Course Outline: **Machine Learning for Biomedical Data**

**Course #:** BMDE 520 / COMP 598-003  
**Term:** Fall  
**Year:** 2022  
**Course pre-requisites:** MATH 203, MATH 223, MATH 323, and one course in computer programming (or equivalent background with permission of instructor). Additional background in Calculus and Probability is recommended, e.g. MATH 222 and MATH 204, but not required.  
**Course co-requisite(s):** None  
**Course schedule (day and time of class):** Tuesdays, 2:35 – 5:25 p.m. EDT  
**Number of Credits:** 3  

**Course Location:** DUFF 507/509, in person  

**Course Instructor:** Dr. Danilo BZDOK, Associate Professor, Department of Biomedical Engineering  
**Email:** danilo.bzdok@mcgill.ca + tas_bmde520_ml_in_biomed+noreply@googlegroups.com  
**Telephone number for office appointments:** 514-398-6738  
**Office location:** B.03, Mila Quebec AI Institute  
**Office hours for students:** TBD

**Course Overview**

Modern datasets are rapidly increasing in size and complexity in industry and government, and especially in the many branches of biomedicine. This course will provide a conceptual overview of practically relevant aspects of different modeling tools, mostly focusing on classical machine learning and deep learning, while comparing these to key principles of mainstream statistics (regression-type modeling) and Bayesian modeling. The goal will be to provide essential modeling intuitions, rather than diving deep into details of mathematical proofs and theorems, while we will contemplate outcomes and consequences of the important formal properties of the discussed modeling tools. We will focus applications and examples on the recently emerged high-dimensional biomedical measurements, centering on high-resolution brain imaging and whole-genome common variant genetics, while also tapping into other modalities of the expanding richness of biomedical information. The majority of discussed tools are applicable and effective in various scenarios that involve modern datasets with many variables per observation.

**Course Description**

Theoretical and practical course in machine learning applied to the expanding richness of biomedical data, including multi-dimensional biomedical measurements centering on high-resolution brain imaging and whole-genome common variant genetics.

**Learning Outcomes**  
By the end of this course students are expected to:  
1) Understand core properties that distinguish different classical and emerging machine learning tools and biomedical data.  
2) Understand and distinguish between several practical examples of how biological research questions and specific analysis tools are applied to biomedical data.  
3) Appreciate the nuances of how distinct quantitative findings and conclusions are enabled by different modeling tools, especially in relation to biomedical research.
4) Demonstrate practical know-how for building machine-learning analysis pipelines in Python.

**Instructional Method**
Lectures and weekly programming exercises with analysis cases in Python

**Required Course Materials**
Course material, prepared by the Lecturer, will be available to registered students via MyCourses.

**Course Content / Outline**

**Week 1: Vocabulary**

- Formal and informal definitions
- Application examples
- Crash course on common technical terms used in machine learning (ML)
- Samples-versus-features ratio
- ML versus statistics
- ML vs artificial intelligence
- Wide data versus long biomedical data

*Required reading:*

**Week 2: Bootstrap**

- The bootstrap as a case example for computer-intensive statistics
- This example is becoming especially interesting now with richly sampled digital-sensor data, questionnaire profiling, biomedical data repositories like body imaging and GWAS approaches in genetics
- History, definition, and answering how bootstrapping extends the toolbox of statistical methods
- The different modelling purposes of the bootstrap
- Procedure
- Point estimates versus interval estimates
- Properties including non-parametric behavior
- Relation to jackknife
- Comparison to MCMC and Bayesian statistics

*Required reading:*

**Week 3: Curse of dimensionality (as a consequence to rich data acquisition)**

- Introduction by anomaly detection
- Unit cube as a function of increasing variables $p$
- Sampling density as a function of $N^{(1/p)}$ for population of the input space
• Examples of k-nearest neighbor
• Principal component analysis and linear regression go awry in high dimensions,
• 3 practical “solutions” to the curse in examples (dimensionality reduction, sparsity, domain-informed bias)

Required reading:

Week 4: Sparsity

• Linear regression versus Lasso
• Lasso versus Ridge versus ElasticNet with regularization behavior and optimization procedures
• Group Lasso
• Sparse group Lasso
• Fused Lasso
• Hierarchical tree sparsity as an example of structured sparsity penalization
• Trace-norm penalization
• Comparison to horseshoe priors in Bayesian regression

Required reading:

Week 5: Bias-variance tradeoff

• Linear regression versus k-nearest neighbors
• Comparison of their behavior in high dimensional datasets
• Computability

Required reading:

Week 6: Cross-validation

• Why more relevant in large-n datasets?
• Procedure
• Purpose (model estimation/selection/assessment)
• Findings from empirical simulation studies
• What is the best in k-fold?
• Restricted model spaces in model regularization

Required reading:

Week 7: Clustering and matrix decomposition techniques

• Structure discovery versus structure inference
• Implicit assumptions of clustering and latent factor procedures
• Why is it ill-posed to search for a best clustering algorithm or clustering solution?
• Why is there no statistical significance of cluster and matrix factorization solutions?
• Cluster validity criteria

Required reading:

Week 8: Diagnostics

• Precision/recall as an alternative to sensitivity/specificity
• Confusion matrix
• F1-score
• ROC curves versus precision/recall curves
• Sample complexity via learning curves in high-bias and high-variance models
• Model calibration

Required reading:

Week 9: Statistical Learning Theory
Important theoretical achievements in classical statistics
- Finite sample theorems
- Hypothesis set
- Approximation error versus estimation error
- VC dimensions
- The fundamental theorem of statistical learning
- Notions of learnability
- Probably Approximate Correct (PAC) Learning
- Uniform convergence
- Non-uniform learnability
- Consistency
- No-free-lunch theorem

Required reading:
- Shalev-Shwartz, S., Ben-David, S., 2014. Understanding machine learning: From theory to algorithms. Cambridge University Press. (chapter 2)
- http://www.no-free-lunch.org/

Week 10: Deep Learning
- Historical overview
- Special empirical properties of deep neural networks (DNNs)
- Autoencoders à la Hinton 2006 Science, why now?
- Types of non-linearity
- Dropout
- Gradient clipping
- Convolutional layers and max-pooling
- Recurrent neural networks
- Long-short term memory neural networks

Required reading:

Week 11: Model scaling with ever deeper and wider biomedical data
- Bayesianism versus frequentism
- Parametric versus non-parametric models
- Inference versus prediction
- Discriminative versus generative models

Required reading:
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning.


**Week 12: Bayesian hierarchical modelling**

- Random effects vs. fixed effects
- Sharing statistical strength by partial pooling between parameters
- Correction for multiple comparisons in the Bayesian context
- Highest posterior density intervals (HPDI)
- Markov Chain Monte Carlo (MCMC) sampling
- Posterior predictive checks vs. types of out-of-sample prediction
- Varying-intercept and varying-slope models
- Bayesian ANOVA

**Required reading:**


**Week 13: Reduced-rank regression and other variants of supervised structure discovery in biomedical data**

- Canonical correlation analysis (CCA) versus partial-least squares (PLS)
- Principal component regression
- Multi-task learning
- Neighbourhood-component analysis
- Semi-supervised autoencoder

**Required reading:**


**Assignments and Evaluations**
Midterm after week 9: 30%
Closed book format, in which objective-type questions will be asked, as well as multiple-choice, and opinion-type questions.

Solutions for coding challenges: 40%
Each week a new coding task will be given that requires to design and implement a data analysis workflow in Python code. The correctness of the programmed solutions from students will be checked and scored.

Final Exam: 30%
Closed book format, in which objective-type questions will be asked, as well as multiple-choice, and opinion-type questions. Covers full material.

McGill Policy Statements

In accord with McGill University’s Charter of Students’ Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

Conformément à la Charte des droits de l’étudiant de l’Université McGill, chaque étudiant a le droit de soumettre en français ou en anglais tout travail écrit devant être noté.

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures” (see www.mcgill.ca/students/srr/honest/ for more information).

L’université McGill attache une haute importance à l’honnêteté académique. Il incombe par conséquent à tous les étudiants de comprendre ce que l’on entend par tricherie, plagiat et autres infractions académiques, ainsi que les conséquences que peuvent avoir de telles actions, selon le Code de conduite de l’étudiant et des procédures disciplinaires (pour de plus amples renseignements, veuillez consulter le site www.mcgill.ca/students/srr/honest/).