22 Week 25
June
Wednesday (174-192)

MODULE 4

Week	June				
Monday		6	13	20	27
Tuesday		7	14	21	28
Wednesday	1	8	15	22	29
Thursday	2	9	16	23	30
Friday	3	10	17	24	
Saturday	4	11	18	25	
Sunday	5	12	19	26	

8.00 # Classification –

8.30 It predicts the categorical class labels (discrete or nominal) and classifies data based on the training set & the values in a classifying attribute and uses it in classifying new data.

9.00

9.30

10.00

10.30

11.00 → Two step process:

11.30

12.00 1. Model Construction : of training set.

12.30 Each tuple / sample is assumed to belong to a predefined class, as determined by class label attribute. The model is represented as classification rules, decision trees, or formulas.

13.00

13.30

14.00

14.30

15.00 2. Model Usage :

15.30 the known label of test sample is compared with the classified result from the model. Accuracy rate

16.00 is the % of test set samples that are correctly classified by the model. Test set is independent of training set.

16.30 If accuracy is acceptable, we use that model to classify data.

17.00

17.30

18.00

Evening

Prediction – it is to identify data points purely on the description of another related data value. It is not necessarily related to future events. It derives a relationship b/w something you know & a thing that you need to predict for future reference.

Meetings | Things To Do | Important Calls

Issues -

1. ~~Data Cleaning~~

Classification & Prediction Issues -

The major issue is preparing the data for classification & Prediction, which involves.

i) Data cleaning: it involves removing of noise & treatment of missing values. Removal of noise is done by applying smoothing techniques and prob. of missing values is solved by replacing a missing value with commonly occuring value.

ii) Relevance Analysis: Database also has irrelevant attributes. Correlation analysis is used to know whether any 2 attributes are related.

iii) Data Transformation ~~& Reduction~~ - Data can be transformed using any of these methods:

- Normalization: scaling of all values for given attribute ~~is~~ in order to make them fall within a specified range
- Generalization: we can use the concept of hierarchies for transformation of data by generalizing it to the higher concept.

iv) Data Reduction - Data can be reduced using concepts of wavelet transformation, binning, histogram analysis & clustering.

Evening

Meetings

Things To Do
- ☐
- ☐
- ☐
- ☐

Important Calls
- ☐
- ☐
- ☐
- ☐

$\frac{1}{4}$ $\frac{1}{4}$ + $\frac{3}{4}$ $\frac{3}{4}$

24 Week 25
June
Friday (176-190)

Week	June				
Monday	22	23	24	25	26
Tuesday		6	13	20	27
Wednesday		7	14	21	28
Thursday	1	8	15	22	29
Friday	2	9	16	23	30
Saturday	3	10	17	24	
Sunday	4	11	18	25	
	5	12	19	26	

numerical

8.00 # Decision Tree

8.30

9.00 → Algorithm:

9.30

10.00 1. Basic Algo (Greedy Approach) –

10.30 ◦ Tree is ~~compared~~ constructed in a top-down recursive divide

11.00 and conquer manner.

11.30 ◦ ~~At~~ start all the training samples are at root.

12.00 ◦ Attributes are categorical

12.30 ◦ Examples/samples are partitioned recursively based on selected

13.00 attributes

13.30 ◦ Test attributes are selected on the basis of heuristic or

14.00 statistical manner.

14.30

15.00 2. Terminating Condition –

15.30 ◦ There are no samples left

16.00 ◦ No remaining attributes for partitioning

16.30 ◦ All samples for a given node belongs to same class.

17.00

17.30 → Entropy (Information Theory) –

18.00 A measure of uncertainity associated with a random variable.

Evening

◦ Calculation – for a discrete random variable Y taking m distinct values $\{y_1, \ldots y_m\}$

$$\text{Info}(D) = H(Y) = \sum_{i=1}^{m} p_i \log(p_i), \text{ where } p_i = P(Y = y_i)$$

Meetings | Things To Do | Important Calls

→ Information needed to classify D (using A to split D) -

$$Info_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{D} Info(D_j)$$

→ Info gained - $Gain(A) = Info(D) - Info_A(D)$

→ Gini Index - If a data set D contains examples from n classes, gini index, $gini(D)$ is defined as.

$$gini(D) = 1 - \sum_{i=1}^{n} P_i^2$$

where p_i is relative frequency of class i in D.

→ If a data set D is split on A into two subsets D_1 and D_2, then $gini(D)$ is defined as

SUNDAY 26

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

→ reduction in impurity -

$$\Delta gini(A) = gini(D) - gini_A(D)$$

the attribute that provides largest $\Delta gini(A)$ (reduction in impurity) is chosen to split the node.

Things To Do ✓ Important Calls

8.00 ⇒ Overfitting: An induced tree may overfit the training data
8.30 ie. - too many branches, some may reflect ~~over~~ anomalies
9.00 due to noise or outliers
9.30 - poor accuracy for unseen samples.
10.00
10.30 ⇒ Methods to avoid overfitting -
11.00
11.30 1. Prepruning: Halt tree construction early - ie. do not
12.00 split a node if this would result in the goodness
12.30 measure falling below a threshold.
13.00
13.30 2. Postpruning: Removes branches from a "fully grown"
14.00 tree ie. to get a sequence of progressively
14.30 pruned trees.
15.00
15.30 * BOAT (Bootstrapped Optimistic Algorithm for Tree Construction) -
16.00
16.30 1. Use a statistical technique called bootstrapping to create several
17.00 smaller samples, each fits in memory.
17.30
18.00 2. Each subset is used to create a tree, resulting in several trees.
Evening
3. These trees are examined & used to construct a new tree T',
it turns out that T' is very close to tree that would be
generated using whole dataset together.

Meetings

✓ Things To Do

- ☐
- ☐
- ☐

✓ Important Calls ✓

- ☐
- ☐
- ☐

Bayesian Classification –

Given training data X, posteriori probability of a hypothesis H, P(H/X), follows Bayes Theorem

$$P(H/X) = \frac{P(X/H) \cdot P(H)}{P(X)}$$

→ Predicts X belongs to C_i iff the probability $P(C_i/X)$ is highest among all $P(C_k/X)$ for all k classes.

→ Reasons to select this method –

1. statistical classifier : performs probabilistic prediction
2. Strong foundation : Bayes theorem
3. Performance : better than decision trees
4. Incremental : each training example can incrementally increase/decrease the prob. that a hypothesis is correct.
5. Standard : they provide a standard of optimal decision making.

→ Naive Bayes Classifier –

It is based on a simplified assumption that attributes are conditionally independent.

$$P(X/C_i) = \prod_{K=1}^{n} P(x_k/C_i) = P(x_1/C_1) \times P(x_2/C_2) \dots P(x_n/C_n)$$

Advantages :

1. Easy to implement
2. Good results obtained in most of cases.

Disadvantages :

1. Based on assumption : class conditional independence.
2. it requires each conditional prob. be non-zero.

29 Week 26
June
Wednesday

→ Bayesian Belief Networks: it allows class conditional independencies between subsets of variables.

A (directed acyclic) graphical model of casual relationships
- Represents dependency among the variables
- Gives a specification of joint probability distribution.

Numerical

Classification by Backpropagation -

Backpropagation is a neural network learning algorithm. It iteratively process a set of training tuples & compare the network's prediction with actual known target value. For each training tuple, the weights are modified to minimize the mean square error ~~then~~, modifications are made in backward direction : from output layer to hidden layer →

Steps -

1. Initialize weights to small random no., associated with biases
2. Propagate the inputs forward (by applying activation fn)
3. Back propagate the error (by updating weights & biases)
4. Terminating condition (when error is small)

Meetings | ✓ Things To Do | ✓ Important Calls

Support Vector Machines -

A relatively new classification method for both linear & non-linear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane.
SVM finds this hyperplane using support vectors (essential training tuples) and margins (defined by support vectors)

let data D be $(x_1, y_1) \dots (x_n, y_n)$ where x_i is the set of training tuples associated with the class labels y_i. There are infinite lines (hyperplanes) separating the two classes but we want to find the best one (the one that minimizes classification error on unseen data) SVM searches for the hyperplane with largest margin ie. maximum marginal hyperplane (MMH).

SVM	Neural Network
• deterministic algo	non-deterministic algo.
• nice generalization properties.	Generalizes well but does not have strong mathematical foundation
• Hard to learn - in batch mode, using quadratic prog. techniques.	Can be easily learned in incremental fashion
• Using kernels can learn very complex fn.	To learn complex functions - use multilayer perceptron.

numerical

k- Nearest Neighbor Algo -

It is based on lazy learning (instance-based learning) i.e. it stores training examples and delay the processing until a new instance must be classified.

→ Algo -

- All instances correspond to points in the n-D space.
- The nearest neighbors are defined in terms of Euclidean distance $dist(X_1, X_2)$
- Target f° could be discrete or real world
- For discrete-valued, k-NN returns the most common value among the k-training examples nearest to x_q

Supervised learning - it is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct ans. After that, the machine is provided with a new set of data so that supervised learning algorithm analyses the training data and produces a correct outcome for labelled data.

It has 2 categories of algo

- classification
- Regression.

Unsupervised learning- It is the training of machine using info that is neither classified nor labeled & allowing the algorithm to act on that info without guidance. Here the task of machine is to group unsorted info according to similarities, patterns and differences without any prior training of data.

It is classified into 2 types of algos -

- Clustering.
- Association.