

Theory of Statistical Data Mining (Statistics 461 Section 21)
 Instructor: Wenxin Jiang
 Office address: 2006 Sheridan Road Room 21
 Phone: 847-491-5081
 E-mail: wjiang@northwestern.edu
 Office Hours: TBA
 Lecture Time: 11am-12:20pm TTH in STAT basement classroom

COURSE DESCRIPTION: This is a sequel to Topics in Statistics - Data Mining (STAT 359, offered every other year).

The STAT 359 counterpart focuses on intuitive ideas and methodology.

The current STAT 461 course, on the other hand, focuses on theoretical framework, definitions, and rigorous mathematical proofs.

We will study the underlying theory and methods for modelling binary responses with multiple explanatory variables. Potential topics include: statistical decision theory, classification, approximation, consistency, mixtures of models, linear combinations, likelihood-based methods, Bayesian methods.

PREREQUISITES:

1. Courses in statistical theory and methodology comparable to STAT 420;
2. Topics in Statistics - Data Mining (STAT 359, offered every other year);
3. Knowledge of measure theoretic probability, real analysis, linear algebra and functional analysis.

TEACHING METHOD: Lectures.

EVALUATION: Homework, Final exam (TBA), and Presentation of a related paper.

READING:

1.(DGL). A Probabilistic Theory of Pattern Recognition
 by Luc Devroye, Lszl Gyrfi and Gbor Lugosi.
 Springer-Verlag, 1996, ISBN 0-387-94618-7

2.(HTF). The Elements of Statistical Learning:
 Data Mining, Inference, and Prediction
 by Trevor Hastie, Robert Tibshirani, Jerome Friedman
 Springer-Verlag, 2001, ISBN 0387952845

SCHEDULE: (Tentative.)

WEEKS	MATERIALS	REFERENCES
1-2	Bayes error nearest neighbor rules consistency slow rate of convergence	DGL Ch 2 Ch 5 Ch 6 Ch 7
3-4	error estimation Vapnik-Chervonenkis theory minimax lower bounds	Ch 8 Ch 12, 13 Ch 14
5-6	epsilon entropy uniform laws of large numbers approximation	Ch 28 Ch 29 Ch 30
7-8	maximum likelihood parametric classification Bayesian method variable selection	Ch 15 Ch 16 Ch 31.2, papers papers

*: Please talk to me about your topic by the end of the 7th week.
A list of papers will be given as potential topics.

HOMEWORK INSTRUCTIONS:

1. SOLVE AND REPORT

We do not have a teaching assistant for this course this quarter. Please grade your homeworks yourself and by each due date, please submit a report on the estimated percentage that you have solved for each problem.

E.g.,

I estimate that I have solved the following problems to the percentages reported below. Name _____ Date_____

Prob.2.7, 100%, Prob. 2.8, 70%, Prob. 2.9, 0%, (etc.)

2. DISCUSSIONS

Each week we will use about 15-30min classtime to discuss the homework problems. I will randomly select the students to describe their solutions in class. The higher the percentage you report for your solution, the more likely will you be selected to present it.

3. ADJUSTMENT

Your estimated percentage of soltution may be adjusted according to how well you explain your solution in class.