STAT 675 – Statistical Machine Learning – Fall 2015

Instructor

	Office	Email	Office Hours
Darren Homrighausen	Stat 204	darrenho@stat.colostate.edu	By Appointment

Lecture	TR 2:00 PM – 3:15 PM Statistics 006
Required Text	None (See first lecture for recommended texts)
Web Site	www.stat.colostate.edu/∼darrenho/SML
Prerequisites	No formal prerequisites. See summary below for expectations.
Expectations	As this is a 3 credit class, there is a CSU expectation that you spend 6 hours outside class on homeworks and review each week.

Upon completing this course, you should be able to tackle modern data analysis problems by: (1) selecting the appropriate methods and justifying your choices; (2) implementing these methods programmatically (using, say, the R programming language) and evaluating your results; (3) having some basis theoretical tools to understand the properties of modern machine learning methods and statistical machine learning techniques.

Administrative Remarks

Honor Pledge

This class operates under the tenets of the CSU honor pledge: *I will not give, receive, or use any unauthorized assistance.*

Software

R. In this class you will be provided the opportunity to work with `R`; a widely used statistical computing platform. It is free, moderately well documented, and has a vast user community (download it from `www.r-project.org`).

Python. We may additionally dabble in `Python`. It is a flexible platform that is increasingly being used for data analysis tasks.

Evaluation

Scribe duties. Every lecture, one person will be elected as scribe. This person's job will be to write up (using a LaTeX template I will provide) the day's lecture. There will be topics that I didn't go into much detail in the lecture but I think would be useful to have more exposition on. These topics will be notated as e.g.

$$\text{and } \mathbb{X}^\dagger \text{ is the Moore-Penrose pseudo inverse}^\sharp$$

The scribe will do the necessary research to provide at least definitions about these concepts and ideally some useful, related results. Continuing this example, we might state the four conditions on the pseudo inverse that make it a Moore-penrose, how to produce it via an SVD, and maybe how it can be used to construct a projection (ie: $\mathbb{X}\mathbb{X}^\dagger$).

The goals of the scribe duties are that it will...

- encourage critical thinking of the lecture materials
- produce a nice set of notes for us all to have for reference
- provide useful LaTeX practice
- give me some sense of how well the material is being understood
- provide incentive to collaborate with your fellow students to fill in gaps in your notes/knowledge.

Homeworks. There will be periodic homeworks that build on topics from lecture. These homeworks will not be formally graded. However, I'd like to produce nice solutions for each problem. Hence, I will be collecting nicely typeset (ie: LaTeX) solutions using a template I provide. I will randomly assign groups of three people to homework teams.

Presentations. During the semester, there will (hopefully) be topics that arise that you find extraordinarily interesting. I want each person to give two short presentations over the course of the semester. These talks can be informal and can be on anything related to lectures or homeworks.

Possible topics could be a proof of a result I mention, additional properties of an introduced concept, or some data analysis results using the methods from the class.

Exams. There will be no exams.

Miscellaneous

Disability Resources. If you require a special accommodation, such as needing more time to finish exams, contact me **and** disability services.

Class Topics

Applied Topics

Preliminary materials

Linear regression methods:

i. Model selection

ii. Regularization

iii. Squared-error risk estimation

iv. Compressed sensing

v. Sparse additive models

Classification:

i. Discriminant analysis

ii. Support vector machines

iii. Kernel methods (Mercer kernels, Reproducing Kernel Hilbert Spaces)

iv. Tree methods (along with boosting/bagging/bootstrap)

v. Neural networks

(Note: In most cases, methods can be used for either classification or regression with minor changes)

Dimension reduction:

i. Principal Components

ii. Nonlinear embeddings

Clustering:

i. K-means

ii. Hierarchical clustering

Theoretical Topics

Concentration of measure:

i. Moment Bounds versus MGF bounds. Hoefdings, McDiarmid, Nemirovski, (generic) chaining

ii. Empirical processes

iii. Covering numbers

iv. VC dimension

Minimax bounds:

i. Function spaces and information theory

II. Normal means

iii. Upper bounds

iv. Lower bounds

Note: This is a rough overview of possible topics, not a schedule. The order of the topics, or even whether they will be included at all, is subject to change.