CUNY School of Professional Studies

Syllabus
School of Professional Studies
DATA 622: Introduction to Machine Learning: Supervised Learning

Instructor Name: Raman Kannan
Instructor Email Address: Raman.Kannan@sps.cuny.edu
Degree Program: M.S. in Data Science
Credits: 3 graduate credits
Prerequisites: *605,606*
Type of Course: ONLINE – Asynchronous
Weekly Meetup: Mondays 8 PM to 9 PM Blackboard collaborate.
Office Hours: Mondays 7 PM to 7:30 PM venue Blackboard collaborate by appointment only.

Course Description

The course will develop basic understanding of supervised learning techniques, a generic process method to execute classification exercise, leveraging the Statistics and Linear Algebra students have acquired in 605 and 606.

This is a clinical course and emphasis is on computational techniques: designing, running and refining supervised learning models introduced in the course. We will be primarily using R as our programming environment. All assignments will need to be submitted in as R-Markdown documents or executable R scripts on cloud computing resources the instructor will provide free of cost. Students are required to demonstrate

1. *good understanding of what supervised learning/classification models are; and*
2. *proficiency and experience in running classification exercise on any given dataset*
3. *familiarity and mastery over technical terminology expected of a practitioner.*

Program Learning Outcomes

1. Conceptual Understanding of Classical Supervised Learning Techniques. Understand the statistical/algebraic foundation of classification algorithms, relative strengths and weakness of various techniques, theoretical and practical criteria in adopting a model.
2. Data Understanding. Collect, describe, model, explore and verify data.
3. Engineering. Use industry standard environment and process to conduct repeatable and reproducible classification exercise.
4. Experimentation and Analysis: Run prescribed process to optimize model using multiple classification algorithms.
5. Presentation. deliver summary results of the experiments and explain key decisions they made in designing the model and model output.

Learning Objectives:

1. Develop deep understanding of both parametric/non-parametric classification algorithms.

2. Prepare datasets for classification algorithms, plan and conduct modeling exercise on given datasets. Understand Supervised Learing process.
3. Apply techniques introduced to reduce overfitting and mitigate the adverse effects of variance and bias. Techniques include combining classifiers, cross validation, penalization and resampling methods.

Assignments and Grading

Grades in this course are determined by the percentage of points obtained.

Course assignments	**Percentage of Final Grade**	**Points**
Homework – Run the model exercise+know how of Algorithms	30.00%	30
✓ There will be 1 individual modeling assignment (15%) and 3 knowledge tests 5% each. ✓ Grading based on process discipline (3%) and ability to prepare data (4%) , run the model (2%), interpreting the results (3%) and summarizing the results (3%)	15% x 1 = 15% 5% x 3 = 15%	
Class Discussion	12.00%	12
✓ You must participate in weekly forums and discussions. ✓ Discussions are applied analysis from the texts. ✓ You must post a response by Wednesday at midnight (ET) ✓ You must respond to at least one of your colleagues' contributions by Saturday at midnight (ET), ✓ You should provide meaningful feedback on the analysis.	12 x 1% = 12%	
Homework – demonstrate ability to improve performance	40%	40
✓ There will be 1 individual modeling assignment (25% each) and 3 knowledge test 5% each ✓ Grading based on computing specified performance measures (12%), and 8% for systematically seeking to improve performance using prescribed methods	25x1+5x3	
✓ Open book, open notes review test – t1	9.00%	9
✓ Open book, open notes review test – t2	9	9

Total	**100%**	**100**

Knowledge tests (5%) are given once in two weeks.
Projects are handed out 15th Sep and Oct 15th and are due Oct 15th and Nov 15th.
Submission Protocol
In order to be graded you have to follow these strictly – dont change names or case.
All work must be submitted on IBM Cloud. Email submission or blackboard submissions cannot be graded.
Create two directories p1 and p2. For each project you have to submit the following files. For p1, submit the files in p1 directory and p2 files in p2 directory.
To be graded each project must have the following files:
All data (csv)
software (notebooks,RMD, scripts and
a report (a PDF or txt format is acceptable) including description, analysis and results of the work.
If you work in R you may submit R or r or RMD or rmd files
If you work in python you may submit .py or notebook files
You can also submit shell scripts and or java applications
Include a studentid_p1_howto.txt file with instructions to compile(if needed) and run.
All files must be named as follows

studentid_p1_howto.txt
studentid_p1_setup.R or studentid_p1_setup.r
studentid_p1_process.R
studentid_p1_setup.py
studentid_p1_process.rmd,
studentid_p1_scrape.sh,
studentid_p1_scrub.sh
studentid_p1_report.txt,
studentid_p1_report.pdf
studentid_p1_dataset_name.csv
studentid_p1_references.url
studentid_p1_sow.txt

Other formats are NOT graded. ZIP files, archives containing all these files are NOT graded.
OnIt submissions accessible ~studentid/p1/studentid_p1_type.ext where ext is R[r],py,ipy,rmd,sh,txt,csv,pdf,url.
Any algorithmic exercise, must include the sow (statement of work). The SOW must include your analytical setup: H_o, acceptance criteria (statistical measures you will use to reject H_o or fail to reject H_o, criteria you will use for that experiment to evaluate the significance of features.

Late Policy for Homework:

Late work will not be graded. All assignment due dates and times as posted in Blackboard.

Course Materials (All Open Source)

URL:http://www-bcf.usc.edu/~gareth/ISL/

AUTHORS:Trevor Hastie, Robert Tibshirani, Jerome Friedman

TITLE:An Introduction to Statistical Learning, CODE:ISL RECOMMENDED

URL:http://faculty.marshall.usc.edu/gareth-james/ISL/ (booksite, pdf,data,R etc)

http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf (pdf)

AUTHORS:Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Following books are optional NOT required. Highly recommended as they are free and deep in their treatment and presentation:

URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

TITLE:Elements of Statistical Learning, CODE:ESL (AMBITIOUS STUDENTS)

(books below are of the class of books by Christopher M. Bishop and book by Kevin P Murphy)

TITLE:Understanding Machine Learning, CODE:UML (AGGRESSIVE)

URL:https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

AUTHORS:Shai Shalev Shwartz and Shai Ben David

There are many books that are very good but are available only in print.

SWARM Intelligence – our collective conscience-wisdom

https://www.quora.com/What-is-the-best-book-to-learn-ML

stackoverflow

Papers

papers, articles, industry case studies and white papers will be distributed through out the course.

Relevant Software and Other Tools:

Students should have R Studio & R installed in their computers. Relevant libraries that are required will be posted in the assignments and course materials. Students are expected to submit R-Markdown files for their assignments. ***Or submit R Script on IBM Cloud. Instructor will provide access to IBM Cloud once registration is stabilized, fully sponsored by IBM Power Systems Academic Initiative. Either R or RMD may be submitted on blackboard.***

Instructor will demonstrate all work exclusively on IBM PSAI Cloud Node.

Students are encouraged to work in their preferred environment – python, java/spark, Azure M/L in either the cloud or their local computing environment. Students may also do their work on IBM Cloud and R that instructor will provide. All work must be submitted on IBM Cloud.

Course Outline:

Please note that this schedule is subject to change depending on our progress, questions, requests, etc.

Module	Weekly Meetup	Topics	Reading (included in the weekly materials)
1	08/31/20	Introduction to Learning, Machine Learning Generic Concepts applicable to all	

		supervised learners: Performance, Comparative Analysis, Occam's Razor,No Free Lunch Theorem, Process	
2	09/14/20	Logistic Regression: Sigmoid Transformation Binary Classifier [0 or 1]	ISLR Chapter 4
3	09/21/20	Naive Bayes: Conditionally Independence Multi-class	ISLR Chapter 4
4	09/29/20 (TUESDAY)	Linear Discriminant Analysis (LDA) and QDA Multi-class	CUNY Closed on 28th.. 29th Tuesday follows Monday MAKE A NOTE.
5	10/05/20	Birds of a Feather flock together. Instance based Learner:kNN non-parametric multi-class	
6	10/14/20 (THURSDAY)	Decision Tree non-parametric, logic based, multi-class	CUNY Closed on 12th.. 14th Thursday follows Monday MAKE A NOTE.
7	10/19/20	Support Vector Machine Margins, Binary Classifier	
8	10/26/20	Bias and Variance, tradeoff Correction method penalization	
9	11/02/20	Resampling, Cross Validation methods	
10	11/09/20	Gradient Descent, MLE Loss Functon, Cost Optimization	
11	11/16/20	Single classifier: Iterative Refinement Boosting Diversified Combination:Bagging Ensemble: Random Forest	
12	11/23/20	Combining heterogeneous classifiers Stacking	
13	11/30/20	Comparing Classifier performance	
14	12/07/20	Big Data: ML over clusters, opportunities for parallel/distributed processing	
15	12/14	Class review: reflection from students course evaluation to CUNY	All

Course Meeting Time:

See the course website.

The Course schedule follows the official SPS CUNY Academic schedule:

https://sps.cuny.edu/academics/academic-calendar/2020-2021-calendar/fall-2020

Grade Distribution

Quality of Performance	Letter Grade	Range %	GPA/ Quality Pts.

Excellent - work is of exceptional quality	A	93 - 100	4.0
	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3.0
Below Average	B-	80 - 82.9	2.7
Poor	C+	77 - 79.9	2.3
	C	70 - 76.9	2.0
Failure	F	< 70	0.0

Accessibility and Accommodations

The CUNY School of Professional Studies is firmly committed to making higher education accessible to students with disabilities by removing architectural barriers and providing programs and support services necessary for them to benefit from the instruction and resources of the University. Early planning is essential for many of the resources and accommodations provided. For more information, please see:

Disability Services on the CUNY SPS Website.

Online Etiquette and Anti-Harassment Policy

The University strictly prohibits the use of University online resources or facilities, including Blackboard, for the purpose of harassment of any individual or for the posting of any material that is scandalous, libelous, offensive or otherwise against the University's policies. Please see: "Netiquette in an Online Academic Setting: A Guide for CUNY School of Professional Studies Students."

Academic Integrity

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the educational mission of the City University of New York and the students' personal and intellectual growth. Please see:

Academic Integrity on the CUNY SPS Website.

Student Support Services

If you need any additional help, please visit Student Support Services.