Artificial Intelligence Program
College of Engineering
University of the Philippines, Diliman
Second Semester, AY 2023-2024

DS 397: Advanced Computational Methods in Data Science

SYLLABUS AND COURSE CONTENT

COURSE DESCRIPTION:

An algorithmic and computational perspective on machine learning and data science methods

COURSE RATIONALE:

This course teaches the algorithmic and computational aspects of popular methods in machine learning and data science, as opposed to the statistical and mathematical approach. Students are required to apply programming skills in handling, storing, and transforming raw data to make inferences (supervised learning) or distill information (unsupervised learning). This ensures that students have the ability to translate theory into practice by coding learning algorithms on a computer. The course is also necessary for students to learn how to implement new algorithms on their own and publish their work in repositories (e.g. GitHub), which are essential to conduct research and dissemination in the field of data science.

COURSE OBJECTIVES:

After completing this course, the student should be able to:

1. Implement common supervised and unsupervised learning algorithms in code via computational executable notebooks.
2. Understand learning algorithms through existing code implementations.
3. Solve issues arising from numerical implementations of learning methods.
4. Evaluate the performance of popular learning algorithms in terms of time and space complexities.
5. Select the right algorithms for data mining and learning in different cases depending on the available time and computational resouces.
6. Design new learning algorithms incorporating readable code implementation, good programming practice, sufficient testing, maintenance, and version control.

CLASS POLICIES AND REQUIREMENTS:

1. Grading System

In this course, we will adopt the following grading system:

[92, 100]	[88, 92)	[84, 88)	[80, 84)	[76, 80)	[72, 76)	[68, 72)	[64, 68)	[60, 64)	[0, 60)
1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00

Your final grade will depend only on the following requirements:

- Machine Problems 50%
- Data Science Project 30%
- Journal Critique 20%

TOTAL: 100%

A student will receive an **INC** if submissions are incomplete. Students may receive a 5.00 if his/her marks are failing at the end of the course. This course will not give a grade 4.00. Details about each requirement will be discussed below.

NOTE: In this course, students are **allowed** to use ChatGPT or Gemini. But make sure to cite it properly, ensure that it runs well, its output is correct, and you must demonstrate that you understand the code using annotations.

2. **Software**

All DS 397 students are required to have access to Python. During live coding sessions in class, codes will be implemented via Jupyter Notebooks. Students are encouraged to follow along with the coding in their own laptops. Any environment will do, e.g. Google Colab, Visual Studio Code, Anaconda, Jupyter Lite. Students are also encouraged to maintain their own GitHub repositories for the class.

3. **How the Course is Handled**

This course will be handled in full face-to-face mode. Our schedule is 9AM to 12NN, Saturdays at the CSRC Conference Room, University of the Philippines, Diliman.

4. **Machine Problems (MP)**

There will be two MPs given to you to test what you have learned. They should be treated as exams. They must be done ***INDIVIDUALLY AT HOME***. You are allotted one week from the release date of the MP until the submission date. The MP will consist of problems to be solved using code.

For each MP, the submission **must be a PDF file** containing the code written as a Jupyter Notebook.

5. **Data Science Project**

This is a **Team Project** that must be presented at the end of class. Groups can consist of 1 to 3 members. The project requires five things:

- A data set
- A project objective
- A student-original code of a customized machine learning model that achieves the objective
- A rigorous evaluation of the model and visualization results
- A written report and an oral presentation

The aim of the project is to test if the students can write original code to test a customized machine learning model that could not be implemented using libraries alone. Customizing a single model is already sufficient. New ideas can be inspired by, but are not limited to, the following:

- Combine ideas from two different learners (e.g. a neural net with an initial kernel PCA layer)
- Invent your own regularizer or any loss function for any learning architecture.
- Apply a different search procedure for training a model (e.g. train a neural net using PSO)
- Invent your own feature selection, reduction, transformation method specific to your data set.
- Data-level fusion (handling heterogeneous data) or model-level fusion (ensemble learning)
- Apply a custom distance metric (e.g. dynamic time warping inside K-means time-series clustering)
- Re-purpose a supervised learner into an optimization problem (e.g. surrogate modelling)

The criteria for grading the Data Science Project are the following:

- Original idea and original code 40%
- Written Report (Clarity and correctness) 30%
- Oral Presentation (Clarity and mastery) 30%

NOTE: If you wish your work to be **published** in a reputable journal, teams can add an extensive comparison of their proposed method with other baseline models on a data set benchmark. If the results are good, teams are encouraged to write a manuscript complete with an Introduction, Literature Review, Proposed Methodology, Results and Discussion, and Conclusion. These will not be graded anymore. We will let the proverbial *Reviewer 2* judge it!

6. **Journal Critique**

Each student must nominate a recent journal paper about Data Science / Machine Learning / Artificial Intelligence by e-mailing a copy to the instructor. **Review papers are not allowed**. If the instructor approves, he will then give a set of essay-type questions for the student to answer in the form of a critique. The Journal Critique should be a single document containing all the answers, itemized per question. Note: The questions are customized for each selected paper.

7. Course Content

Topic	Course Topic	Course Readings/ Learning Resources
1	**I. Linear Regression** A. Linear Least-squares B. Ridge Regularization C. Cross-validation D. Linear Basis Functions E. Numerical Issues in Matrix Inversions	Bishop (2006) Hastie et al. (2008)
2	**II. Linear Classification** A. Rosenblatt's Perceptron Algorithm B. Logistic Regression C. Confusion Matrix and Classification Performance Metrics	Bishop (2006) Shalev-Shwartz (2014)
3	**III. Kernel Machines** A. Support Vector Machines (SVM) and Sequential Minimal Optimization (SMO) B. Kernel functions for various data types C. Kernel Ridge Regression	Bishop (2006) Cristianini and Shawe-Taylor (2000)
4	**IV. Gaussian Process Regression and Bayesian Curve Fitting** A. Bayesian perspective on Linear Regression B. Bayesian Curve Fitting C. Gaussian Process Regression (GPR)	Bishop (2006) Rasmussen and Williams (2006)
5	**V. Neural Nets and Backpropagation** A. Neural net regression and classification B. Backpropagation C. Stochastic gradient descent	Marsland (2014) Shalev-Shwartz (2014)
6	**VI. Tree-based Learning and Ensemble Learning** A. Decision Tree Algorithm B. Bagging models C. Boosting models D. Stacking models	Hastie et al. (2008) Marsland (2018)
7	**VII. Dimensionality Reduction via Matrix Decomposition** A. Eigenvalue and Singular Value Decomposition B. Data visualization C. PCA and Kernel PCA D. LDA	Bishop (2006) Hastie et.al. (2008)
8	**VIII. Nearest Neighbor Graphs and their Applications** A. Nearest neighbor graphs (NNG) B. k-Nearest Neighbors (kNN) for classification and regression C. Notions of Distance D. Applications of NNG graphs to manifold learning: Isomap and Spectral Embedding	Marsland (2014)
9	**IX. Feature Selection Algorithms** A. Recursive Feature Elimination algorithm B. Min Redundancy, Max Relevance (MRMR) algorithm C. Embedded Methods	Liu and Motoda (2007)
10	**X. Clustering Algorithms** A. K-means clustering B. Expectation Maximization algorithm C. Clustering for time series, texts, and image compression D. Gaussian Mixture Models (GMMs) E. Density-based Clustering	Bishop (2006) Marsland (2014)
11	**XI. Density Estimation, Anomaly Detection, and Sampling** A. Kernel Density Estimation (KDE) B. Anomaly Detectors: One-Class SVM, Local Outlier Factor, KDE C. Sampling: Markov Chain Monte Carlo, Metropolis-Hastings	Hastie et al. (2008) Marsland (2014)
12	**XII. Hyper-parameter Optimization Algorithms** A. Random Search and Grid Search B. Review on GPR C. Bayesian Optimization (BayesOpt) D. Particle Swarm Optimization (PSO), Genetic Algorithm (GA) E. AutoML	Garnett (2023) Simon (2013)

REFERENCES

Bishop (2006). Pattern Recognition and Machine Learning. Springer.

Cristianini and Shawe-Taylor (2014). Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

Garnett, R. (2023). Bayesian Optimization. Cambridge University Press. Available online: https://bayesoptbook.com/

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Ed. O'Reilly Media, Inc.

Hastie, T., Tibshirani, R., Friedman, J. (2008). The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Ed. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R. 2nd Ed. Springer.

Liu, H. and Motoda, H. (2007). Computational Methods of Feature Selection. Chapman and Hall/CRC.

Marsland, Stephen (2014). Machine Learning: An Algorithmic Perspective. 2nd Ed. Chapman and Hall/CRC.

Moroney, L. (2020). Al and Machine Learning for Coders: A Programmer's Guide to Artificial Intelligence. O'Reilly Media, Inc.

Rasmussen and Williams (2006). Gaussian Processes for Machine Learning. MIT Press.

Shalev-Shwartz, S., Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Simon, D. (2013). Evolutionary Optimization Algorithms. Wiley.

Suzuki, J. (2021). Statistical Learning with Math and Python. Springer Singapore.

INSTRUCTOR'S INFORMATION:

Assoc. Prof. Karl Ezra S. Pilario
Department of Chemical Engineering
Artificial Intelligence Program
College of Engineering
kspilario@up.edu.ph

Google Scholar: https://scholar.google.com.ph/citations?user=n41zoQ8AAAAJ
Research Gate: https://www.researchgate.net/profile/Karl-Ezra-Pilario