Program Description

The application of machine learning continues to grow rapidly in a variety of industrial settings. In pharmaceutical development in particular, Machine Learning methodologies have been adopted to accelerate process optimization and material characterization, a trend that is anticipated to rise with the increased rate of data digitization. The goal of the program is to develop the skill set necessary to evaluate and apply the growing portfolio of algorithms from the open source ecosystem and deploy them in the appropriate context. In addition to covering a multitude of Machine Learning algorithms (generalized linear regression, non-linear regression, regularization methods, random-forest, neural networks, Markov processes, etc.), the program will emphasize the generation of contextualized visualization tools to adequately present results to a wider non-expert audience as well as health authorities, a key aspect of the successful implementation of Machine Learning in pharmaceutical development.

Required Courses

<table>
<thead>
<tr>
<th>Course #</th>
<th>Course name</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA 541</td>
<td>Statistical Methods</td>
</tr>
<tr>
<td>MA 576</td>
<td>Optimization Models in Data Science</td>
</tr>
<tr>
<td>ChE/MT 542</td>
<td>Data Science in Pharmaceutical Development</td>
</tr>
<tr>
<td>ChE/MT 543</td>
<td>Machine Learning in Pharmaceutical Development</td>
</tr>
</tbody>
</table>

All courses available on campus and online

Program Objectives

Graduates will be able to:
1. Identify data streams typically found in pharmaceutical organizations (but can be extended to other industries) that will benefit from the adoption of Machine Learning and the corresponding dissemination of results to improve and streamline the quality of decision making;
2. Organize teams to develop machine learning algorithms targeted to specific workflows aimed at increasing access and adoption of sophisticated analysis;
3. Diagnose existing Data Science workflows for potential improvements in reliability, efficiency and speed;
4. Explore and evaluate the adoption of state of the art machine learning models to address pharmaceutical development problems in the organization or improve on existing tools.

Who Can Participate in this Program?

Students in Chemical, Mechanical, Industrial, Material Science, Civil and Environmental Engineering programs. Also, students in Physics, and Chemistry. Graduates can pursue careers in Pharmaceutical or related industries (Chemical, Food, etc.).

Program Outcomes

Students will be able to:
1. Understand and explain the underlying concepts powering relevant Machine Learning algorithms that will be covered throughout the program.
2. Understand and explain the applicability of the most frequently used loss functions for model fitting problems.
3. Identify appropriate Machine Learning algorithms to solve problems relevant to pharmaceutical development and rigorously evaluate the available data sets for applicability by applying tools in data visualization analysis and feature engineering.
4. Relate between sequential decision problems and decision trees, Bayesian networks and other Bayesian models arising in statistical learning and incorporate the model uncertainty in the underlying decision.
5. Apply reproducible research and version control to efficiently manage Machine Learning projects in a collaborative environment in situations and organizations typically found in industry.
**Instructors**

**Dr. Zhong Xiao**, Lecturer  
Department: Mathematical Sciences; Office: Kidde 221  
Phone: 201.216.xxxx; Email: Xiao.Zhong@stevens.edu  
Xiao Sophia Zhong holds a Ph.D. in Computer Science and Technology from Zhejiang University, China, and a Ph.D. in Mathematics with emphasis on Statistics from Missouri S&T. She worked at AstraZeneca R&D Boston, Whitehead Institute of MIT, and Tsinghua University, China. Her research focuses on artificial intelligence, pattern recognition, data mining, financial engineering, statistical applications, and simulations.

**Prof. Darinka Dentcheva**  
Department: Mathematical Sciences; Office: Peirce 302  
Phone: 201.216.8640; Email: ddentche@stevens.edu  
Darinka Dentcheva holds a PhD and a Doctor of Sciences (Habilitation) degree from Humboldt University Berlin, Germany. Her current research interests are in optimization under uncertainty and risk. She is passionate about education and has participated in the development of new graduate curricula and courses. Darinka Dentcheva is Associate Editor of several scientific journals, a member of international scientific bodies, and a recipient of multiple research awards and recognitions.

**Dr. Jose Tabora**  
Company: Bristol-Myers Squibb; Email: jose.tabora@bms.com  
Dr. Tabora interviewed by AIChE: [https://www.youtube.com/watch?v=UBv0xsGim-o&feature=youtu.be](https://www.youtube.com/watch?v=UBv0xsGim-o&feature=youtu.be)  
Tabora holds a Ph. D. from the University of Virginia. For the past 25 years he has worked for Merck, Eli Lilly and BMS in Pharmaceutical product Development. Through his professional life he has championed the adoption of machine learning and data science in various workflows of pharmaceutical process development. He is a Fellow of the AIChE.

**Dr. Jacob Albrecht**  
Company: Bristol-Myers Squibb; Email: jacob.albrecht@bms.com  
Jacob Albrecht is a Principal Scientist in Product Development at Bristol-Myers Squibb, with a Ph.D. in chemical engineering from MIT. With over a decade of pharmaceutical development experience, he has championed the adoption of modern approaches to data science and machine learning both within BMS and through pharmaceutical industry consortia.

**Program Contact Information**: Professor Adeniyi Lawal  
Office: 100A McLean Main Campus  
Phone: 201.216.8241; Email: alawal@stevens.edu
Who Can Take this Course?

Students in Chemical, Mechanical, Industrial, Material Science, Civil and Environmental Engineering programs. Also, students in Physics, and Chemistry. Graduates can pursue careers in Pharmaceutical or related industries (Chemical, Food, etc.).

Course Objectives

The objective of this course is to introduce the basic, classical, and comprehensive statistical concepts, theories, approaches, and models to the students. After taking this course, the students are expected to have thorough understanding about the important and popular statistical methods which have various applications in machine learning and data science. They will be trained in the different phases of the professional statistician’s work, namely: data collection, description, analysis, testing, and presentation of the conclusions.

Course Outcomes

The students will develop their ability to:

1. Understand deeply the basic and classical statistical concepts, theories, methodologies, inferences, and models which are important in data science and machine learning
2. Be familiar with the complete data analyzing process commonly employed in academia and industry
3. Be familiar with the statistical skills applied to each stage of data analysis procedure, including mathematical transformation of scientific statement, data collection, data description, data preprocessing, model selection, output interpretation, conclusion and decision making
4. Be aware of the similarities and differences among various statistical models, estimation and prediction methods, commonly used and advanced approaches to data analysis
5. Be able to choose the appropriate statistical tools to understand and explore different types of data arising from engineering and sciences
6. Comprehend the assumptions and restrictions made on classical statistical methodologies, and understand the importance of developing new technologies in machine learning and data science
Instructor

Xiao Sophia Zhong, Lecturer at Department of Mathematical Sciences of Stevens Institute of Technology. She was a Visiting Assistant Professor at GSOM of Clark University and COES of Louisiana Tech University. She also worked at AstraZeneca R&D Boston, Whitehead Institute of MIT, and Tsinghua University, China. Her research focuses on artificial intelligence, pattern recognition, data mining, financial engineering, statistical applications, and simulations. She holds a Ph.D. degree in Computer Science and Technology from Zhejiang University, China, and a Ph.D. degree in Mathematics with emphasis on Statistics from Missouri S&T.

Topics

1. Population, samples, stratified sampling, cumulative distribution function, density curves, measures of location and dispersion
2. Parameter estimation, method of moments, method of maximum likelihood
3. Bayesian approach to parameter estimation, efficiency, Cramer-Rao lower bound, sufficiency
4. Neyman-Pearson paradigm, significance level, p-value, uniformly most powerful tests, confidence interval
5. Generalized likelihood ratio tests, likelihood ratio tests for multinomial distribution
6. Poisson dispersion test, tests for normality
7. Comparing two independent samples, comparing paired samples, nonparametric methods
8. Experimental design, one-way analysis of variance, two-way analysis of variance
9. Analysis of categorical data, relationships with scatterplots, logistic regression
10. Statistical properties of least squares estimates, multiple linear regression
11. Model selection, LASSO, ridge regression
12. Basis functions, regression splines
13. PCA
14. Resampling methods, Bootstrap

Pre-requisite:
MA540

Program:
M.Sc. Data Science in Pharmaceutical Development, Applied Mathematics

Available on campus and online

Credit Applied towards Graduate Degree or Certificate

Contact Information

Dr. Xiao Sophia Zhang
Office: Kidde 222
Main Campus
Department: Mathematical Sciences
Email: xzhang6@stevens.edu
Course Description
This course will introduce the students to basic models and methods in machine learning and, more generally, in data-driven optimization and sequential decision making. The course will provide introduction to stochastic subgradient calculus for non-smooth convex functions, foundation of regularization in optimization problems and its role in numerical methods of optimization. Typical models of statistical estimation, such as various regression models, classification, and other problems in data science are analyzed as sample average approximation models. The second portion of the class discusses sequential decision models focusing specifically on their role in the process of statistical learning. Both Markov and Non-Markov finite-horizon decision problems will be introduced. Applications will cover decision trees, hierarchical clustering, Bayesian classifiers, and others as time permits. Some attention will be placed on bias- and variance reduction techniques.

Course Objectives
The objective of this course is to provide an overview of basic optimization models arising in data science and machine learning. The students will learn the associated mathematical questions and challenges arising in this context and will be made aware of the existing mathematical methodology that helps address those challenges.

Who Can Take this Course?
Students in Chemical, Mechanical, Industrial, Material Science, Civil and Environmental Engineering programs. Also, students in Physics, and Chemistry. Graduates can pursue careers in Pharmaceutical or related industries (Chemical, Food, etc.).

Course Outcomes
The students will develop their ability to:
1. Calculate subgradients of non-smooth convex function and specifically of the most frequently occurring loss functions in statistical model fitting problems. This enables application of stochastic subgradient methods for specific applications.
2. Use regularization techniques in optimization problems and devise numerical methods using regularization.
3. Formulate a sample-based optimization problem and be aware of the necessary probabilistic guarantees for its solution.
4. Formulate finite-horizon Markov decision problems for fully observable problems, formulate the corresponding dynamic programming equations and solve the problem using iteration methods.
5. Formulate finite-horizon Markov decision problems for partially observable problems, describe and properly update the belief states and solve the problem using iteration methods.
6. Formulate multi-stage stochastic optimization problems.
7. Relate between sequential decision problems and decision trees, Bayesian networks and other Bayesian models arising in statistical learning.
Instructor

Dr. Darinka Dentcheva holds a PhD and a Doctor of Sciences (Habilitation) degree from Humboldt University Berlin, Germany. Her current research interests are in optimization under uncertainty and risk. She is passionate about education and has participated in the development of new graduate curricula and courses. Darinka Dentcheva is Associate Editor of several scientific journals, a member of international scientific bodies, and a recipient of multiple research awards and recognitions.

Topics

1. Subgradient calculus for non-smooth functions
2. Review of optimality conditions
3. Moreau-Yoshida regularization; the proximal point and the bundle method
4. Properties of expectation functionals; Static stochastic optimization problems in DS
5. Stochastic subgradient methods and other methods of unconstrained optimization
6. Sample average approximation (SAA) and asymptotic analysis
7. Markov decision problems introduction
8. Dynamic programming equations for finite horizon problems
9. Hierarchical Clustering and other applications
10. Partially observable systems; Belief states
11. Hidden Markov chains; Bayes operator
12. Non-Markov sequential decision problems
13. Boosting, bagging and ensemble learning

The project will encompass elements of SAA and sequential problems in a context of statistical learning with application of some of the variance reduction techniques.
# Course Description

The increased availability of digital data in the pharmaceutical industry has enabled the adoption of machine learning models to advance pharmaceutical development efforts. In addition, the emergence of open source languages with extensive libraries of powerful algorithms has transformed the way in which data science is adopted and practiced in pharmaceutical development organizations. This class provides the students with an introduction to pharmaceutical development aimed at contextualizing the incorporation of data science methodologies acquired in mathematical foundation courses (see requirements below). Industrial case studies in the public domain will be used as practice examples to demonstrate the incorporation of data science principles to industrially relevant applications.

# Course Objectives

1. Students will develop an understanding of comprehensive pharmaceutical development, specifically research workflows that are data intensive for which application of data science is particularly informative;  
2. Students will acquire and be able to demonstrate the requisite skills on how to incorporate data science algorithms into the optimization and characterization of unit operations contextualized by mechanistic relationships;  
3. Students will be able to create and effectively describe tools designed to communicate results and inferences from data science analysis to non-data science experts to enable decision making.

# Who Can Take this Course?

Students in Chemical, Mechanical, Industrial, Material Science, Civil and Environmental Engineering programs. Also, students in Physics, and Chemistry. Graduates can pursue careers in Pharmaceutical or related industries (Chemical, Food, etc.).

# Course Outcomes

Students will have a detailed understanding of the overall pharmaceutical development space (active synthesis and formulation). They will be able to:  
1. Apply data exploration analysis including dimension reduction, of structured and un-structured data sets to enable the formulation of mechanistically informed hypotheses;  
2. Understand how to identify and manipulate features and responses from data sets (typically experimental or unit operation sensors and instruments) to build quantitative relationships for predictive purposes to enable model based decisions;  
3. Evaluate experimental plans using Design of Experiments tools following mechanistic understanding of chemical processes and unit operations;  
4. Plan and analyze experiments for the calibration of spectroscopic instruments (FTIR, Raman, UV-Vis) for monitoring pharmaceutical processes and product quality;  
5. Work effectively on a team to define a project, identify specific tasks required for completion and distribute them among members to collaboratively complete the project;  
6. Present the results of an investigation clearly and succinctly.
Instructor

Dr. Jose Tabora holds a Ph. D. from the University of Virginia. For the past 25 years he has worked for Merck, Eli Lilly and BMS in Pharmaceutical product Development. Through his professional life he has championed the adoption of machine learning and data science in various workflows of pharmaceutical process development. He is a Fellow of the AIChE.

Dr. Tabora interviewed by AIChE:
https://www.youtube.com/watch?v=UBv0xsGim-o&feature=youtu.be

Topics

The topics in the syllabus are to be interpreted to be applications to specific pharmaceutical development case studies as described in the references which will be worked out during class. The theory behind the topics is covered in the co-requisite for the course.

1. Overview of pharmaceutical development and examples of data science applications; Introduction to Python
2. Data exploration analysis
3. Linear regression
4. Design of experiments
5. Generalized linear models and multiple linear regression (MLR)
6. Principal component analysis (PCA)
7. Principal component regression and partial least squares (PLS)
8. Bayesian analysis of linear models
9. Bayesian applications to mechanistic models
10. Bayesian hierarchical models
11. Process robustness, failure analysis
12. Quality by Design (QbD) paradigm for pharmaceutical development and PHARMA 4.0

Pre-requisite: Instructor Permission
Co-requisite: MA541
Cross-listing: MT542
Program:
Chemical Engineering
Available on campus and online
Credit Applied towards Graduate Degree or Certificate
Contact Information
Professor Adeniyi Lawal
Office: 100A McLean Main Campus
Phone: 201.216.8241
Email: alawal@stevens.edu
Course Description

The fourth industrial revolution, referred to as Industry 4.0, will transform all aspects of research, development and manufacturing across all product modalities.

This transformation is characterized by increased adoption of digitization, robotics, and intelligent systems. Artificial intelligence and machine learning in particular are anticipated to play crucial roles in this revolution by enabling the rational integration of the ever-increasing amounts of digital data.

The pharmaceutical development space is rapidly adapting to Pharma 4.0 by integrating machine learning and artificial intelligence in novel and unprecedented ways to the increasing flow and density of digital information. This class will provide a detailed exploration of machine learning algorithms such as neural networks and gradient boosting applied to novel data workflows in pharmaceutical development. These applications result in enhanced quantitative characterization of processes and material properties and enable incorporation of the large data sets anticipated to be generated by the adoption of Industry 4.0 practices.

Course Objectives

1. Students will be able to conduct data aggregation, curation, and pre-processing (data wrangling) of raw data sets to enable optimal application of machine learning algorithms.
2. Students will acquire and be able to demonstrate the appropriate and successful application of different ML/AI algorithms to address pharmaceutical development problems.
3. Students will be able to determine optimal approaches for ML/AI depending on the nature of the problem and the amount and quality of the available data.
4. Students will acquire skills to evaluate and adopt different ML approaches to data sets and use outcome metrics to determine appropriateness of implementation depending on requisite predictive power.

Who Can Take this Course?

Students in Chemical, Mechanical, Industrial, Material Science, Civil and Environmental Engineering programs. Also, students in Physics, and Chemistry. Graduates can pursue careers in Pharmaceutical or related industries (Chemical, Food, etc.).

Course Outcomes

Upon completion of the course, students will have a detailed understanding of the overall pharmaceutical development space (active synthesis and formulation). The students will be able to:

1. Incorporate and evaluate open source software for ML/AI applications.
2. Understand how to recognize common deficiencies of real-case datasets typically encountered in the Pharmaceutical Development space and be able to address them prior to the application of machine learning algorithms.
3. Recognize the applicability of supervised v. unsupervised methods depending on the properties of the data set and the context and goals of the analysis in a realistic Pharma environment.
4. Work effectively on a team to define a project, identify specific tasks required for completion and distribute them among members to collaboratively complete the project.
5. Present the results of an investigation clearly and succinctly.
Instructor

Dr. Jacob Albrecht is a Principal Scientist in Product Development at Bristol-Myers Squibb, with a Ph.D. in chemical engineering from MIT. With over a decade of pharmaceutical development experience, he has championed the adoption of modern approaches to data science and machine learning both within BMS and through pharmaceutical industry consortia.

Topics

The topics in the syllabus are to be interpreted to be applications to specific Pharmaceutical R&D case studies as described in the references which will be worked out during class. The theory behind the topics is covered in the corequisite for the course.

1. Pharma 4.0 with attention to new forms of digital information typical data sets and data wrangling
2. Analysis of raw material in pharmaceutical process via K-Nearest Neighbor clustering
3. Estimation of pharmaceutical performance with K-Nearest Neighbor clustering and
4. K-Means clustering
5. Modeling pharmaceutical processes with neural networks-1
7. Applications of decision trees in the optimization of chemical organic synthesis
8. Random forest for reaction performance estimation
9. Characterization of drug substance properties using support vector regression
10. Optimization of drug product properties using support vector machines
11. Development of pharmaceutical process models with boosting methods
12. Applications of genetic algorithms to optimize spectroscopy calibrations
13. Applications of natural language processing to classify pharmaceutical documents

Pre-requisite: Instructor Permission
Co-requisite: MA576
Cross-listing: MT543

Program:
Concentration in Master’s Degree in Chemical Engineering; Materials Science
Available on campus and online
Credit Applied towards Graduate Degree or Certificate

Contact Information

Professor Adeniyi Lawal
Office: 100A McLean Main Campus
Phone: 201.216.8241
Email: alawal@stevens.edu