ISyE 7406 **Data Mining & Statistical Learning** **Spring 2020**

Instructor: Dr. **Yajun Mei** (Pronounced as "YA-JUNE MAY"), Email: `ymei@isye.gatech.edu`; Tel: 404-894-2334. Office: Groseclose 343; Office Hours: after class, 1:15-1:45pm TR.

Class Meets: Tuesdays and Thursdays 12:00-1:15pm in MRDC 2404.

Office Hours: Tuesdays and Thursdays 1:15-1:45pm.

Textbooks: 1. Hastie, Tibshirani, and Friedman, "*The Elements of Statistical Learning,*" (http://statweb.stanford.edu/∼tibs/ElemStatLearn/). 2. "*An Introduction to Statistical Learning*" by James, et al., see (http://www-bcf.usc.edu/∼gareth/ISL/)

Homepage: Canvas. We will use Canvas extensively, e.g., homeworks, lecture notes, R code, datasets.

Catalog description: Topics include neural networks (Ch 11), support vector machines (Ch 12), classification trees (Ch 9), boosting (Ch 10) and discriminant analyses (Ch 4).

Grading: The course grade is based on homework (20%), class participation (5%), midterm project report (20%, individual, 48-hour take-home to analyze data, **March 3-5**), team project (55% = proposal 3% + oral presentation 20% + report 33%). There will be no regular exams, and the projects are in lieu of exams. Since this is an advanced statistical course that aims to enhance research experience, **the write-up/presentation is also part of grading** for data analysis in homeworks, midterm and final project.
Only for the distance learning students: *one-week delay for all assignments.*

Homeworks: $3 \sim 5$ Homeworks will be assigned, and **due back at Canvas** after one week (we will use Canvas extensively to save trees). Active, live collaboration/discussion is allowed and encouraged on the homework, but **each student must write down the solution in her or his own individual way** and there should be no two identical solutions to any problem. In particular, **do not consult the solution sets from classmates, previous years or online** when working this year's problems. Late homework might be accepted with 50% penalty within 72 hours of the deadline (and no penalty if you have a valid reason, i.e., a note from the Dean's Office). No homework is accepted after 72 hours of the deadline.

Team Project: The detailed guideline for the project will be handed out later. You are encouraged to work in a team of $2-4$ students, but it would be fine if you plan to do it all by yourself. If you decide to work in a team, you will need to **submit only one report per team**. You can choose a project related to your own research interest, and please feel free to discuss the topic with the instructor before starting on the project. Be aware of the following deadlines *The DL students do not need to do oral presentation, but still need to turn in the proposal, the presentation file, and the final report, with one-week delay policy.*

1. **March 12 (Thursday):** the ($1 \sim 3$ page) Project Proposal is due at Canvas, and the purpose is to get you started. It also allows the instructor to provide feedbacks.
2. **10:00am on April 14 (Tuesday)**: the Presentation file of your team project is due for on-campus students at Canvas (either pptx or pdf version will be fine). All team members are expected to present.
3. **April 21 (Tuesday):** the **Final Written Report** is due at Canvas. In your write-ups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

Software: You are expected to self-learn or use any software you like, e.g., Python, R, SAS, Matlab, etc. Handouts with R (or Python) code lines will be provided from time to time.

TA: TBD.

Academic Honor Code: It is your responsibility to get familiar with the Georgia Tech Honor Code. If you have any questions, please consult me or `http://osi.gatech.edu/content/honor-code`

Please turn page over →

ISyE 7406 **Data Mining & Statistical Learning** **Spring 2020**

Brief Schedule

Yajun Mei (ymei@isye.gatech.edu)

Catalog description: Topics include neural networks (Ch 11), support vector machines (Ch 12), classification trees (Ch 9), boosting (Ch 10) and discriminant analyses (Ch 4).

Tentative Schedule is as follows.

week	Date	Lect#	Content
1	Jan 07	1	Intro. to data mining
	09	2	Overview of Supervised learning (Ch2)
2	Jan 14	3	Bootstrapping, Cross-validation, KNN
	16	4	Linear Methods for Regression (Ch3)
3	Jan 21	5	OLS, Ridge, LASSO, Dantzig; PCA, PLS
	23	6	
4	Jan 28	7	Linear Methods for Classification (Ch4)
	30	8	LDA, QDA, Naive Bayes, Logistic Regression
5	Feb 04	9	Case Study: Viral set point
	06	10	Local Smoothers & Additive Models
6	Feb 11	11	LOESS, Kernel, Spline (Ch5,6)
	13	12	(Generalized) Additive Models (Ch9.1)
7	Feb 18	13	Tree (Ch9.2)
	20	14	
8	Feb 25	15	Ensemble Methods
	27	16	BMA, bagging, stacking (Ch 8), Boosting (Ch 10)
9	Mar 03	17	Neural Networks (Ch11) **distribute midterm take-home exam/data-set**
	05	18	**midterm report due at 12noon today**
10	Mar 10	19	Support Vector Machines (Ch12)
	12	20	**Project Proposal due**
11	Mar 17		**No class, Spring break**
	19		
12	Mar 24	21	Unsupervised Learning
	26	22	Associate Rules (Ch 14.2)
13	Mar 31	23	Cluster Analysis (Ch 14.3): K-means Clustering (Ch 14.3.6)
	Apr 02	24	EM algorithm (Ch 8.5), MM algorithm
14	Apr 07	25	Selected Topic, TBD (Time series/streaming/graph/network data?)
	09	26	If needed, optional Project Presentation
15	Apr 14	27	Project presentation **Presentation File of project due at 10:00am today**
	16	28	Project presentation
16	Apr 21	29	Project presentation **Final written report due, last class**