Faculty of Science and Technology

Savitribai Phule Pune University

Maharashtra, India

http://unipune.ac.in

Honours* in Data Science

Board of Studies

(Computer Engineering)

(with effect from A.Y. 2020-21)

http://unipune.ac.in/university_files/syllabi.htm

Savitribai Phule Pune University

Honours* in Artificial Intelligence and Machine Learning

With effect from 2020-21

Year & Semester	Course Code and Course Title	Teaching Scheme Hours / Week			Examination Scheme and Marks						Credit Scheme		
		Theory	Tutorial	Practical	Mid-Semester	End-Semester	Term work	Practical	Presentation	Total Marks	Theory / Tutorial	Practical	Total Credit
TE & V	Computational Statistics	04	--	--	30	70	--	--	--	100	04	--	04
	Computational Programming Laboratory	--	--	02	--	--	50	--	--	50	--	01	01
	Total	04	-	02	100		50	-	-	150	04	01	05
Total Credits =05													
TE & VI	Artificial Intelligence	04	--	--	30	70	--	--	--	100	04	--	04
	Total	04	-	-	100		-	-	-	100	04	-	04
Total Credits =04													
BE & VII	Machine Learning	04	--	--	30	70	--	--	--	100	04	--	04
	Machine Learning Laboratory	--	--	02	--	--	50	--	--	50	--	01	01
	Total	04	-	02	100		50	-	-	150	04	01	05
Total Credits =05													
BE & VIII	Soft Computing and Deep Learning	04	-	--	30	70	--	--	--	100	04	--	04
	Seminar	--	02	--	--	--	-	--	50	50	02	--	02
	Total	04	-	02	100		-	--	50	150	06	-	06
Total Credits =06													
Total Credit for Semester V+VI+VII+VIII = 20													

*** <u>To be offered as Honours for Major Disciplines as–</u>**

1. Computer Engineering
2.Electronics and Telecommunication Engineering
3.Electronics Engineering

For any other Major Disciplines which is not mentioned above, it may be offered as Minor Degree.

Reference: https://www.aicte-india.org/sites/default/files/APH%202020_21.pdf / page 99-100

Savitribai Phule Pune University
Honours* inAI & ML
Third Year of Engineering (Semester V)
Computational Statistics

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Lecture: 04 Hours/Week	04	Mid_Semester(TH): 30 Marks End_Semester(TH): 70 Marks

Companion Course : Computational Statistics Laboratory

Course Objectives:

To introduce several statistical techniques found to be serving as tools even today in the development of machine learning and artificial intelligence based computer algorithms.

- To imbibe strong foundation of statistics in students for implementation in computation.
- To understand modern computational methods used in statistics.
- To get detailed approach of simulation, estimation and visualization of statistical data
- To understand the role of computation as a tool of discovery in data analysis.
- To be able to appropriately apply computational methodologies to real world statistical problems.
- To learn the data processing techniques required to get applied on machine learning algorithms.

Course Outcomes:

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

- **Identify** the suitable method of statistics on the given data to solve the problem of any heuristic approach of prediction.
- **Apply** appropriate statistical concepts and skills to solve problems in both familiar and unfamiliar situations including those in real-life contexts.
- **Design and analyze** real world engineering problems by applying various statistical modeling techniques.
- **Formulate** suitable statistical method required as pre-processing technique for finding the solution of machine learning algorithm.
- **Model and solve** computing problem using correlation, and resampling using appropriate statistics algorithms.

#Exemplar/Case Studies- Elaborated examples/Case Studies are included at the end of each unit to explore how the learned topics apply to real world situations and need to be explored so as to assist students to increase their competencies, inculcating the specific skills, building the knowledge to be applicable in any given situation along with an articulation. One or two sample exemplars or case studies are included for each unit; instructor may extend the same with more. **Exemplar/Case Studies** may be assigned as self-study by students and to be excluded from theory examinations.

Course Contents

Unit I	Introduction to Statistics	(07 Hours)

What is statistics, Statistical Data- Categorical, Numerical (Continuous), Univariate and Bivariate Analysis, Mean, Median, Mode, Standard Deviation, Harmonic Mean, Data Visualization-Line, Scatter, Box plots, Histogram, Statistical Thinking.

#Exemplar/Case Studies	Know about the great statistician- Ronald Fisher

Unit II	Distributions	(9 Hours)

Probability Distributions, Characterizing a Distribution, Discrete Distributions, Normal Distributions, Continuous Distributions Derived from the Normal Distribution, Poisson Distribution, Other

Continuous distributions- Lognormal, Weighbull, Exponential, Uniform.		
#Exemplar/Case Studies	Know about the great statistician and father of Indian statistical institute- Praful Chandra Mahanalobis	
Unit III	**Hypothesis Tests and Statistical Tests**	**(08 Hours)**
Typical Analysis procedures, Hypothesis Concept, Errors, p-Value, and Sample Size, Confusion Matrix, Sensitivity and Specificity, ROC-AUC Curve, Test on Numerical Data- Distribution of a Sample Mean, Comparison of Two Groups, Comparison of Multiple Groups		
#Exemplar/Case Studies	Do watch brief history of Statistics on YouTube https://www.youtube.com/watch?v=J8W37byz_uw	
Unit IV	**Statistical Methods**	**(08 Hours)**
Standard Deviation, Normalization- Feature Scaling, Min-Max scaling, Bias, Variance, Regularization, Ridge Regression, Lasso Regression, Cross Validation Techniques- K-fold, LOOCV, Stratified K-fold, Grid Search CV, CV Error		
#Exemplar/Case Studies	Euclid's Elements	
Unit V	**Statistical Processing**	**(08 Hours)**
Dimensionality Reduction Techniques- Principal Component Analysis, Discriminant Analysis, Feature Selection- Chi2 square method, Variance Threshold, Recursive Feature Elimination, Outliers detection methods, Resampling-Random, under-sampling and over re-sampling		
#Exemplar/Case Studies	Anomalies	
Unit VI	**Statistical Modeling**	**(08 Hours)**
Linear Regression models, Correlation coefficient, Rank Correlation, Residual Error, Mean Square Error, RMSE, Multilinear Regression, Polynomial Features, Gradient Descent, Logistic Regression, Bayesian Statistics, Bayes' Theorem, Monte Carlo Method		
#Exemplar/Case Studies	Biography of Thomas Bayes	

Learning Resources

Text Books:

- Thomas Haslwanter, "An Introduction to Statistics with Python with Applications in the Life Sciences", Springer International Publishing Switzerland 2016, ISBN 978-3-319-28315-9, ISBN 978-3-319-28316-6 (eBook)
- Allen B. Downey, "Think Stats", Second Edition, O'Reilly Media, ISBN: 978-1-491-90733-7

Reference Books:

- Thomas Haslwanter, "An Introduction to Statistics with Python with Applications in the Life Sciences", Springer International Publishing Switzerland 2016, ISBN 978-3-319-28315-9, ISBN 978-3-319-28316-6 (eBook)
- Peter Bruce and Andrew Bruce, "Practical Statistics for Data Scientists", First Edition, O'Reilly Media, ISBN-978-1-491-95296-2
- Allen B. Downey, "Think Stats", Second Edition, O'Reilly Media, ISBN: 978-1-491-90733-7
- José Unpingco, "Python for Probability, Statistics, and Machine Learning", Springer International Publishing Switzerland, ISBN 978-3-319-30715-2, DOI 10.1007/978-3-319-30717-6, ISBN 978-3-319-30717-6 (eBook)
- Claus Weihs, Olaf Mersmann, Uwe Ligges, "Foundations of Statistical Algorithms", CRC Press, ISBN-978-1-4398-7887-3 (eBook - PDF)

e-Books:

- http://file.allitebooks.com/20151204/Foundations%20of%20Statistical%20Algorithms.pdf
- http://onlinestatbook.com/Online_Statistics_Education.pdf
- https://upload.wikimedia.org/wikipedia/commons/8/82/Statistics.pdf
- http://cnx.org/content/col10522/1.38/pdf

- http://www.greenteapress.com/thinkstats/thinkstats.pdf

MOOC/ Video Lectures available at:

- https://www.udemy.com/course/introduction-to-bayesian-statistics/ (Free Course)
- https://www.youtube.com/watch?v=xxpc-HPKN28
- https://www.udacity.com/course/intro-to-statistics--st101# (Free Course)
- https://nptel.ac.in/courses/111/105/111105090/
- https://nptel.ac.in/courses/111/105/111105077/

Websites Resources:

- https://analyticsvidhya.com
- https://towardsdatascience.com
- https://medium.com
- https://stackabuse.com
- https://machinelearningmastery.com

Savitribai Phule Pune University
Third Year of Engineering (Semester V)
Computational Statistics Laboratory

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Practical: 2 Hours/Week	01	Term work: 50 Marks

Guidelines for Laboratory Conduction

- **Lab Assignments:** Following is list of suggested laboratory assignments for reference. Laboratory Instructors may design suitable set of assignments for respective course at their level. Beyond curriculum assignments and mini-project may be included as a part of laboratory work. The instructor may set multiple sets of assignments and distribute among batches of students. It isappreciated if the assignments are based on real world problems/applications. The Inclusion of few optional assignments that are intricate and/or beyond the scope of curriculum will surely be the value addition for the students and it will satisfy the intellectuals within the group of the learners and will add to the perspective of the learners. For each laboratory assignment, it is essential for students to draw/write/generate flowchart, algorithm, test cases, mathematical model, Test data set and comparative/complexity analysis (as applicable). Batch size for practical and tutorial may be as per guidelines of authority.
- **Term Work**–Term work is continuous assessment that evaluates a student's progress throughout the semester. Term work assessment criteria specify the standards that must be met and the evidence that will be gathered to demonstrate the achievement of course outcomes. Categorical assessment criteria for the term work should establish unambiguous standards of achievement for each course outcome. They should describe what the learner is expected to perform in the laboratories or on the fields to show that the course outcomes have been achieved. It is recommended to conduct internal monthly practical examination as part of continuous assessment.
- **Assessment:**Students' work will be evaluated typically based on the criteria like attentiveness, proficiency in execution of the task, regularity, punctuality, use of referencing, accuracy of language, use of supporting evidence in drawing conclusions, quality of critical thinking and similar performance measuring criteria.
- **Laboratory Journal-** Program codes with sample output of all performed assignments are to be submitted as softcopy. Use of DVD or similar media containing students programs maintained by Laboratory In-charge is highly encouraged. For reference one or two journals may be maintained with program prints in the Laboratory. As a conscious effort and little contribution towards Green IT and environment awareness, attaching printed papers as part

of write-ups and program listing to journal may be avoided. Submission of journal/ term work in the form of softcopy is desirable and appreciated.

Suggested list of assignments
(Use suitable programming language/Tool for implementation)

Sr. No	Assignment statement
1	Compute Estimators of the main statistical measures like Mean, Variance, Standard Deviation, Covariance, Correlation and Standard error with respect to any example. Display graphically the distribution of samples.
2	Plot the Normal Distribution for class test result of a particular subject. Identify the Skewness and Kurtosis
3	Load the dataset: birthwt Risk Factors Associated with Low Infant Birth Weight at https://raw.github.com/neurospin/pystatsml/master/datasets/birthwt.csv 1. Test the association of mother's (bwt) age and birth weight using the correlation test and linear regeression. 2. Test the association of mother's weight (lwt) and birth weight using the correlation testand linear regeression. 3. Produce two scatter plot of: (i) age by birth weight; (ii) mother's weight by birth weight. Elaborate the Conclusion ?
4	Apply Basic PCA on the iris dataset. The data set is available at: https://raw.github.com/neurospin/pystatsml/master/datasets/iris.csv • Describe the data set. Should the dataset been standardized? • Describe the structure of correlations among variables. • Compute a PCA with the maximum number of components .• Compute the cumulative explained variance ratio. Determine the number of components$K$by your computed values. • Print the$K$principal components directions and correlations of the$K$principal components with the original variables. Interpret the contribution of the original variables into the PC. • Plot the samples projected into the$K$first PCs. • Color samples by their species.
5	Perform clustering of the iris dataset based on all variables using Gaussian mixture models. Use PCA to visualize clusters.