Lahore University of Management Sciences

CS 535/EE 514 Machine Learning

Fall 2020

Instructor	Agha Ali Raza
Room No.	SBASSE 9-G49A
Office Hours	TBA
Email	agha.ali.raza@lums.edu.pk
Telephone	3306
Secretary/TA	Haris Bin Zia (haris.zia@lums.edu.pk), Muhammad Usama Saleem (21100273@lums.edu.pk)
TA Office Hours	TBA
Course URL (if any)	

Course Teaching Methodology (Please mention following details in plain text)

- Teaching Methodology: We will follow an asynchronous approach for lecture delivery. Exams/quizzes will be conducted over the LMS as timed assessments and take-home exams.
- Lecture details: Recorded lectures would be made available over the course YouTube channel. We may plan a few live sessions if the need arises.

Course basics				
Credit Hours	3 hours			
Lecture(s)	Nbr of Lec(s) Per Week	2	Duration	75 minutes
Recitation/Lab (per week)	Nbr of Lec(s) Per Week		Duration	
Tutorial (per week)	Nbr of Lec(s) Per Week	1 (optional)	Duration	50 minutes

Course distribution	
Elective	This is an elective course.
Open for Student Category	Juniors, seniors and graduates.
Close for Student Category	Please see prerequisites below.

Course description

Machine learning (ML) techniques allow computers to adapt to data and solve new problems that are related to previously encountered problems, more efficiently. Such techniques allow machines to perform useful exploratory and predictive tasks without being explicitly programmed. ML finds its applications in speech recognition and synthesis, machine translation, object recognition, chat bots, question-answering, natural language understanding, anomaly detection, medical diagnosis and prognosis, autonomous vehicles and robots, time series forecasting, and much more. This introductory course covers the theoretical foundations and practical applications of ML and the design, implementation, and analysis of various ML algorithms. Students will learn to compare across and choose the most appropriate algorithms for various problem types and be able to design and implement their solutions. Students will be prepared for both industry and academia as well as for pursuing advanced courses.

Course prerequisites

- Undergrads (Seniors/Juniors) must have passed:
  - An Ugrad/Grad course in Probability (MATH230 (Probability) OR DISC203 (Probability & Statistics) OR CS501 (Applied Probability))
  - And, a programming course (CS200/EE201 (Intro. to Programming))
  - And, a course on Linear Algebra (MATH120 (LA with Diff. Equations))
- Grads are strongly advised to brush up their programming skills and take CS501 (Applied Probability), may be in parallel with ML
- All students must possess strong programming skills and proficiency in algorithm implementation in JAVA/C/Python/MATLAB

Course objectives

The goal of this course is to get the students excited about Machine Learning and to enable them to:

- Develop a strong grip on the theory behind statistical learning
- Understand and rigorously go through the phases of the design, implementation, and evaluation of fundamental ML algorithms
- Choose the appropriate algorithm for each problem type and be able to compare the strengths and weaknesses of algorithms
- Appreciate the end-to-end organic integration of ML in its application areas all the way from data sources, annotation pipelines, and choice of algorithms to societal biases, explainability of models and potentials to impact and even disrupt existing processes

Lahore University of Management Sciences

Learning outcomes

By the end of the course, students should be able to:

- Develop an appreciation for what is involved in learning models from data, and integrating ML in existing real-world processes
- Thoroughly understand the ML pipeline from design and data gathering to meaningful and relevant evaluation
- Learn a wide variety of learning algorithms, and formulate and implement solutions to machine learning problems
- Apply algorithms to real-world problems, optimize the trained models and report on the expected performance

Grading Breakup and Policy (remote)

Programming assignment(s):	25%
Online timed quizzes:	25%
Project:	20%
Reading assignment(s)/homework(s)/Implementation of Research Paper(s):	15%
Online timed final examination:	15%

Examination detail

Midterm Exam	Yes/No:	No
	Duration:	
	Exam Specifications:	
Final Exam	Yes/No:	Yes
	Duration:	2.5 – 3 hours
	Exam Specifications:	Online timed or take-home exam.

Textbook(s)/Supplementary Readings

Textbooks

• Machine Learning, Tom Mitchell, McGraw Hill, 1997 – TM

• The Elements of Statistical Learning: Data mining, Inference, and Prediction, Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, Springer Science & Business Media, 2009 – ESLII

Reference books

• Speech and Language Processing by Jurafsky and Martin, Ed 3 (online draft) – SLP

• Machine Learning: A Probabilistic Perspective, Murphy, Kevin P. MIT press, 2012 – Murphy.

• Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer, 2006 – Bishop.

• Introduction to Machine Learning, Ethem Alpaydin, Ed 2, MIT Press, 2010 – Alpaydin.

• Deep Learning, Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016 – Goodfellow

Course policies

• **Plagiarism:** All work MUST be done independently. In certain assignments students will be allowed to have discussions with peers, in which case they must mention the name and roll number of the student with whom the discussion took place and the nature of the discussion. Even in those assignments, all implementations need to be done independently. Any plagiarism or cheating of work from others or the internet will be immediately referred to the DC. If you are confused about what constitutes plagiarism, it is YOUR responsibility to consult with the instructor or the TA in a timely manner. No "after the fact" negotiations will be possible.

• **Quizzes:** Quizzes will be unannounced. There is no makeup for a missed quiz.

• **Non-uniform weightage:** All Quizzes may not carry the same weight. Similarly, various assignment may carry different weight. These weights may not be announced prior to the submission of the components and will be determined by the course instructor based on factors including (but not limited to) the length, difficulty level, amount of help available, etc. for each component.

• **Programming:** Strong programming skills are expected for this course. Please keep in mind that this is a programming intensive course and you will be spending a lot of time designing and coding up your solutions. All code must be intelligently documented. Undocumented code may not be given any credit.

• **Assignments:** There is negative marking for skipped assignments. Assignments are a basic building block of this course and it will be ensured that students, who pass the course, have significant hands-on experience.

SSE Council on Equity and Belonging

In addition to LUMS resources, SSE's **Council on Belonging and Equity** is committed to devising ways to provide a safe, inclusive and respectful learning, living, and working environment for students, faculty and staff. To seek counsel related to any issues, please feel free to approach either a member of the council or email at cbe.sse@lums.edu.pk.

Mental Health Support at LUMS

For matters relating to counselling, kindly email student.counselling@lums.edu.pk, or visit https://osa.lums.edu.pk/content/student-counselling-office for more information. You are welcome to write to me or speak to me if you find that your mental health is impacting your ability to participate in the course. However, should you choose not to do so, please contact the Counselling Unit and speak to a counsellor or speak to the OSA team and ask them to write to me so that any necessary accommodations can be made.

Lahore University of Management Sciences

Harassment Policy
SSE, LUMS and particularly this class, is a harassment free zone. Harassment of any kind is unacceptable, whether it be sexual harassment, online harassment, bullying, coercion, stalking, verbal or physical abuse of any kind. Harassment is a very broad term; it includes both direct and indirect behaviour, it may be physical or psychological in nature, it may be perpetrated online or offline, on campus and off campus. It may be one offense, or it may comprise of several incidents which together amount to sexual harassment. It may include overt requests for sexual favours but can also constitute verbal or written communication of a loaded nature. Further details of what may constitute harassment may be found in the LUMS Sexual Harassment Policy, which is available as part of the university code of conduct. LUMS has a Sexual Harassment Policy and a Sexual Harassment Inquiry Committee (SHIC). Any member of the LUMS community can file a formal or informal complaint with the SHIC. If you are unsure about the process of filing a complaint, wish to discuss your options or have any questions, concerns, or complaints, please write to the Office of Accessibility and Inclusion (OAI, oai@lums.edu.pk) and SHIC (shic@lums.edu.pk) —both of them exist to help and support you and they will do their best to assist you in whatever way they can. You can find more details regarding the LUMS sexual harassment policy here. **To file a complaint, please write to harassment@lums.edu.pk.**
Rights and Code of Conduct for Online Teaching
A misuse of online modes of communication is unacceptable. TAs and faculty will seek consent before the recording of live online lectures or tutorials. Please ensure if you do not wish to be recorded during a session to inform the faculty member in a timely manner. Please also ensure that you prioritize formal means of communication (email, LMS) over informal means to communicate with course staff.

Course overview

W	Topics	Recommended Readings
1.	**Course overview** ▪ What is ML? The Turing test, traditional CS vs. ML, history of ML, AI vs. ML ▪ ML application areas ▪ Learning: Supervised, unsupervised, semi-supervised ▪ Regression and classification with simple examples (linear regression and decision trees) – cost functions (MSE (mean squared error), MAE (mean absolute error)) ▪ A deterministic binary classifier – REGEX using decision trees **Evaluation metrics** - Confusion matrix (contingency tables) – binary and multi-label - Gold labels and annotation of data ▪ Inter-annotator agreements - Type I and type II errors - Accuracy, Sensitivity (recall), Specificity, False acceptance rate, False rejection rate, Precision - The need for a combine measure - Types of averages: AM, GM, HM - F-β-measure, F-1-measure - ROC, AUC - EER (equal error rate) - Multiclass Classification ▪ Any-of (multi-label) classification ▪ One-of (multinomial) classification- ▪ Micro and Macro averaging	• Murphy chapter 1 • Alpaydin, chapter 1 • SLP3: 4.7-4.9
2.	**Supervised Learning** - Instance and Label spaces ▪ Label spaces for classification (binary and multiclass) and regression - Features and feature vectors ▪ Sparse and dense feature vectors, one-hot vectors - Loss functions and goals of optimization ▪ Zero-One ▪ Squared ▪ Absolute - Hypothesis and hypothesis classes ▪ The No Free Lunch theorem ▪ Choosing the hypothesis class H and hypothesis $h \in H$ ▪ Various Algorithms for traversing hypothesis classes:	• Murphy: 1.1, 1.2, 1.4.2, 1.4.3, 1.4.9

	(a) Pick h randomly (b) try every h (c) just output the label of the training data (memorizer) - Generalization in Learning ▪ Memorizers ▪ Smoothing and Priors ▪ Tradeoff between Bias and Variance -Sampling from the distribution $P(X, Y)$ ▪ Representative datasets ▪ Training, validation and testing	
3.	**Supervised Learning Setup: Training/Validation/Test data** - How to split the dataset D? ▪ Time series data ▪ Independent and Identically Distributed (IID) - The weak law of large numbers ($\epsilon_{TE} \to \epsilon \; as \; \lvert D_{TE} \rvert \to +\infty$) - How to prevent overfitting to test data? Do's and Don'ts - Validation sets (dev sets) and Cross Validation - Goals of Cross Validation: Model selection, training and performance estimation - Types of Cross Validation and Pros and Cons ▪ Exhaustive - Leave-p-out - Leave-one-out ▪ Non-Exhaustive - k-fold - Holdout - Repeated random subsampling ▪ Nested - $k * l$ fold - k-fold with validation and test sets - Bootstrapping - Stratified cross validation - Time series cross validation (forward chaining – Rolling origin)	• Murphy: 1.1, 1.2, 1.4.2, 1.4.3, 1.4.9
4.	**Linear Regression** - Derivation and implementation - Cost functions: Mean Square Loss, 0/1 loss, Absolute Loss - Learning Parameters using Batch Gradient Descent - Multivariate Linear Regression and Gradient Descent - Feature Scaling, local minima and saddle points - Hyperparameters: Learning rate - Polynomial regression	• ESLII Ch3 • Murphy 7-7.5.1, 7.5.4
5.	**Logistic Regression** (A linear, discriminative, parametric classifier) - Derivation and implementation of a simple sentiment classifier - Non-linear activation functions: the Sigmoid - Hyperplanes, linear and non-linear decision boundaries - Cost function: Derivation of the cross-entropy loss function (log loss) - Learning algorithm: Batch, Stochastic and Mini-batch Gradient Descent - Multiclass classification: One-vs-all, simultaneous parameter learning - The SoftMax activation function and multivariate log loss	• SLP3 Ch5, ESLII Ch4, • Murphy 8, 8.1, 8.2, 8.3.1, 8.3.2, 8.3.4, 8.6, 13.3-13.3.1, 8.3.2, 8.3.6, 13.5.3 • Ben Taskar's notes • TM chapter: Naive Bayes and Logistic Regression • Nice blogpost on Gradient Descent, Adagrad, Newton's method
6.	**Bayes Theorem and the Bayes Optimal Classifier** - Review of probability, conditional probability and derivation of the Bayes Theorem - Maximum a posteriori (MAP) and Maximum Likelihood Estimation (MLE) - Example problems and solutions using Bayes Theorem ▪ Monty Hall problem, medical testing, Language Modeling - Bayesian and Frequentist approaches in statistics ▪ Hypothesis spaces ▪ Derivation and comparison of MAP and MLE ▪ Conjugate priors: The Beta and Dirichlet distributions	• SLP Ch4 • Murphy 2.2, 3.1-3.4 • ESLII 6.6.3

	▪ Laplace smoothing - The Bayes Optimal Classifier - Comparison of Generative and Discriminative classifiers	
7.	**The Naïve Bayes Classifier** (A linear, generative, parametric classifier) - Independence, mutual exclusion and conditional independence - Derivation and implementation - The bag-of-words model - Data sparsity and Out-of-vocabulary (OOV) items - Laplace Add-1 smoothing - Sentiment analysis - Gaussian Naïve Bayes - Text generation using Naïve Bayes ▪ The Shannon visualization method for N-grams	• Ben Taskar's notes on Naïve Bayes • TM chapter on Naive Bayes (ch 1-3) • Xiaojin Zhu's notes on Multinomial Naïve Bayes • Mannings' description of Multinomial Naive Bayes
8.	**The K-Nearest Neighbor Classifier** (An instance-based, lazy, discriminative, non-linear, non-parametric classifier) - Definition - KNN decision boundaries and Voronoi Tessellations - Similarity/Distance measures and constraints ▪ Minkowski Distances (Manhattan, Euclidean and Chebyshev) - The KNN algorithm and implementation: KNN regression and classification - Space and Time complexity - Bias/variance tradeoff as $K \to 1$ and $K \to n$ - Choice of the hyperparameter: K - KNN error bounds as $n \to \infty$ - The curse of dimensionality ▪ Challenges ▪ Lower dimensional subspaces and manifolds	• Murphy: 1.1, 1.2, 1.4.2, 1.4.3, 1.4.9 • Videos of different values of k • Video describing nearest neighbors • A nice explanation of nearest neighbors
9.	**Decision/Regression Trees** (A discriminative, non-linear, non-parametric classifier) - ID3 algorithm - Gini Index - Entropy splitting function (aka Information Gain) - Regression Trees	• Decision Tree wiki page • Ben Taskar's old notes • Murphy: 16.2
10.	**Basics of Neural Networks: The Perceptron** (A discriminative, linear, parametric classifier) - Revision: Lines, planes, hyperplanes and vectors ▪ Lines and planes: Normal form and slope-intercept form ▪ Hyperplanes ▪ Decision boundaries with perpendicular weight vectors ▪ Distance between a hyperplane and a point ▪ Geometric interpretation of absorbing the bias in the vectors - The McCulloch-Pitts Neuron and its limitations - The Perceptron and its limitations - The Heaviside step function and comparison with other activation functions - Linear separability as a constraint in in low and high dimensional spaces - The perceptron learning algorithm and its geometric interpretation - Proof of convergence • Relation between margin and rate of convergence	• The Perceptron Wiki page • Murphy 8.5.4
11.	**Maximum Margin Classifiers: Support Vector Machines (SVMs)** (A discriminative, linear/non-linear classifier) - Hard Margin Linear Support Vector Machines: Derivation - The perceptron problem and the optimal separating hyperplane - Constrained optimization and Lagrange Multipliers - Soft Margin Linear Support Vector Machines: Derivation - Hinge-loss - Intuitive introduction of Kernels	• Murphy: 14.5 - 14.5.2.2 • Ben Taskar's Notes on SVMs
12.	**ML Debugging, Over- / Underfitting** - k-fold cross validation - Regularization	• Ben Taskar's under- and overfitting • MLaPP: 1.4.7 • Andrew Ng's lecture – ML debugging

	- How to debug ML algorithms - How to recognize high variance/high bias scenarios - What to do about high variance/high bias **Bias Variance Trade-off** - Bias, Variance and Noise - Error decomposition into Bias, Variance, Noise	• • Ben Taskar's Notes on Bias Variance • Notes by Scott Foreman-Roe • ELSII Chapter 2.9 • Murphy: 6.2.2
13.	**Kernels** - Kernelizing algorithms: Why, how and requirements (**W** as linear combinations, inner products) - RBF Kernel, Polynomial Kernel, Linear Kernel - Classifying indirectly in very high dimensional spaces using the Kernel trick - Kernel SVM	• Ben Taskar's Notes on SVMs • Murphy 14-14.2.1, 14.4, 14.4.1, 14.5.2, 14.4.1 • Derivation of kernel Ridge regression by Max Welling • Kernel Cookbook by David Duvenaud • Laurent El Ghaoui's lectures on duality • "Idiot's guide to SVM"
14.	**Neural Networks** - The Neuron and linear decision boundaries. Can we do better? - Changing the representation of the data • Kernels • Neural networks - Multi-layer Perceptron → Neural Networks → Deep Learning - Universal approximators - Layers: Depth vs. width - Overfitting and the Stochastic Gradient Descent (SGD) - Formal Notation for Logistic Regression and NNs - Vectorizing LR and NNs - Forward propagation (LR, NN): Simple, vectored with a single instance, vectored with m instances - Activation functions, gradients, and pros and cons ▪ Sigmoid ▪ Tanh ▪ ReLU/Leaky ReLU - Backward propagation (LR, NN): Simple, vectored with a single and m instances	• SLP Ch7, ESLII Ch11
15.	**Unsupervised Learning** - Vector Quantization - K-means - GMMs - Expectation Maximization	• SLP Ed2 Ch9, ESLII Ch13
16.	**Big Challenges and Opportunities in AI and ML** - The case for Explainable AI - The case for Fair AI ▪ Societal biases ▪ Imbalanced classification - Machine Learning for Development (ML4D)	
	Other topics – to be covered if we have time	
17.	**Ensemble methods** - Bagging and Random Forests - Boosting and AdaBoost	• ESLII 8.7, ESLII Ch10, 15, 16
18.	**Graphical Sequence Processing Models** - Hidden Markov Models (HMMs) - Maximum Entropy Markov Models (MEMMs) - Undirected Graphical Models (Markov Random Fields) - Conditional Random Fields (CRFs) - Directed Graphical Models (Bayes Nets)	• SLP A, ESLII Ch17
19.	**Deep Neural Networks** - RNNs and LSTMs	• SLP Ch9