**Machine Learning for Biomedical Data**

**SYLLABUS SYSM-578 FALL-2022 (Sept – Dec)**

**Mondays (3-Credits)**

|  |
| --- |
| This course will cover practical and conceptual aspects of machine learning in application to high-throughput biomedical data using Python. Throughout the course, students will get an understanding of opportunities and limitations of machine learning in the context of pre-clinical and clinical research. The course is designed as a combination of online resources, practical assignments and live sessions that will be conducted online. Throughout the course, we will review several project examples that demonstrate successes and limitations of conventional machine learning (ML) methods and deep learning (DL) using data from public repositories. As a result of completing this course, each participant should be able to differentiate between various methods, apply the correct method to a data or problem statement and develop a completed project using ML or DL. |

**Course Director**: Pine Biotech

**Instructors:**

**Elia Brodsky**

CEO of Pine Biotech (<http://pine-biotech.com/pine-biotech-multi-omics/>)

Extensive experience in bioinformatics, as well as entrepreneurship and business development with life sciences. Educated in various fields such as bioinformatics, mathematics, business analytics, and finance. Extensive hands-on involvement and collaborations with Life Science Companies and Research Teams in the domain of *data science*. Project mentor for student research projects.

**Course Start date**: August 30, 2022.

**Readings**

The readings will be in the form of relevant publications that will be provided. All additional content, including coding practice for learning will be available through the Omics Logic portal: [learn.omicslogic.com](http://learn.omicslogic.com/)

1. OmicsLogic Learn portal: <https://learn.omicslogic.com>
2. T-BioInfo server for Big Data Analysis: <https://server.t-bio.info>
3. Recommended reading: Deep Learning in Omics Data Analysis and Precision Medicine (book - <https://www.ncbi.nlm.nih.gov/books/NBK550335/>)
4. Overview of Machine Learning Part 1: Fundamentals and Classic Approaches <https://www.sciencedirect.com/science/article/pii/S1052514920300629?via%3Dihub>

**Publications for Literature review:**

“Modeling precision treatment of breast cancer”: an analysis of over 70 different Breast Cancer cell lines and over 90 different therapeutic agents, Daemen et al., 2013, <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-10-r110>

“Whole transcriptome profiling of patient-derived xenograft models as a tool to identify both tumor and stromal specific biomarkers,” James R. Bradford et. al.; DOI: 10.18632/oncotarget.8014. (<https://www.oncotarget.com/article/8014/text/>)

“Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin”, Stuart et. al. 2014, <http://www.cell.com/cell/fulltext/S0092-8674(14)00876-9>

**Enquiries**

*Informal enquiries* about the course can be made to:

Dr. Sona Vasudevan,

Department of Biochemistry and Molecular Biology, Georgetown University School of Medicine, 3300, Whitehaven Street, NW, Washington, DC 20007. Phone: (202)-687-2242. Fax: (202)-687-0057. E-mail: [*sv67@georgetown.edu*](mailto:sv67@georgetown.edu)*.*

**Contact Information/Office hours:**

**Sona Vasudevan, Ph.D.**

3300, Whitehaven Street,

Georgetown University School of Medicine

Harris Building, Suite 1200

Phone: 202-687-2242

E-Mail: [sv67@georgetown.edu](mailto:sv67@georgetown.edu)

By appt only.

**SYLLABUS SYSM-578 FALL 2021 (Sept – Dec)**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Lecture Topic** | **Date** | **Readings** |
| 1 | Introduction to the course | 08/30 | <https://learn.omicslogic.com> |
| 2 | Data Processing and Exploratory Analysis | 09/06 | <https://learn.omicslogic.com/courses/course/course-7-bioml-machine-learning-for-biomedical-data> |
| 3 | Machine Learning Methods: unsupervised and supervised types of the analysis | 09/13 | <https://learn.omicslogic.com/courses/course/python-course-2-introduction-to-data-science-bioml> |
| 4 | Dimensionality Reduction: Ordination and Embedding | 09/25 | Lesson 4 |
| 5 | Unsupervised Learning: Clustering | 10/04 | Lesson 5 |
| 6 | Supervised Learning: Discriminant Analysis and Classification | 10/11 | Lesson 6 |
| 7 | Feature selection | 10/18 | Lesson 6 |
| 8 | Regression | 10/25 | https://learn.omicslogic.com/Learn/course-5-transcriptomics/lesson/10-t2-regression-and-factor-regression-analysis |
| 9 | Generalized Linear Models | 11/1 | Practical notebook and slides will be provided |
| 10 | Network analysis | 11/22 | Practical notebook and slides will be provided |
| 11 | Deep Learning | 11/29 | Lesson 9 |
| 12 | Model Accuracy and Validation | 12/06 | Lesson 6 |
| 13 | Project Examples and Case Studies | 12/08 | <https://learn.omicslogic.com/courses> |
| 14 | How to design your project | 12/10 | <https://learn.omicslogic.com/courses/course/course-9-designing-a-bioinformatics-research-project> |
| 15 | **Final Project Submissions and Exam** | 12/13 | <https://learn.omicslogic.com/projects> |

**Course Grading Information:**

To evaluate the program for grading (participation in interactive or completion of all recorded sessions, activity monitoring and use of surveys, quiz and exam results).

|  |  |  |  |
| --- | --- | --- | --- |
| Pre-assessment:  Literature review: 3%  Activity and Quiz Results: 5%  Self-assessment: 2% | 10% | Participation  Session Participation: 10%  Coursework Completion: 10% | 20% |
| Mid-assessment: 20%  Literature review: 5%  Activity and Quiz Results: 10%  Self-assessment: 5% | 20% | Final Exam:  Unsupervised ML: 10%  Supervised ML: 10%  Case Study: 10% | 20% |
| Post-assessment: 20%  Literature review: 5%  Activity and Quiz Results: 10%  Self-assessment: 5% | 20% | Project submission:  Topic: 3%  Dataset: 5%  Analysis plan: 2% | 10% |

Literature Review: Literature review of case study publications (3 in total) will be expected to be submitted as a 1-page written essay and a short 5-question quiz.

Activity points: expectation of 10,000 points or more achieved on the learn.omicslogic.com portal

Quiz results: minimum 80% for passing (quizzes are included in the courses online)

Participation: attendance of ZOOM sessions (must have a name in the session and have video on)

Coursework: theoretical and practical tutorials on learn.omicslogic.com. Additional resources will be shared via jupyter notebooks (or Google Colaboratory / colab)

Final Exam: 10 questions for each type of data (30 question total, all multiple choice)

Project: At the end of the course, you will be expected to present your own project proposal, outlining in technical terms the way you will use machine learning approaches to address a real-world problem in biomedical research or precision medicine. The minimal requirement is to submit a project proposal form, where you describe the topic, dataset, approaches and references for your project. The expected outcome is to share 5-20 slides and a workbook (jupyter notebook) where the analysis is performed. Project form should be submitted no later than 1 week before the course ends. Can be filled out at any point during the course and modified along the way. Only 1 final (latest) submission will be evaluated. All project submission has to be done via the learn.omicslogic.com portal. To learn more about the project, follow this link: <https://learn.omicslogic.com/courses/course/course-9-designing-a-bioinformatics-research-project>

**Course Topics and Outcomes:**

## **In-depth review of statistical concepts, ML and Deep Learning:**

* Review statistical approaches: hypothesis testing, ANOVA and Regression
* Exploratory analysis and data visualization
* NMDS, PCA, t-SNE, UMAP
* Clustering (hierarchical, k-means, DBscan, Fuzzy, PAM)
* Classification: Decision Trees, Random Forest, Support Vector Machine, Naive Bayes
* AutoML, Model evaluation and Validation
* Feature selection strategies (feature significance and greedy methods)
* Deep Learning: Encoders, Perceptron, Neural Networks

## **What will you learn to do in Python? How much Python background should you already have?**

The course is suitable for anyone without coding background or without python background, however you will be expected to learn python syntax, read code provided to you in an online environment and assemble working code in a jupyter notebook. Introductory material on getting started will be made available on th associated online resource for the course: learn.omicslogic.com

### Coding challenges:

* Loading data from csv, txt or xlsx sources and converting it to various data structures (*dataframe, matrix, lists* and *vectors*)
* Summarizing **categorical** and **continuous** datasets
* Data preparation using **log-normal transformation** and **quantile normalization**
* Statistical tests and outputs (*p-value, t-value, standard error, FDR, logFC*)
* Popular packages like **pandas, numpy,** and **sklearn**
* Visualization using **matplotlib** and **seaborn**
* Reading, understanding and loading code examples

**General Coding and Data sharing practices:**

* Organizing your scripts with comments and functions
* Setting up a development environment (IDE)
* Dealing with errors and troubleshooting code (debugging)
* Preparing data summaries and submitting curated data and meta-data tables to sharing repositories (**FAIR principles**)
* Sharing your analysis in **jupyter notebooks**, on **github** or **google colab**
* Creating interactive visualization in **plot.ly**

**Requirements:**

The course is available for those who are just getting started and does not require in-depth knowledge of programming or machine learning. Some background in the basics of molecular biology, preferred introduction to bioinformatics.

## **Intended Outcomes and learning Objectives:**

* Understanding of analytical methods for processing, visualization and analysis of complex biomedical data
* Learning terminology for machine learning and artificial intelligence in biomedical discovery
* Becoming familiar with project examples where ML was used effectively to achieve meaningful results
* Hands-on practice in application of standard unsupervised and supervised learning methods to various types of data, such as genomic, transcriptomic, metagenomics, imaging and clinical
* Understand the ML taxonomy and the commonly used machine learning algorithms for analyzing “omics” data
* Understand differences between ML algorithms categories and to which kind of problem they can be applied to
* Understand different applications of ML in application to different -omics studies and project design objectives
* Use popular Python packages for data visualization, analysis and ML
* Interpret and visualize the results obtained from ML analyses on omics datasets
* Apply the ML techniques to analyze public domain or their own datasets