APSTA-GE.2011 Marc Scott
Supervised & Unsupervised Machine Learning January 2023
Class Meeting: Online on Zoom, on the following days/times Office: 207W Kimball Hall

Week 1:	Wed. Jan. 4: 9a-12p	Fri. Jan. 6: 12:30-3:30p
Week 2: Mon. Jan. 9: 12:30-3:30p	Wed. Jan. 11: 9a-12p	Fri. Jan. 13: 9a-12p
Week 3:	Wed. Jan. 18: 9a-12p	Fri. Jan. 20: 9a-12p

email: marc.scott@nyu.edu

Office Hours: Prof. Scott's Office Hour begins 1 hour after any class ending at 12pm and 30 minutes after the two classes that end at 3:30pm. Beginning Jan. 5, there will be an additional hour 12-1pm on Tuesdays and Thursdays, and there will be the possibility of appointments.

Course Assistant: Nora Tang <xt367@nyu.edu> is our course assistant. Nora will have office hours Tuesdays and Thursdays 9am-10am and by appointment.

Text: There is no required text. Selected chapters from several sources will be made available.
Software: R (required)
This course will use **NYU Brightspace** for distribution of information (handouts, announcements), **and** for turning in assignments. We also use Bookdown as a comprehensive learning environment (notes, practice assignments and labs all integrated into one document). More details will be given in course announcements.

COURSE DESCRIPTION:
Supervised and unsupervised machine learning, also known as classification and clustering, are important statistical techniques commonly applied in many social and behavioral science research problems. Both seek to understand social phenomena through the identification of naturally occurring homogeneous groupings within a population. Supervised learning techniques are used to sort new observations into pre-existing or known groupings, while unsupervised learning techniques sort the population under study into natural, homogeneous groupings based on their observed characteristics. Both help to reveal hidden structure that may be used in further analyses. This course will compare and contrast these techniques, including many of their variations, with an emphasis on applications.

PREREQUISITES:
The course is an advanced masters level statistics course with substantial computing requirements. Prior coursework covering the General Linear Model (as well as logistic regression), Maximum Likelihood Estimation, Probability, including multivariate distributions, as well as statistical computing is required. Note: we do not use specific course numbers as prerequisites in Albert, as there are many reasonable alternatives. However, APSTA-GE 2003, 2351, 2352 is a common pathway to success in this class. If you wish to take this class and believe that you have equivalent prerequisites, you should have emailed me with the precise descriptions of the classes you have that meet these requirements ***before registering***.

COURSE REQUIREMENTS:

Participation:	15%	You are expected to attend class and participate. This will be assessed with end of class short quizzes, Ed Discussion, and potentially a few post-lab assignments.
Homework problems:	30%	There will be several assigned problems intended to give you practical experience with the methods discussed.
Data Analysis Projects:	55%	There will be two data analysis projects (worth 25% and 30%).

Data Analysis Project 2: This can be an original analysis of data that you provide (subject to my review beforehand) in groups of up to four. Alternatively, I can make available a dataset and instructions for analysis that is more guided, and groups may be up to size two. If you choose to analyse your own data, you need to form a group in the first week of class and begin discussing your analysis plan. This is essential so that you do not get bogged down with data processing problems (like missing data) that are quite common.

GRADING: the grading system is as follows:

A: 94.5% and above	A-: 89.5-94.49%	
B+: 86.5-89.49%	B: 82.5-86.49%	B-: 79.5-82.49%
C+: 76.5-79.49%	C: 72.5-76.49%	C-: 69.5-72.49%
D+: 67-69.99%	D: 64-66.99%	D-: 60-63.99%
F: less than 60%		

Class Expectations:

This course is highly interactive, both in terms of working and learning in teams and as a group. It is often based on readings as well, so being prepared is important. However, interaction takes a variety of forms, ranging from one-on-one discussions to group presentations, so that different skills are emphasized at different times. Attendance, online quizzes/questionnaires, and active participation in Ed Discussion will form **the bulk of the participation grade**.

Academic Integrity and Group Work Policies:

All students are responsible for understanding and complying with the **NYU Steinhardt Statement on Academic Integrity.** Unless otherwise indicated, you should work on the assignments by yourself. You may discuss problems with your classmates, the course assistant, and the professor, but your submitted work must be your own (see the Academic Integrity statement here). In particular, you may not copy code from your classmates; this will be deemed plagiarism and sanctioned according to NYU guidelines.

Students with Disabilities:

Students with physical or learning disabilities are required to register with the Moses Center for Student Accessibility, 726 Broadway, 2nd Floor, (212-998-4980 and online at http://www.nyu.edu/csd) and are required to present a letter from the Center to the instructor at the start of the semester in order to be considered for appropriate accommodation.

Inclusion:

NYU values an inclusive and equitable environment for all our students. I hope to foster a sense of community in this class and consider it a place where individuals of all backgrounds, beliefs, ethnicities, national origins, gender identities, sexual orientations, religious and political affiliations, and abilities will be treated with respect. It is my intent that all students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource and strength. If this standard is not being upheld, please feel free to speak with me.

COURSE READINGS: Handouts will be available several days before each class. **Review Beforehand**. **Ask and respond to questions (online) via the predetermined mechanism.**

Late assignment policy: **Assignments are to be handed in on time.**

SCHEDULE AND READINGS ON NEXT TWO PAGES

SCHEDULE FOR January 2023

Date		*Topic*
Jan	4	Introduction; what is a cluster; visualization techniques, including principal components. The supervised learning technique you already know (logistic regression). Introduction to hierarchical clustering methods.
	6	Hierarchical clustering: linkage choices; distance measures; the dendogram. Optimization techniques (k-means); choosing the number of groups; evaluating clusters;
	9	Model-based clustering (including model selection); ***HW 1 DUE (12:30pm)***
	11	Supervised learning – intro via K-Nearest-Neighbor; logistic regression; (Linear) Discriminant function analysis; ***HW 2 DUE (9am)***
	13	Support Vector Machines; kernel trick; Cross-Validation; ***PROJECT 1 DUE (9am); PROJECT 2 DRAFT PROPOSAL DUE***
	18	Tree-based methods; Naïve Bayes Classifier.
	20	Best practices; Caret function; Specific Tree-based method: Random Forests; Neural Networks; **HW 3 DUE (9am);** ***PROJECT 2 FINAL PROPOSAL DUE***
	23	*(Monday by 9am)* ***PROJECT 2 DUE***

READINGS

Recommended text (general resource): Rogers & Girolami (2016). "A First Course in Machine Learning," Second Edition, CRC Press. Available online through NYU Bobst as an e-book here (***may require NYU login***): https://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=4718644

Class 0:

- Review Maximum Likelihood Estimation:
  - https://www.youtube.com/watch?v=93fPFOf547Q
  - https://www.youtube.com/watch?v=Dn6b9fCIUpM
  - https://apsta.shinyapps.io/MLE-Sim/ (do not do last two parts unless very experienced with MLE)
- Review PCA online:
  - https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Classes 1 & 2: **Chapters/Handouts 1 & 2**

Everitt et al., Cluster Analysis (4th Ed.), chapters 1 & 2. Brian F. Manly, *Multivariate Statistical Methods, A Primer (2nd Ed.),* chapter 11.
Everitt & Dunn, *Applied Multivariate Data Analysis*, chapter 6, sections 6.1, 6.2. Brian F. Manly, *Multivariate Statistical Methods, A Primer (2nd Ed.),* chapter 9.

Class 3: **Chapters/Handout 3**

Everitt & Dunn, *Applied Multivariate Data Analysis*, chapter 6, sections 6.3, 6.4. Peter J. Rousseeuw (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
Banfield & Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, Vol. 49, no. 3, 803-821.
Fraley and Raftery (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578-588.
Fraley, C., & Raftery, A. E. (2006). MCLUST version 3: an R package for normal mixture modeling and model-based clustering. *TR 504. UNIV. WASHINGTON/SEATTLE DEPT OF STATISTICS.*
Leisch, F. (2008). Visualizing cluster analysis and finite mixture models. In Handbook of Data Visualization (pp. 561-587). Springer: Berlin
Proust-Lima, Cécile, Viviane Philipps, and Benoit Liquet. *"Estimation of extended mixed models using latent classes and latent processes: the R package lcmm." arXiv preprint arXiv:1503.00890 (2015).*

Class 4 & 5: **Chapters/Handouts 4, 5**

Everitt & Dunn, *Applied Multivariate Data Analysis*, chapter 11 (skip 11.11). Brian F. Manly, *Multivariate Statistical Methods, A Primer (2nd Ed.),* chapter 8. Tabachnick & Fidell, *Using Multivariate Statistics (4th Ed.),* chapter 11.
G. James et al. (2013). *An Introduction to Statistical Learning: with Applications in R*, Springer: New York. Chapter 9

Class 6 & 7: **Chapters/Handouts 6, 7**

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.