

Syllabus

Savitribai Phule Pune University Fourth Year of Computer Engineering (2015 Course) 410243 : Data Analytics

Teaching Scheme
TH : 3 Hours/Week

Credit
03

Examination Scheme :
In-Sem (Paper) : 30 Marks
End-Sem (Paper) : 70 Marks

Prerequisite Courses : 310242-Database Management Systems

Companion Course : 410246-Laboratory Practice I

Course Objectives

- To develop problem solving abilities using Mathematics
- To apply algorithmic strategies while solving problems
- To develop time and space efficient algorithms
- To study algorithmic examples in distributed, concurrent and parallel environments

Course Outcomes

On completion of the course, student will be able to–

- Write case studies in Business Analytic and Intelligence using mathematical models.
- Present a survey on applications for Business Analytic and Intelligence.
- Provide problem solutions for multi-core or distributed, concurrent/Parallel environments.

Course Contents

UNIT I : Introduction and Life Cycle

(08 Hours)

Introduction: Big data overview, state of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach. Data Analytic Life Cycle: Overview, phase 1- Discovery, Phase 2- Data preparation, Phase 3- Model Planning, Phase 4- Model Building, Phase 5- Communicate Results, Phase 6- Operationalize. Case Study: GINA (Refer Chapter 1)

UNIT II : Basic Data Analytic Methods

(08 Hours)

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank-sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods: Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions. (Refer Chapter 2)

(08 Hours)

UNIT III : Association Rules and Regression

Advanced Analytical Theory and Methods: Association Rules- Overview, a-priori algorithm, evaluation of candidate rules, case study-transactions in grocery store, validation and testing, diagnostics. Regression- linear, logistics, reasons to choose and cautions, additional regression models.

(Refer Chapter 3)

(08 Hours)

UNIT IV : Classification

Decision trees- Overview, general algorithm, decision tree algorithm, evaluating a decision tree. Naïve Bayes – Bayes' Algorithm, Naïve Bayes' Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods.

(Refer Chapter 4)

(08 Hours)

UNIT V : Big Data Visualization

Introduction to Data visualization, Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Analytical techniques used in Big data visualization.

(Refer Chapter 5)

(08 Hours)

UNIT VI : Advanced Analytics-Technology and Tools

Analytics for unstructured data- Use cases, Map Reduce, Apache Hadoop. The Hadoop Ecosystem- Pig, HIVE, HBase, Mahout, NoSQL. An Analytics Project-Communicating, operationalizing, creating final deliverables.

(Refer Chapter 6)

□□□

**UNIT I****Chapter 1 : Introduction and Life Cycle 1-1 to 1-23****Syllabus :**

Introduction : Big data overview, state of the practice in Analytics
- BI Vs Data Science, Current Analytical Architecture, drivers of
Big Data, Emerging Big Data Ecosystem and new approach.

Data Analytic Life Cycle : Overview, phase 1 - Discovery, Phase
2 - Data preparation, Phase 3 - Model Planning,
Phase 4 - Model Building, Phase 5 - Communicate Results,
Phase 6 - Operationalize. Case Study : GINA.

✓	Syllabus Topic : Big Data Overview	1-1
1.1	Big Data Overview.....	1-1
1.1.1	Defining Data Science and Big Data.....	1-3
1.1.2	Examples of Big Data Applications.....	1-3
1.1.3	Data Explosion.....	1-4
1.1.4	Data Volume.....	1-5
1.1.5	Data Velocity.....	1-7
1.1.6	Big Data Infrastructure and Challenges.....	1-8
✓	Syllabus Topic : State of the Practice in Analytics BI Vs Data Science	1-10
1.2	State of Practice in Analytics BI Vs Data Science	1-10
✓	Syllabus Topic : BI Vs Data Science	1-10
1.2.1	BI Vs Data Science.....	1-10
✓	Syllabus Topic : Current Analytical Architecture	1-12
1.2.2	Current Analytical Architecture.....	1-12
✓	Syllabus Topic : Drivers of Big Data	1-13
1.2.3	Drivers of Big Data.....	1-13
✓	Syllabus Topic : Emerging Big Data Ecosystem and New Approach	1-14
1.2.4	Emerging Big Data Ecosystem and New Approach.....	1-14
✓	Syllabus Topic : Data Analytic Life Cycle - Overview	1-16
1.3	Data Analytic Life Cycle : Overview.....	1-16
✓	Syllabus Topic : Phase 1 - Discovery Phase	1-17
1.3.1	Phase 1 - Discovery Phase.....	1-17
✓	Syllabus Topic : Phase 2 - Data Preparation	1-17
1.3.2	Phase 2 - Data Preparation.....	1-17
✓	Syllabus Topic : Phase 3 - Model Planning	1-17
1.3.3	Phase 3 - Model Planning.....	1-17
✓	Syllabus Topic : Phase 4 - Model Building	1-18
1.3.4	Phase 4 - Model Building.....	1-18
✓	Syllabus Topic : Phase 5 - Communicate Results	1-18
1.3.5	Phase 5 - Communicate Results.....	1-18

✓	Syllabus Topic : Phase 6 - Operationalize	1-18
1.3.6	Phase 6 - Operationalize.....	1-18
✓	Syllabus Topic : Case Study - GINA	1-19
1.4	Case Study - GINA : Global Innovation Network and Analysis.....	1-19
1.4.1	Phase 1 - Discovery.....	1-19
1.4.2	Phase 2 - Data Preparation.....	1-20
1.4.3	Phase 3 - Model Planning.....	1-20
1.4.4	Phase 4 - Model Building.....	1-21
1.4.5	Phase 5 - Communicate Results.....	1-22
1.4.6	Phase 6 - Operationalize.....	1-22

UNIT II**Chapter 2 : Basic Data Analytic Methods 2-1 to 1-17****Syllabus :**

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank-sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods : Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions.

✓	Syllabus Topic : Statistical Methods for Evaluation	2-1
2.1	Statistical Methods for Evaluation.....	2-1
✓	Syllabus Topic : Hypothesis Testing	2-1
2.1.1	Hypothesis Testing.....	2-1
✓	Syllabus Topic : Difference of Means	2-2
2.1.2	Difference of Means.....	2-2
2.1.2(A)	Student's t-test.....	2-3
2.1.2(B)	Welch's t-test.....	2-4
✓	Syllabus Topic : Wilcoxon Rank-Sum Test	2-4
2.1.2(C)	Wilcoxon Rank-Sum Test.....	2-4
✓	Syllabus Topic : Type I and Type II Errors	2-5
2.1.3	Type I and Type II Errors.....	2-5
✓	Syllabus Topic : Power and Sample Size	2-6
2.1.4	Power and Sample Size.....	2-6
✓	Syllabus Topic : ANNOVA	2-6
2.1.5	ANNOVA.....	2-6
✓	Syllabus Topic : Advanced Analytical Theory and Methods	2-7
2.2	Advanced Analytical Theory and Methods.....	2-7
✓	Syllabus Topic : Overview of Clustering	2-8
2.2.1	Overview of Clustering.....	2-8
✓	Syllabus Topic : K-Means	2-8
2.2.2	K- Means.....	2-8

✓	Syllabus Topic : Use Cases	2-9
2.2.3	Use Cases.....	2-9
✓	Syllabus Topic : Overview of the Methods	2-9
2.2.4	Overview of the Method.....	2-9
✓	Syllabus Topic : Determining the Number of Clusters	2-11
2.2.5	Determining the Number of Clusters.....	2-11
✓	Syllabus Topic : Diagnostics	2-13
2.2.6	Diagnostics.....	2-13
✓	Syllabus Topic : Reasons to Choose and Cautions	2-14
2.2.7	Reasons to Choose and Cautions	2-14

UNIT III

Chapter 3 : Association Rules and Regression

3-1 to 3-18

Syllabus :

Advanced Analytical Theory and Methods : Association Rules- Overview, a-priori algorithm, evaluation of candidate rules, case study-transactions in grocery store, validation and testing, diagnostics. Regression- linear, logistics, reasons to choose and cautions, additional regression models.

✓	Syllabus Topic : Advanced Analytical Theory and Methods : Association Rules - Overview	3-1
3.1	Advanced Analytical Theory and Methods : Association Rules- Overview	3-1
✓	Syllabus Topic : Apriori Algorithm	3-2
3.2	Apriori Algorithm.....	3-2
3.2.1	Limitations of Apriori Algorithm.....	3-4
✓	Syllabus Topic : Evaluation of Candidate Rules	3-5
3.3	Evaluation of Candidate Rules	3-5
✓	Syllabus Topic : Case Study - Transactions in Grocery Store	3-6
3.4	Case Study - Transactions in Grocery Store.....	3-6
✓	Syllabus Topic : Validation and Testing	3-10
3.5	Validation and Testing	3-10
✓	Syllabus Topic : Diagnostics	3-11
3.6	Diagnostics.....	3-11
3.7	Regression Analysis.....	3-12
✓	Syllabus Topic : Linear Regression	3-13
3.7.1	Linear Regression	3-13
✓	Syllabus Topic : Logistic Regression	3-14
3.7.2	Logistic Regression	3-14
✓	Syllabus Topic : Reasons to Choose and Cautions	3-16
3.7.3	Reasons to Choose and Cautions	3-16
✓	Syllabus Topic : Additional Regression Models	3-17
3.7.4	Additional Regression Models	3-17

UNIT IV

Chapter 4 : Classification

4-1 to 4-22

Syllabus :

Decision trees - Overview, general algorithm, decision tree algorithm, evaluating a decision tree. Naïve Bayes -Bayes' Algorithm, Naïve Bayes' Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods.

4.1	Decision Trees : Introduction	4-1
✓	Syllabus Topic : Decision Trees- Overview	4-2
4.1.1	Decision Trees : Overview	4-2
✓	Syllabus Topic : General Algorithm	4-5
4.1.2	General Algorithm	4-5
✓	Syllabus Topic : Decision Tree Algorithms	4-9
4.1.3	Decision Tree Algorithms	4-9
✓	Syllabus Topic : Evaluating a Decision Tree	4-11
4.1.4	Evaluating a Decision Tree.....	4-11
✓	Syllabus Topic : Naïve Bayes	4-13
4.2	Naïve Bayes.....	4-13
✓	Syllabus Topic : Bayes' Algorithm	4-14
4.2.1	Bayes' Algorithm	4-14
✓	Syllabus Topic : Naïve Bayes Classifier	4-14
4.2.2	Naïve Bayes Classifier	4-14
✓	Syllabus Topic : Smoothing	4-16
4.2.3	Smoothing.....	4-16
✓	Syllabus Topic : Diagnostics	4-17
4.2.4	Diagnostics.....	4-17
✓	Syllabus Topic : Diagnostics of Classifiers	4-18
4.2.5	Diagnostics of Classifiers	4-18
✓	Syllabus Topic : Additional Classification Methods	4-21
4.2.6	Additional Classification Methods	4-21

UNIT V

Chapter 5 : Big Data Visualization

5-1 to 5-30

Syllabus :

Introduction to Data visualization, Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Analytical techniques used in Big data visualization

✓	Syllabus Topic : Introduction to Data Visualization	5-1
5.1	Introduction to Data Visualization	5-1



✓	Syllabus Topic : Challenges to Big Data Visualization5-2
5.2	Challenges to Big Data Visualization.....5-2
✓	Syllabus Topic : Conventional Data Visualization Tools5-5
5.3	Conventional Data Visualization Tools.....5-5
✓	Syllabus Topic : Techniques for Visual Data Representations5-7
5.4	Techniques for Visual Data Representations.....5-7
✓	Syllabus Topic : Types of Data Visualization5-8
5.5	Types of Data Visualization.....5-8
✓	Syllabus Topic : Visualizing Big Data5-17
5.6	Visualizing Big Data.....5-17
✓	Syllabus Topic : Tools used in Data Visualization5-18
5.7	Tools used in Data Visualization.....5-18
5.8	Open-Source Data Visualization Tools.....5-20
5.9	Data Visualization with Tableau.....5-21
5.10	Introduction to : Pentaho, Flare, Jasper Reports, Dygraphs, Datameer Analytics Solution and Cloudera, Platfora, NodeBox, Gephi, Google Chart API, Flot, D3, Visual.ly.....5-22
✓	Syllabus Topic : Analytical Techniques used in Big Data Visualization5-27
5.11	Analytical Techniques used in Big Data Visualization.....5-27

UNIT VI

Chapter 6 : Advanced Analytics - Technology and Tools 6-1 to 6-24

Syllabus :

Analytics for unstructured data - Use cases, MapReduce, Apache Hadoop. The Hadoop Ecosystem - Pig, HIVE, HBase, Mahout, NoSQL. An Analytics Project - Communicating, operationalizing, creating final deliverables.

✓	Syllabus Topic : Analytics for Unstructured Data6-1
6.1	Analytics for Unstructured Data.....6-1

✓	Syllabus Topic : Use Cases6-2
6.1.1	Use Cases.....6-2
✓	Syllabus Topic : Apache Hadoop6-3
6.1.2	Apache Hadoop.....6-3
6.1.2(A)	Modules (Components) of Hadoop.....6-3
✓	Syllabus Topic : MapReduce6-4
6.1.3	MapReduce.....6-4
6.1.4	Hadoop Distributed File System (HDFS).....6-5
✓	Syllabus Topic : The Hadoop Ecosystem6-6
6.2	The Hadoop Ecosystem.....6-6
✓	Syllabus Topic : Pig6-6
6.2.1	Pig.....6-6
✓	Syllabus Topic : HIVE6-7
6.2.2	HIVE.....6-7
✓	Syllabus Topic : HBase6-8
6.2.3	HBase.....6-8
6.2.3(A)	HBase Data Model.....6-9
6.2.3(B)	HBase Regions.....6-10
✓	Syllabus Topic : Mahout6-11
6.2.4	Mahout.....6-11
✓	Syllabus Topic : NoSQL6-12
6.3	NoSQL.....6-12
6.3.1	Types and Examples of NoSQL Database.....6-15
6.3.2	Comparative Study of SQL and NoSQL.....6-18
6.3.3	NoSQL Data Models.....6-19
✓	Syllabus Topic : An Analytics Project - Communicating, Operationalizing6-20
6.4	An Analytics Project - Communicating, Operationalizing.....6-20
✓	Syllabus Topic : Creating Final Deliverables6-21
6.5	Creating Final Deliverables.....6-21
6.5.1	Developing Core Material for Multiple Audiences.....6-22
6.5.2	Project Goals.....6-23
6.5.3	Approach.....6-23
6.5.4	Model Description.....6-24
•	Lab ManualL-1 to L-87
•	Questions and Answers for In - Semester ExaminationQ-1 to Q-26