# STAT 740 : Foundation of Statistical Machine Learning – Fall 2020

**Instructor:** Wen Zhou (Email: riczw@stat.colostate.edu)
**Meeting Place:** We will primarily meet virtually and the live lectures will be delivered through zoom. Recorded lectures will be uploaded to repository accessible to the class. The zoom link will be send to the class.
**Meeting Hours:** TR 12:30pm-1:45pm
**Office Hours:** TR 11:00am-12:00pm or by appointment; a zoom link will be send for office hour.
**Webpage:** `https://www.stat.colostate.edu/~riczw/teach/STAT740_F20/Stat740.html`

**Prerequisites:** Calculus, linear algebra, probability theory, linear models and regressions.

**Objectives:** Last two decades have witnessed an explosion of available information in almost all fields. Massive data of ultra high-dimensional, complex structured or even unstructured, dynamical, and heterogeneous have been continuously producing, collecting, storing, and becoming more affordable to industrial institutions, academic researchers, investors, and individuals. Learning from these unprecedented massive data and making accurate predictions bring great challenges to both algorithm-driven machine learning methods and traditional statistics. Statistical machine learning or statistical learning emerged in response to the challenges by emphasizing on statistical models and concepts, and particularly, the assessment of uncertainty. A vast range of fields have been influenced by the development in statistical machine learning such as artificial intelligence, bioinformatics, control, environmental science, economics, finance, game theory, genomics, genetics, information communication, management science, networking, signal processing, etc.

The aim of this course is to provide some theoretical foundation for methodologies widely used in statistical machine learning or "data sciences", as well as to introduce some fundamental methods including some modern models/algorithms and their theoretical bases. Illustrations of the methods in some applications will be discussed. Given the time limit, some topics to be covered include

- Fundamental topics
  - Introduction/review to (convex) optimization
  - Linear model: introduction/review, regressions, model selection, regularization(s)
  - Classification, decision trees
  - Bagging, ensembles, and random forest
  - Support vector machines (SVMs)
- Advanced topics
  - Kernel methods
  - Boosting
  - Generalizing error and PAC framework: concentration of measures
  - Ranking and multiclass problems
  - Neural network
  - Dimension reductions and clustering
  - Graphical models
- If time allows, one or some of the following topics might be covered
  - Robustification and distributional robust learning
  - Bandit problems

- On-line learning
- Maxent models: maximum entropy models and logistic regression

**Learning Outcomes and Expectations:** The students are able to understand the theory behind various existing statistical machine learning methods, and they are expected to be capable to provide theoretical justifications on newly developed methods and understand their practical limitations. Upon completing this course, the student can tackle modern data analysis problems by:

- selecting the appropriate models/methods and justifying the choices;
- implementing these methods programmatically using language like `R` and evaluating the results **statistically**; and
- having solid theoretical tools/senses to understand the properties of modern machine learning methods.

The students are highly recommended to spend at least six-eight hours outside of the instructional time on reading, homework, and project.

**Course Work**

**Homework:** Homework will be assigned approximately every week (it will be assigned on Thursday in general and due on the following Friday), and each assignment will carries equal weight.

**Exams:** One take home midterm exam will be assigned to cover materials up to the kernel methods (or SVMs, depending on the progress of the course).

**Final project:** A final project will be used to strengthen students' understanding on the materials. The options include presenting newly-developed statistical machine learning methods from literature, or empirical data analysis using lectured methods, or developing new statistical machine learning methods. The final formality will be discussed and determined in class.

**Grading:** Homework (40%), midterm (20%), and final project (40%). There is no quota or limit to the number of potential A's or any other grade.

**Course Policies**

1. Late homework: No credit unless a prior permission is granted.
2. Any grading dispute must be submitted in writing to me within one week after the work is returned. No changes will be made after this deadline.
3. **Academic honesty**: It is important that your course work represents only your ideas. I encourage discussion of homework in broad, conceptual terms where one student is trying to educate another without giving away the answer. Copying solutions or computing code from other students or other sources is plagiarism. At a minimum, all students involved will receive a 0 on the assignment in question for any type of academic dishonesty.
4. Resources for Disabled Students: Support and services are offered to student with functional limitations due to visual, hearing, learning, or mobility disabilities as well as to students who have specific chronic health conditions. See the Resources for Disabled Students web page for more information (rds.colostate.edu). If you need specific accommodations due to a disability, please meet with me outside of class to discuss your needs as early in the semester as possible. In accordance with RDS procedures, accomodations must be arranged in advance–no retroactive remedies are allowed.

**Textbook** (Chronically ordered and **recommended only**)

The Elements of Statistical Learning. Hastie, Tibshirani, and Friedman. Springer, 2009 (2nd Edition). 1st Edition is in 2001.

Statistics for High-Dimensional Data: Methods, Theory and Applications. Büuhlmann and Van de Geer. Springer, 2011.

Foundations of Machine Learning. Mohri, Rostamizadeh, and Talwalkar. The MIT Press, 2012.

Understanding Machine Learning: From Theory to Algorithms. Shalev-Shwartz and Ben-David. Cambridge University Press, 2014.

Statistical Learning with Sparsity: The Lasso and Generalizations. Hastie, Tibshirani, and Wainwright. Chapman and Hall/CRC, 2015.

Computer Age Statistical Inference: Algorithms, Evidence and Data Science. Efron and Hastie. Cambridge University Press, 2016.

An Introduction to Statistical Learning: with Applications in `R`. James, Witten, Hastie, and Tibshirani. Springer, 2017 (7th Edition). 1st Edition is in 2013.

Applied Predictive Modeling. Kuhn and Johnson. Springer, 2018 (2nd Edition). 1st Edition is in 2013.

High-Dimensional Probability: An Introduction with Applications in Data Science. Vershynin. Cambridge University Press, 2018.

High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Wainwright. Cambridge University Press, 2019.

**Disclaimer** The instructor reserves the right to make amendments to the syllabus and schedule as the semester develops. It is your responsibility to attend lectures and keep track of the proceedings.