St Xavier's College (Autonomous), Ahmedabad

M. Sc. Programme in Big Data Analytics

Course Structure
Semester-1

All courses compulsory
1. Statistical Methods
2. Probability & Stochastic Process
3. Linear Algebra & Linear Programming
4. Computing for Data Sciences
5. Database Management

Semester-2

Compulsory courses:
1. Foundations of Data Science
2. Advance Statistical Methods
3. 3 Machine Learning I
4. Datamining

Elective courses: (choose 1 from 2)
1. Multivariate Statistics
2. Operations Research

Semester-3

Compulsory courses:
1. Modelling in Operations Management
2. Enabling Technologies for Data Science
3. Value Thinking

Elective courses: (choose 2 from 4)

1. Introduction to Econometrics & Finance
2. Machine Learning II
3. Time series Analysis & Forecasting
4. Bio informatics

Semester-4

1. Internship based project.

Semester-1

1. Basic Statistical Methods : 60hours

A) **Data Collection & Visualization :**(25 hrs – Theory 17 hrs + Lab 8 hrs)

Concepts of measurement, scales of measurement, design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data, principles of data visualization, different methods of presenting data in business analytics.

B) **Basic Statistics :** (25 hrs – Theory 17 hrs + Lab 8 hrs)

Frequency table, histogram, measures of location, measures of spread, skewness, curtosis, percentiles, box plot, correlation and simple linear regression, partial correlation, probability distribution as a statistics model, fitting probability distributions, empirical distributions, checking goodness of fit through plots and tests.

C) **Contingency Tables :** (10 hrs – Theory 8 hrs + Lab 2 hrs)

Two way contingency tables, measures of association, testing for dependence.

Suggested books :

1. Statistics : David Freedman, Pobert Pisani & Roger Purves, WW.Norten & Co. 4th Edition 2007.
2. The visual display of Quantitative Information : Edward Tufte, Graphics Press, 2001.
3. Best Practices in Data Cleaning : Jason W. Osborne, Sage Publications 2012.

Evaluation: Theory: 70% + Practical/Lab. : 30%

2. Probability & Stochastic Process : (60 Hours)

A) **Basic Probability :** (20 hrs – Theory 18 hrs + Lab 2 hrs)

Concepts of experiments, Outcomes, Sample space, Events, Combinatorial probability, Birthday paradox, Principle of inclusion & exclusion, Conditional probability, Independence, Bayes Theorem.

B) **Probability Distribution:** (20 hrs – Theory 16 hrs + Lab 4 hrs)

Random Variables : discrete and continuous probability models, some probability distributions : Binomial, Poisson, Geometric, Hypergeometric, Normal, exponential, Chi-square, expectation, variance and other properties of the distribution.

C) **Stochastic Process :** (10 hrs – Theory 4 hrs + Lab 6 hrs)
Markov Chains, Classification of states, Stationery distribution, limit theorems, Poisson process, illustrations and applications.

D) **Introduction to Time Series :** (10 hrs – Theory 4 hrs + Lab 6 hrs)
Components of time series, Smoothing auto correlation, stationarity, concepts of AR, MA, ARMA & ARIMA models with illustrations.

Suggested Books:

1. A First Course in Probability : Shelden M. Ross, 2014.
2. Introduction to Stochastics Process : Paul G. Hoel, Sydney C. Port & Charles J. Stone, Waveland Press, 1987.
3. Time Series Analysis and Its Applications : Robert H. Shumway and David S. Stoffer, Springer 2010.

Evaluation: Theory: 70% + Practical/Lab: 30%

3. Linear algebra & linear programming (60 hrs):

A) **Linear Algebra :** (40 hrs – Theory 28hrs + Lab 12 hrs)
Linear equations and matrices, matrix operations, solving system of linear equations, Gauss-Jordan method, Concept & Computation of determinant and inverse of matrix, Eigen values and eigen vectors, Illustrations of the methods, Positive semi definite and position definite matrices, illustrations.

B) **Linear Programming** (20 hrs – Theory 14hrs + Lab 6hrs)
Definition of the problem, convex sets, corner points, feasibility, basic feasible solutions, Simplex method

Suggested Books:

1. Linear Algebra and Its Application : Gilbert Strang, 4th Edition, Academic Press.

2. Linear Programming : G. Hadley, Addison-Wesley.

Evaluation: Theory: 70% + Practical/Lab: 30%

4. Computing for Data Sciences (60hrs):

A) **Computer Package : (**20 hrs – Theory 14hrs + Lab 6 hrs)
Usage of R with illustration.

B) **Concepts of Computation (**20 hrs – Theory 2 hrs + Lab 18 hrs)
Algorithms, Convergence, Complexity with illustrations, Some sorting & searching algorithms, Some numerical methods e.g. Newton-Raphson, Steepest ascent.

C) **Computing Methodologies: (**20 hrs – Theory 8hrs + Lab 12 hrs)
Monte-Carlo simulations of random numbers and various statistical methods, memory handling strategies for big data.

Suggested Books: No specific book.

Evaluation: Theory: 40% + Practical/Lab: 60%

5. Database Management (60 hrs):

A) **Basic Concepts : (15** hrs – Theory 15hrs)
Different data models, ER and EER diagram, schema, table.

B) **Relational data base : (**20 hrs – Theory 8 hrs + Lab 12 hrs)
Structure, various operations, normalization, SQL, Parallel and distributed data base.

C) **Implementation : (**25 hrs – Theory 13hrs + Lab 12 hrs)
ORACLE, Concept of data base security.

Suggested Books:

1. Database system concepts : Abraham Silberschartz, Henry F. Korth and S. Surarshan, McGraw Hill, 2011.

Evaluation: Theory: 60% + Practical/Lab: 40%

Semester-2

1. Foundations of data science (programming for big data) 60 hrs:

A) **Graph Theory : (**10 hrs – Theory 4hrs + Lab 6 hrs)
Basic Concepts, Algorithms for connectedness, Shortest path, Minimum Sampling Tree

B) **High Dimensional Space : (12** hrs – Theory 6hrs + Lab 6 hrs)
Properties, Law of large numbers, Sphere and cube in high dimension, Generating points on the surface of a sphere, Gaussians in High dimension, Random projection, Applications.

C) **Random Graphs : (12** hrs – Theory 6hrs + Lab 6 hrs)
Large graphs, G(n,p) model, Giant Component, Connectivity, Cycles, Non-Uniform models, Applications.

D) **Singular Value Decomposition (SVD): (5** hrs – Theory 1hrs + Lab 4 hrs)
Best rank k approximation, Power method for computing the SVD, Applications.

E) **Random Walks : (5** hrs – Theory 1hrs + Lab 4 hrs)
Reflection Principle, Long leads, Changes of Sign, Illustrations.

F) **Algorithm for Massive Data Problems : (16** hrs – Theory 6hrs + Lab 10 hrs)
Frequency Moments of data streams, matrix algorithms.

Suggested book :

1. Foundations of Data Science : John Hopcroft & Ravindran Kannan.

Evaluation: Theory: 40% + Practical/Lab: 60%

2. Advance Statistical Methods : (60 hrs)

A) **Estimation : (15 hrs** – Theory 13hrs + Lab 2 hrs)
Unbiasedness, Consistency, UMVUE, Maximum likelihood estimates. (15 hrs)

B) **Test of Hypotheses : (15 hrs** – Theory 13hrs + Lab 2 hrs)
Two types of errors, test statistic, parametric tests for equality of means & variances.

C) **Linear Model : (15 hrs** – Theory 9hrs + Lab 6 hrs)
Gauss Markov Model, least square estimators, Analysis of variance.

D) **Regression : (15 hrs** – Theory 9hrs + Lab 6 hrs)
Multiple linear regression , forward, backward & stepwise regression, Logistic Regression.

Suggested Books :

1. Statistical Inference : P. J. Bickel and K. A. Docksum, 2nd Edition, Prentice Hall.
2. Introduction to Linear Regression Analysis : Douglas C. Montgomery

Evaluation: Theory: 70% + Practical/Lab: 30%

3. Machine Learning (60 Hrs)

A) **Linear Regression (10 hrs** – Theory 6hrs + Lab 4 hrs)
Linear Regression with Multiple variables, applications.

B) **Logistic Regression : (10 hrs** – Theory 4hrs + Lab 6hrs)
Model, Classification, Problem of over-fitting, Applications.

C) **Neural Networks : (9 hrs** – Theory 3hrs + Lab 6 hrs)
Representation Learning, Different Models like single and multi-layer perceptron, back propagation, Application.

D) **Machine Learning System Design : (8 hrs** – Theory 2hrs + Lab 6 hrs)
Evaluating a learning algorithms, handling skewed data, using large data sets. (5 hrs)

E) **Support Vector Machines : (5 hrs** – Theory 3hrs + Lab 2 hrs)
Model, Large Margin Classification, Kernels, SVMs in practice. (5 hrs)

F) Unsupervised Learning. **(5 hrs** – Theory 3hrs + Lab 2 hrs)

G) Dimensionality Reduction. **(8 hrs** – Theory 6hrs + Lab 2 hrs)

H) Anomaly Detection. **(5 hrs** – Theory 3hrs + Lab 2 hrs)

Suggested Books :

1. Machine Learning : Tom Mit chell

Evaluation: Theory: 50% + Practical/Lab: 50%

4. Data Mining: (60 Hrs)

A) **Introduction : (5 hrs** – Theory 5hrs)
Knowledge discovery from databases, scalability issues.

B) **Data Warehousing: (8 hrs** – Theory 2 hrs + Lab 6 hrs)
General principles, modeling, design, implementation and optimization.

C) **Data Preparation : (5 hrs** – Theory 1hrs + Lab 4 hrs)
Pre-processing, sub-sampling, feature selection.

D) **Classification and Prediction: (18 hrs** – Theory 10 hrs + Lab 8 hrs)
Bayes learning, decision trees, CART, neural learning, support vector machines, associations, dependence analysis, rule generation.

E) **Cluster Analysis and Deviation Detection: (14 hrs** – Theory 6hrs + Lab 8 hrs)
Partitioning algorithms, Density bases algorithm, Grid based algorithm, Graph theoretic clustering.

F) Temporal and spatial data mining. **(10 hrs** – Theory 6 hrs + Lab 4 hrs)

Suggested Books

1. Data Mining Techniques : A. K. Pujari, Sangam Books Ltd., 2001

2. Mastering Data Mining : M. Berry and G. Linoff, John Wiley & Sons., 2000

Evaluation: Theory: 50% + Practical/Lab: 50%

5. Multivariate Statistics : (60 Hrs)

A) Representation of multivariate data, bivariate and multivariate distributions, multinomial distribution, multivariate normal distribution, sample mean & sample dispersion matrix, concepts of location & depth in multivariate data. (20 hrs – Theory 12 hrs + Lab 8 hrs)
B) Principal Component Analysis (10 hrs – Theory 6 hrs + Lab 4 hrs)
C) Classification (10 hrs – Theory 6 hrs + Lab 4 hrs)
D) Factor Analysis(10 hrs – Theory 6 hrs + Lab 4 hrs)
E) Clustering(10 hrs – Theory 6 hrs + Lab 4 hrs)

Suggested Books :

1. Applied Multivariate Statistical Analysis : Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002

Evaluation: Theory: 60% + Practical/Lab: 40%

6. Operations Research : (60 Hrs)

A) Review of Linear Programming. (5 hrs – Theory 5 hrs)
B) Non-Linear Programming. (10 hrs – Theory 6 hrs + Lab 4 hrs)
C) Assignment Models. (5 hrs – Theory 1 hrs + Lab 4 hrs)
D) Transportation Models. (15 hrs – Theory 11 hrs + Lab 4 hrs)
E) Queuing Models : Characteristics of Queuing Process, Poisson Process, Birth-Death Process, Single-Server Queues, Multi-ServerQueues, Queues with Truncation, Finite-Source Queues, Numerical Techniques & Simulation. (25 hrs – Theory 19 hrs + Lab 6 hrs)

Suggested Books :

1. Operations Research : Prem Kumar Gupta & D. S. Hira

2. Fundamentals of Queuing Theory : Donald Gross, John F. Shortle, James M. Thompson & Carl M. Harris, Fourth Edition, Wiley

Evaluation: Theory: 70% + Practical/Lab: 30%

Semester-3

1. **Modelling In Operations Management : (60 Hrs)**
   A) Venture analytics (10 hrs – Lab 10 hrs)
   B) Banking analytics (10 hrs – Lab 10 hrs)
   C) Marketing analytics (10 hrs – Lab 10 hrs)
   D) Healthcare analytics (10 hrs – Lab 10 hrs)
   E) Retail analytics (10 hrs – Lab 10 hrs)
   F) Supply chain analytics (10 hrs – Lab 10 hrs)

Suggested Books: None

Evaluation: Practical / Lab / Report: 100%

2. **Enabling Technologies For Data Science (60 hrs) :**
   A) Big data and Hadoop : Hadoop architecture, Hadoop Versioning and configuration, Single node & Multi-node Hadoop, Hadoop commands, Models in Hadoop, Hadoop daemon, Task instance, Illustrations.
   B) Map-Reduce : Framework, Developing Map-Reduce program, Life cycle method, Serialization, Running Map-Reduce in local and pseudo-distributed mode, Illustrations.
   C) HIVE: Installation, data types and commands, Illustrations.
   D) SQOOP : Installation, Importing data, Exporting data, Running, Illustrations
   E) PIG: Installation, Schema, Commands, Illustrations.

Suggested Books:

1. Hadoop in Action : Chuck Lam, 2010, ISBN : 9781935182191

2. Data-intensive Text Processing with Map Reduce : Jimmy Lin and Chris Dyer, Morgan & Claypool Publishers, 2010

Evaluation: Practical / Lab : 100%

3. **Value Thinking (60 hrs) :**
   This course involves watching few movies (list provided below) and reading few books (list provided below) that deals mostly with argumentative logic, evidence, drawing inference from evidences. After watching the movies and reading the books, there will be general discussion amongst the students. Couple of case studies that involve mostly logical thinking will also be presented. Each student will prepare a term paper. Evaluation will be on the basis of this term paper and participation in group discussion.

Movies:

1. Twelve Angry Men
2. Roshoman by Kurosawa
3. Trial of Nuremberg
4. Mahabharata by Peter Brook

Suggested Books:

1. The Hound of the Baskervilles by Arthur Conan Doyle
2. Five Little Pigs by Agatha Christie
3. The Purloined Letter by Edger Allan Poe
4. The Case of the Substitute Face

Case Studies:

4. Introduction to Econometrics & Finance (60 hrs):

A) Analysis of Panel Data. **(19 hrs** – Theory 16 hrs + Lab 3 hrs)

B) Generalized Method of Moments (GMM). **(18 hrs** – Theory 16 hrs + Lab 3 hrs)

C) **Simultanuous Equations System : (7 hrs** – Theory 4 hrs + Lab 3 hrs)
Least Squares, Bias Problem, Estimation Method.

D) **Cointegration : (8 hrs** – Theory 2 hrs + Lab 6 hrs)
Concept, two variable model, Engle-Granger Method, Vector autoregressions (VAR), Vector error correlation model (VECM).

E) ARCH/GARCH/SV models, some important generalizations like EGARCH & GJR models, ARCH –M models. **(8 hrs** – Theory 2 hrs + Lab 6 hrs)

Suggested Books :

1. The Econometrics of Financial Markets : J. Campbell, A.Lo and C. Mackinlay

2. Econometric Analysis : William H. Greene

Evaluation: Theory: 70% + Practical/Lab: 30%

5. Machine Learning II: (60 hrs)

A) Decision Tree Classification : (6 hrs – Theory 3 hrs + Lab 3 hrs)
Entropy, Gini index, Algorithms, Regression Trees.

B) Probabilistic Classifiers : (6 hrs – Theory 3 hrs + Lab 3 hrs)
Generative and Conditional classifiers.

C) Hyper plane classifiers: (6 hrs – Theory 3 hrs + Lab 3 hrs)
Loss functions, Stochastic gradient algorithms, Perceptron algorithms.

D) Application of to Pattern Recognition Problems. (6 hrs – Theory 3 hrs + Lab 3 hrs)

E) Clustering : (6 hrs – Theory 3 hrs + Lab 3 hrs)
Performance criteria, K-means clustering, EM algorithm
F) Collaborative filtering (6 hrs – Theory 3 hrs + Lab 3 hrs)
G) Combining models (6 hrs – Theory 3 hrs + Lab 3 hrs)
H) Probabilistic graphical models(6 hrs – Theory 3 hrs + Lab 3 hrs)
I) Large Scale Machine Learning : (6 hrs – Theory 3 hrs + Lab 3 hrs)
gradient descent with large data sets
J) Genetic Algorithm. (6 hrs – Theory 3 hrs + Lab 3 hrs)

Suggested Books

1. Machine Learning : Tom Mitchell

Evaluation: Theory: 50% + Practical/Lab: 50%

6. Time Series & Forecasting: (60 hrs)

A) Exploratory Analysis of Time Series. (10 hrs – Theory 4 hrs + Lab 6 hrs)
B) Stationary and Non-Stationary Time Series. (5 hrs – Theory 5 hrs)
C) AR, MA, ARMA, ARIMA models, their properties, estimation of parameters. (20 hrs – Theory 16 hrs + Lab 4 hrs)
D) Tests of Non-Stationarity – Unit Root tests. (5 hrs – Theory 3 hrs + Lab 2 hrs)
E) Forecasting, Smoothing, Minimum MSE Forecast, Forecast Error. (10 hrs – Theory 8 hrs + Lab 2 hrs)
F) Modelling Seasonal Time Series. (5 hrs – Theory 3 hrs + Lab 2 hrs)
G) Missing Data Problem in Time Series. (5 hrs – Theory 3 hrs + Lab 2 hrs)

Suggested Books

1. Introduction to Statistical Time Series : W. A. Fuller
2. Introduction to Time Series Analysis : P. J. Brockwell and R. A. Davis

Evaluation: Theory: 70% + Practical/Lab: 30%

7. Bioinformatics (60 hrs) :

A) Sequence Allignments. (4 hrs – Theory 2 hrs + Lab 2 hrs)
B) Advance Allignment Methods. (4 hrs – Theory 2 hrs + Lab 2 hrs)
C) Gibbs Sampling. (8 hrs – Theory 2 hrs + Lab 6 hrs)
D) Population Genomics. (4 hrs – Theory 2 hrs + Lab 2 hrs)
E) Genetic Mapping. (4 hrs – Theory 2 hrs + Lab 2 hrs)
F) Disease Mapping. (4 hrs – Theory 2 hrs + Lab 2 hrs)
G) Gene Recognition. (4 hrs – Theory 2 hrs + Lab 2 hrs)
H) Transcriptome & Evolution. (4 hrs – Theory 2 hrs + Lab 2 hrs)
I) Protein Structure. (4 hrs – Theory 2 hrs + Lab 2 hrs)
J) Protein Motifs. (4 hrs – Theory 2 hrs + Lab 2 hrs)

K) Hidden Markov Model. (4 hrs – Theory 2 hrs + Lab 2 hrs)
L) Lattice Model. (4 hrs – Theory 2 hrs + Lab 2 hrs)
M) Algorithms. (8 hrs – Theory 6 hrs + Lab 2 hrs)

Suggested Books

1. Introduction to Computational Molecular Biology : C. Setubal & J. Meidanis, PWS Publishing, Boston, 1997

Evaluation: Theory: 50% + Practical/Lab: 50%

Semester-4

Internship based project

A real life project has to be undertaken at an industry for 20 weeks. Each student will have two supervisors: one from academic institution and one from the industry. The project shall involve handling data extensively and use of methodologies learnt during the course work to derive meaningful inferences. A final project report has to be submitted and an "open" presentation has to be made.

Project evaluation may be as follows.

Report from two supervisors: 200 marks (100 each)
Project report: 200 marks
Presentation: 100 marks.
Total: 500 marks