SEVEN DAYS SKILL DEVELOPMENT

BIOINFORMATICS WORKSHOP ON

"MACHINE LEARNING APPLICATION IN BIOLOGICAL DATA ANALYSIS: LINEAR REGRESSION & K-MEANS CLUSTERING"

OCTOBER 15-21, 2019

SPONSORED BY

COUNCIL OF SCIENTIFIC & INDUSTRIAL RESEARCH, NEW DELHI (GOVT. OF INDIA)

ORGANIZED BY

CSIR-CENTRAL INSTITUTE OF MEDICINAL & AROMATIC PLANTS, LUCKNOW

Workshop Contents:

- *Fundamentals of Machine Learning*
- *Types of Machine Learning*
- *Supervised Learning – Regression*
- *Linear Regression Modeling*
- *Simple & Multiple Linear Regression*
- *Independent Variables & Dependent Variables*
- *Variables Collinearity & Correlation Matrix*
- *Model Development & Equation Derivatization*
- *Residuals/Sum of Square Error Significance*
- *Unsupervised Learning - k-Means Clustering*
- *Hands-on practicals in Spreadsheet, R & Python language*

Convenor
Dr. Feroz Khan

Coordinator
Dr. Laiq-ur Rahman

Chairman
Dr. Abdul Samad
Director, CSIR-CIMAP

About CSIR-CIMAP, Lucknow

CSIR-Central Institute of Medicinal & Aromatic Plants (CSIR-CIMAP) is a premiear multidisciplinary research institute of Council of Scientific & Industrial Research (CSIR), New Delhi, India with its major focus on exploiting the potential of medicinal and aromatic plants (MAPs) by cultivation, bioprospection, chemical characterization, extraction, and formulation of bioactive phytomolecules. With a strenght of 100 scientists, 162 technical officers, 129 support staff and nearly 300 doctoral and post-doc scholars at its head-quarter in Lucknow and research centers at Bengalure, Hyderabad, Pantnagar, and Purara. CSIR-CIMAP has played a key role in positioning India as a global leader in production of mints, vetiver and other aromatic grasses, and in ensuring indigenous production of artemisinin – a WHO approved antimalarial. CSIR-CIMAP houses a National Gene Bank on MAPs, which is one of the three of its kind in India. CSIR-CIMAP has played a key role in successfully commercializing an ayurvedic herbs based antidiabetic formulation, which has now benefitted millions. The institute is presently accredited by ICS-UNIDO and Indian-Ocean Rim Association (IORA) as a focal point for research and training on Medicinal Plants among 21 participating member countries. For more details please see the CSIR-CIMAP website www.cimap.res.in

About the Workshop

Artificial Intelligence (AI) has become widespread recently especially in Biological data modeling, pattern analysis and classification. Machine learning (ML) is a subset of AI, which focuses mainly on computer learning from its experience and making predictions based on it, without being programmed. ML is so universal today that you probably use it dozens of times a day without knowing it. Researchers think we are moving towards human-level AI. The IBM reported that 2.5 million terabytes data is being generated every day. In the last few years, enormous progress made in data generation e.g., a single genome sequencing generate multiple hundred gigabytes in size. Therefore, a specific set of technical skills required to deal with such a large biological data. That's where Bioinformatics comes in. Commonly ML algorithms can be divided into following categories based on their purposes, namely supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In the workshop major focus will be on supervised learning algorithms, where computer is trained with labeled data (a set of training examples or with input and output value). In this, algorithm try to model relationships and dependencies between the target prediction output and the input features such that to predict the output values for new data. The main types of supervised learning problems include regression and classification problems. In this, common algorithms are Linear Regression, Logistical Regression, Random Forest, Gradient Boosted Trees, Support Vector Machine (SVM), Neural Networks (NN), Decision Trees, Naive Bayes, and Nearest Neighbor. In this workshop series, we will first focus on Linear Regression with brief introduction to other methods. Subsequently, we will understand unsupervised learning algorithm, where computer is trained with unlabeled data. In this, algorithms are mainly used in pattern detection and descriptive modeling. There are no output categories or labels, so these algorithms try to use techniques on the input data to mine for rules, detect patterns, and summarize or group data points. The main types of common algorithms are Clustering algorithms, t-SNE (t-Distributed Stochastic Neighbor Embedding), PCA (Principal Component Analysis), and Association rule learning algorithms. In this workshop, we will learn in detail about k-means clustering and brief introduction to others by using Biological/Biomedical data points examples. Bioinformatics is one of the application of ML. Bioinformatics is the interdisciplinary science of interpreting biological data using information technology and computer science algorithms. In biological sequence analysis and classification, we normally used ML algorithms e.g., genome sequencing, gene finding, genome annotation, sequence comparison, transcriptome analysis, microarray analysis, regulatory sequence analysis, computation proteomics such as electrophoresis analysis, and protein identification through mass spectrometry.

The Aim of Workshop

To familiarize researchers/academicians/students with the basics & application of Machine Learning algorithms (Linear Regression & k-Means Clustering) used in the laboratory experimental biological/biochemical data analysis and interpretation. Also explain the underlying principles, mathematical functions, data mining algorithms with suitable easy to understand examples. In parallel, hands-on practical exercises/examples for technical skill development will be scheduled after lectures. The workshop covers first lectures by data mining experts, example demo presentations, and hands-on practical examples/exercises. The workshop would cover the following aspects:

- Biological/Biochemical data, resources, & types
- Introduction to Machine Learning
- Types of Machine Learning
- Supervised Learning - Linear Regression
- Simple Linear Regression analysis on Biological data points
- Multiple Linear Regression analysis on Biological data points
- Features selection, Multiple Variable Collinearity calculation & visualization
- Regression line fitting – Equation generation & visualization
- Model Validation- Residuals/Sum of square Error, Significance/fitness plot
- Clustering Introduction
- Data classification with k-Means Clustering (Unsupervised Machine Learning)
- Example – k- Means Clustering & data classification
- Practicals exercises/Hands-on experience in Spreadsheet, R and Python language.

The participants may access online bioinformatics resources/literatures/reviews related to machine learning applications in biological and biochemical data analysis. The hands-on skill development training gained may help in studying big data analysis (labeled or unlabeled) and entrepreneurship in biological/biomedical/clinical data mining research domains/services.

Eligibility

Graduates/PG/Ph.D. fellows/Post-Doc scholars/ Project fellows/ Scientists/Technical Officers/ Company Professionals/Entrepreneurs/Academicians can attend.

Certification

Workshop participants will receive a certificate of participation from the CSIR-CIMAP, Lucknow after successful completion of the workshop. The digital certificate will also be emailed after the successful completion of workshop.

Feedback

After workshop, participants may be asked to submit the given feedback form. Participants may be asked share their workshop experiences and further improvement points verbally during valedictory session.

Technical requirements

Participants may carry Laptop/Tablets (optional). Windows OS platform, MS Word, MS Excel spreadsheet analysis software, R and Python languages will be used for analyses/hands-on examples. A working version of Windows, MS Office software is necessary to follow these examples. No coding or statistics background is required to participate in the workshop, however basic cell biology and biostatistics will be beneficial.

Accommodation

Accommodation will be provided to the interested participants on twin shared basis. Preference will be given to external participants.

Registration Fee

Rs.12,000/- per participant (without accommodation) and **Rs.15,000**/- per participant (with accommodation, twin shared basis). The registration fee includes Registration Kit (Brochure, Program schedule, & Stationary items), a CD/DVD (Lecture presentations/Tutorials, Practicals/Examples/Exercises, & Group photograph), Feedback form, Morning High Tea, Lunch, Evening Tea, and Dinner. On request, vehicle will be arranged for local site seeing/shopping.

Registration fee can be pay through online mode to SBI bank A/c No. 00000030267691783, SBI Main branch, Hazratganj, Lucknow (IFSC code: SBIN0000125) or through Demand Draft in favor of '**Director, CIMAP**', payable to Lucknow. Complete registration form along with the fee details should reach us on or before deadline i.e., October 14, 2019 upto 8:00 PM. Registration to the skill development workshop will be on 'First-come-First-serve' basis.

For any query, kindly contact Dr. Feroz Khan (f.khan@cimap.res.in) or Dr. Laiq-Ur Rahman (l.rahman@cimap.res.in).

For further details please contact:

Director

CSIR-Central Institute of Medicinal & Aromatic Plants,

P.O.-CIMAP, Kukrail Picnic Spot Road, Lucknow-226015, India

Ph.: +91 522 2718639, 2718641, 2718505

E-mail: director@cimap.res.in , Website: www.cimap.res.in

सी एस आई आर–केन्द्रीय औषधीय एवं सगंध पौधा संस्थान
कुकरैल पिकनिक स्पाट रोड, लखनऊ–226015, भारत
फोन नः +91 - 522 - 2359623, फैक्सः +91 - 522 - 2342666
टेलेफैक्सः +91 - 522 - 2357136
director@cimap.res.in

CSIR-CIMAP

CSIR - Central Institute of Medicinal and Aromatics Plants
Kukrail Picnic Spot Road, Lucknow - 226 015, India
Phone: +91 - 522 - 2359623; Fax: +91 - 522 - 2342666
Telefax: +91 - 522 - 2357136
director@cimap.res.in

Registration Form

Full Name: _____

Designation/Position: _____

Affiliation/Institute/Univ.:_____

Address: _____

_____

Area of Interest/Expertise: _____

E-mail: _____

Mobile No.: _____

Payment Details:

Registration Fee Amount: Rs._____

Mode of payment (Online/DD): _____

Online Transaction/DD No._____ Date _____

Bank Name: _____

Signature