

Chapter 1 : Data Warehouse (DWH) Fundamentals with Introduction to Data Mining 1-1 to 1-44

Syllabus : DWH characteristics, Dimensional modeling : Star, Snowflakes, OLAP operation, OLTP vs OLAP Data Mining as a step in KDD, Kind of patterns to be mined, Technologies used, Data Mining applications.

Self-learning Topics : Data Marts, Major issues in Data Mining.

1.1	DWH Characteristics.....	1-1
1.1.1	Definition Data Warehouse	1-1
1.1.2	Benefits of Data Warehousing.....	1-1
1.1.3	Features of a Data Warehouse	1-2
1.2	Dimensional modelling : Star, Snowflakes.....	1-3
1.2.1	What is Dimensional Modelling?.....	1-3
1.2.2	Difference between Data Warehouse Modeling and Operational Database Modeling.....	1-3
1.2.3	Comparison between Dimensional Model and ER model.....	1-3
1.2.4	Information Package Diagram	1-4
1.2.5	Star Schema.....	1-5
1.2.6	STAR schema Keys.....	1-6
1.2.7	The Snowflake Schema	1-7
1.2.8	Star Flake Schema	1-7
1.2.9	Differentiate between Star Schema and Snowflake Schema.....	1-8
1.2.10	Fact Tables and Dimension Tables.....	1-8
1.2.11	Factless Fact Table	1-9
1.2.12	Fact Constellation Schema or Families of Star	1-10
1.2.13	Examples on Star Schema and Snowflake Schema.....	1-12
1.3	OLAP operation.....	1-26
1.3.1	OLAP operations or OLAP Techniques.....	1-26
1.3.1(A)	Consolidation or Roll Up.....	1-27
1.3.1(B)	Drill-down.....	1-28
1.3.1(C)	Slicing and dicing.....	1-29
1.3.1(D)	Dice.....	1-29
1.3.1(E)	Pivot / Rotate	1-30
1.3.1(F)	Other OLAP operations.....	1-30
1.3.2	Examples of OLAP	1-30
1.4	OLTP vs OLAP	1-34
1.5	Data Mining as a step in KDD	1-35
1.5.1	Definition.....	1-35
1.5.2	KDD Process (Knowledge Discovery in Databases)	1-36
1.5.3	Architecture of a Typical Data Mining System	1-37

1.6	Kind of Patterns to be Mined.....	1-38
1.6.1	Data Mining Functionalities	1-38
1.7	Technologies Used.....	1-40
1.7.1	Statistics.....	1-40
1.7.2	Machine Learning.....	1-40
1.7.3	Information Retrieval (IR).....	1-41
1.7.4	Database Systems and Data Warehouses	1-41
1.7.5	Decision Support System.....	1-41
1.8	Data Mining Applications	1-42
1.9	Self-learning Topics	1-42
1.9.1	Data Marts.....	1-42
1.9.2	Major Issues in Data Mining.....	1-43

Chapter 2 : Data Exploration and Data Preprocessing

2-1 to 2-53

Syllabus : Types of Attributes, Statistical Description of Data, Measuring Data Similarity and Dissimilarity. Why Preprocessing ? Data Cleaning, Data Integration, Data Reduction : Attribute Subset Selection, Histograms, Clustering, Sampling, Data Cube aggregation, Data transformation and Data Discretization : Normalization, Binning, Histogram Analysis. Self-learning Topics Data Visualization, Concept hierarchy generation.

2.1	Types of Attributes.....	2-1
2.1.1	Attributes Types.....	2-1
2.2	Statistical Description of Data.....	2-3
2.2.1	Central Tendency.....	2-4
2.2.2	Dispersion of Data.....	2-6
2.2.3	Graphic Displays of Basic Statistical Descriptions of Data	2-7
2.3	Measuring Similarity and Dissimilarity.....	2-15
2.3.1	Data Matrix versus Dissimilarity Matrix	2-15
2.3.2	Proximity Measures for Nominal Attributes.....	2-16
2.3.3	Proximity Measures for Binary Attributes.....	2-16
2.3.4	Dissimilarity of Numeric Data : Minkowski Distance	2-18
2.3.5	Proximity Measures for Ordinal Attributes	2-19
2.3.6	Dissimilarity for Attributes of Mixed Types.....	2-20
2.3.7	Cosine Similarity.....	2-21
2.4	Why Preprocessing ?	2-21
2.4.1	Why Pre-processing is Required ?	2-21
2.4.2	Different Forms of Data Pre-processing.....	2-22
2.5	Data Cleaning	2-22
2.5.1	Reasons for "Dirty" Data	2-22

2.5.2	Steps in Data Cleansing.....	2-23
2.5.3	Missing Values.....	2-24
2.5.4	Noisy Data.....	2-25
2.5.4(A)	Binning.....	2-25
2.5.4(B)	Outlier analysis by clustering.....	2-29
2.5.4(C)	Regression.....	2-30
2.5.5	Inconsistent Data.....	2-31
2.6	Data Integration.....	2-31
2.6.1	Introduction to Data Integration.....	2-31
2.6.1(A)	Entity Identification Problem.....	2-32
2.6.1(B)	Redundancy and Correlation Analysis.....	2-32
2.6.1(C)	Tuple Duplication.....	2-35
2.6.1(D)	Data Value Conflict Detection and Resolution.....	2-35
2.7	Data Reduction.....	2-36
2.7.1	Data Cube Aggregation.....	2-37
2.7.2	Dimensionality Reduction.....	2-37
2.7.2(A)	Attribute subset selection.....	2-38
2.7.3	Data Compression.....	2-39
2.7.4	Numerosity Reduction.....	2-40
2.7.4(A)	Histogram Analysis.....	2-40
2.7.4(B)	Clustering.....	2-41
2.7.4(C)	Sampling.....	2-41
2.8	Data transformation and Data Discretization.....	2-42
2.8.1	Data Transformation.....	2-42
2.8.2	Data Discretization.....	2-43
2.8.3	Data Transformation by Normalization.....	2-43
2.8.4	Discretization by Binning.....	2-46
2.8.5	Discretization by Histogram Analysis.....	2-47
2.9	Self-learning Topics.....	2-47
2.9.1	Data Visualisation.....	2-47
2.9.2	Concept Hierarchies.....	2-52
2.9.2(A)	Concept hierarchy generation for categorical data.....	2-52

Chapter 3 : Classification

3-1 to 3-78

Syllabus : Basic Concepts; Classification methods : 1. Decision Tree Induction: Attribute Selection Measures, Tree pruning.
 2. Bayesian Classification : Naïve Bayes Classifier. Prediction: Structure of regression models; Simple linear regression, Accuracy and Error measures, Precision, Recall, Holdout, Random Sampling, Cross Validation, Bootstrap, Introduction of Ensemble methods, Bagging, Boosting, AdaBoost and Random forest. Self-learning Topics : Multiple linear regression, logistic regression, Random forest, nearest neighbour classifier, SVM

3.1	Basic Concept : Classification	3-1
3.1.1	Classification Problem	3-1
3.1.2	Classification Example	3-2
3.1.3	Classification is a Two Step Process.....	3-2
3.1.4	Difference between Classification and Prediction.....	3-4
3.1.5	Issues Regarding Classification and Prediction.....	3-4
3.2	Classification Methods.....	3-5
3.2.1	Decision Tree Induction	3-5
3.2.1(A)	Appropriate Problems for Decision Tree Learning.....	3-5
3.2.1(B)	Decision Tree Representation.....	3-5
3.2.1(C)	Attribute Selection Measure.....	3-6
3.2.1(D)	Algorithm for Inducing a Decision Tree	3-9
3.2.2	Tree Pruning.....	3-10
3.2.3	Examples of ID3.....	3-11
3.3	Bayesian Classification : Naive Bayes Classifier	3-45
3.3.1	Bayes' Theorem.....	3-45
3.3.2	Basics of Bayesian Classification.....	3-45
3.3.3	Naive Bayes Classifier : Examples	3-45
3.3.4	Rule based Classification	3-57
3.3.5	Other Classification Methods.....	3-58
3.4	Prediction.....	3-58
3.4.1	Structure of Regression Model.....	3-59
3.4.2	Linear Regression.....	3-59
3.4.2(A)	Simple linear regression	3-59
3.5	Model Evaluation and Selection	3-60
3.5.1	Accuracy and Error Measures	3-61
3.5.2	Holdout.....	3-63
3.5.3	Random Sub-sampling.....	3-63
3.5.4	Cross-Validation (CV).....	3-64
3.5.5	Bootstrapping.....	3-65
3.6	Introduction of Ensemble methods.....	3-65
3.6.1	Bagged (or Bootstrap) trees.....	3-65
3.6.2	Boosted trees.....	3-66
3.6.3	Ada Boost.....	3-67
3.7	Regression	3-68
3.7.1	Multiple Linear Regression.....	3-68
3.7.2	Logistic Regression.....	3-68

3.7.3	Random forest.....	3-72
3.7.4	K-Nearest-Neighbor Classifiers.....	3-73
3.7.5	Support Vector Machine (SVM).....	3-75
3.7.5(A)	Tuning Hyperparameters.....	3-77

Chapter 4 : Clustering and Outlier Detection

4-1 to 4-72

Syllabus : Cluster Analysis : Basic Concepts; Partitioning Methods: K-Means, K Medoids ; Hierarchical Methods: Agglomerative, Divisive, BIRCH; Density-Based Methods: DBSCAN. What are outliers? Types, Challenges; Outlier

Detection Methods : Supervised, Semi Supervised, Unsupervised, Proximity based, Clustering Based. Self-learning Topics
Hierarchical methods : Chameleon, Density based methods: OPTICS, Grid based methods: STING, CLIQUE.

4.1	Cluster Analysis.....	4-1
4.1.1	What is Clustering ?.....	4-1
4.1.2	Categories of Clustering Methods.....	4-2
4.1.3	Different Distance Measures that can be used to Compute Distances between Two Clusters.....	4-3
4.1.4	Difference between Classification and Clustering.....	4-4
4.2	Partitioning Methods : K-Means, K Medoids.....	4-5
4.2.1	K-means Clustering : (Centroid Based Technique).....	4-5
4.2.2	K-Medoids (Representative Object-based Technique).....	4-19
4.2.3	Sampling Based Method.....	4-24
4.3	Hierarchical Methods : Agglomerative, Divisive, BIRCH.....	4-24
4.3.1	Agglomerative Hierarchical Clustering.....	4-26
4.3.2	Divisive Hierarchical Clustering.....	4-53
4.3.3	BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).....	4-54
4.3.4	Advantages and Disadvantages of Hierarchical Clustering.....	4-57
4.4	Density-Based Methods: DBSCAN.....	4-57
4.4.1	DBSCAN (Density Based Methods).....	4-58
4.5	What is an Outlier ?.....	4-60
4.5.1	Applications.....	4-60
4.6	Types of Outliers.....	4-61
4.6.1	Global Outliers.....	4-61
4.6.2	Contextual (or Conditional) Outliers.....	4-62
4.6.3	Collective Outliers.....	4-62
4.7	Challenges of Outlier Detection.....	4-63
4.8	Outlier Detection Methods.....	4-63
4.8.1	Supervised, Semi - Supervised, Unsupervised Methods.....	4-63
4.8.2	Statistical Methods, Proximity-based Methods and Clustering-based Methods.....	4-64

4.9	Proximity based Approaches.....	4-64
4.9.1	Distance-based Outlier Detection and a Nested Loop Method.....	4-65
4.9.2	A Grid based Method.....	4-65
4.9.3	Density based Outlier Detection	4-66
4.10	Clustering based Approaches.....	4-68
4.11	Self-learning Topics	4-70
4.11.1	Hierarchical methods : Chameleon.....	4-70
4.11.2	Density based methods : OPTICS.....	4-70
4.11.3	Grid based methods : STING, CLIQUE.....	4-72

Chapter 5 : Frequent Pattern Mining

5-1 to 5-48

Syllabus : Basic Concepts : Market Basket Analysis, Frequent Itemset, Closed Itemset, and Association Rules, Mining Methods: The Apriori Algorithm: Finding Frequent Itemset Using Candidate Generation, Generating Association Rules from Frequent Itemset, Improving the Efficiency of Apriori, A pattern growth approach for mining Frequent Itemset, Mining Frequent Itemset using vertical data formats; Introduction to Advance Pattern Mining : Mining Multilevel Association Rules and Multidimensional Association Rules.

Self-learning Topics : Association Mining to Correlation Analysis, lift, Introduction to Constraint-Based Association Mining

5.1	Basic Concept : Market Basket Analysis.....	5-1
5.1.1	What is Market Basket Analysis?.....	5-1
5.1.2	How is it Used ?.....	5-1
5.1.3	Applications of Market Basket Analysis	5-2
5.2	Frequent Itemsets, Closed Itemsets and Association Rules	5-2
5.2.1	Frequent Itemsets.....	5-2
5.2.2	Closed Itemsets.....	5-3
5.2.3	Association Rules.....	5-4
5.2.3(A)	Large Itemsets.....	5-4
5.3	Frequent Pattern Mining	5-5
5.4	Frequent Itemset Mining Method	5-5
5.4.1	Apriori Algorithm for Finding Frequent Itemsets using Candidate Generation.....	5-5
5.4.2	Generating Association rules from frequent itemsets.....	5-7
5.4.3	Advantages and Disadvantages of Apriori Algorithm.....	5-7
5.4.4	Solved Examples on Apriori Algorithm	5-7
5.4.4	Improving the Efficiency of Apriori.....	5-28
5.5	A Pattern Growth Approach for Mining Frequent Itemsets (FP-Growth).....	5-28
5.5.1	Definition of FP-tree.....	5-28
5.5.2	FP-Tree Algorithm	5-29
5.5.3	FP-Tree Size	5-30

5.5.4	Example of FP Tree.....	5-30
5.5.5	Mining Frequent Patterns from FP Tree.....	5-34
5.5.6	Benefits of the FP-Tree Structure.....	5-39
5.6	Mining Frequent Itemsets using Vertical Data Formats.....	5-39
5.7	Mining Closed and Maximal Patterns.....	5-40
5.8	Mining Multilevel Association Rules.....	5-41
5.9	Mining Multidimensional (MD) Association Rules.....	5-42
5.10	Association Mining to Correlation Analysis.....	5-45
5.11	Pattern Evaluation Measures.....	5-45
5.12	Introduction to Constraint based Association Mining.....	5-47

Chapter 6 : Business Intelligence

6-1 to 6-20

Syllabus : What is BI? Business intelligence architectures; Definition of decision support system; Development of a business intelligence system using Data Mining for business Applications like Fraud Detection, Recommendation System

Self-learning Topics : Clickstream Mining, Market Segmentation, Retail industry, Telecommunications industry, Banking & finance CRM, Epidemic prediction, Fake News Detection, Cyberbullying, Sentiment Analysis etc.

6.1	What is Business Intelligence?.....	6-1
6.2	Business Intelligence Architectures.....	6-1
6.2.1	The Three Major Components of BI Architecture.....	6-2
6.2.2	Different Components of a Business Intelligent System.....	6-3
6.3	Definition of Decision Support System.....	6-3
6.4	Development of a Business Intelligence System.....	6-5
6.5	Business Intelligence.....	6-6
6.5.1	Business Intelligence Issues.....	6-7
6.6	Fraud Detection.....	6-8
6.7	Recommendation System.....	6-10
6.8	Clickstream Mining.....	6-11
6.8.1	Clickstream Data : Collection and Restoration.....	6-11
6.8.2	Clickstream Data: Visualisation and Categorisation.....	6-12
6.9	Market Segmentation.....	6-12
6.9.1	Market Segmentation for Market Trend Analysis.....	6-12
6.9.2	Sales Trend Analysis.....	6-13
6.10	Retail Industry.....	6-13
6.11	Telecommunications Industry.....	6-14
6.12	Banking and Finance.....	6-15

6.13	CRM.....	6-16
	6.13.1 Data Mining Challenges and Opportunities in CRM.....	6-17
6.14	Epidemic Prediction.....	6-18
6.15	Fake News Detection.....	6-18
6.16	Cyberbullying.....	6-19
6.17	Sentiment Analysis.....	6-19
