - The accuracy estimate is then the mean of the estimates for each of the classifiers
 - $K=10$ has been found to be adequate and accurate

- Leave-one-out method

- In this method, one of the training samples is taken out and the model is generated using the remaining training data
- Once the model is built, the one remaining sample is used for testing and the result is coded as 1 or 0 depending of it was classified correctly or not.
- The average of such results provides an estimate of the accuracy
- This method is useful when the dataset is small

Bootstrap method

- In this method, given a data set of size n , a bootstrap sample is randomly selected uniformly with replacement (i.e., a sample may be selected more than once) by sampling n times and used to build a model
- It can be shown that only 63.2% of these samples are unique
- This is repeated and the average of error estimates is obtained
- The bootstrap method is unbiased and many iterations are needed for good error estimate

Improving accuracy of classification methods

- Techniques used for improving the accuracy of classification results are
 - Bootstrapping
 - Bagging
 - Boosting
- All three involve combining several classification results from the same training data
- Bootstrapping method
 - In this method, on the average, only 63.2% of the objects are unique and so all the samples are different from each other
 - The bootstrap samples are then used for building decision trees which are then combined to form a single decision tree




- Bagging

- Name is derived from Bootstrap and aggregating
- This method combines classification results from multiple models or results of using the same method on several different sets of training data
- Used to improve the stability and accuracy of a complex classification model with limited training data by using sub-samples obtained by resampling, with replacement for generating models
- Involves simple voting (with no weights)



- Boosting

- This involves applying weights to different instances of training data depending on the quality of the previous result
- More weight is given to those instances that are not correctly classified and less weight to those that are correctly classified
- One approach is that different weights can be applied to data in different classes
- For ex., the weights could be inversely proportional to the accuracy of prediction in each class

- 
- Thus classes predicted with less accuracy will get a higher weight and data from these classes will be selected more frequently in re-sampling
 - Benefits of bootstrapping, bagging and boosting are
 - These techniques can provide a level of accuracy that cannot be obtained by a large single-tree method
 - Creating a single decision tree from a collection of trees in bagging and boosting is not difficult
 - These methods can avoid the problem of overfitting

Other evaluation criteria for classification methods

- Speed
- Robustness
- Scalability
- Interpretability
- Goodness of the model
- Flexibility
- Time complexity

Classification software

- C4.5 version 8 of the basic decision tree tool, developed by J.R.Quinlan
- C 5.0/See5 from Rulequest research
- CART 5.0 and TreeNet from Salford systems
- DTREG
- Model Builder for decision trees from Fair Isaac
- OC₁, Oblique Classifier 1
- Quadstone system Version 5
- Shih Data Miner
- SMILES
- NBC: a Simple Naive Bayes Classifier