16:954:581: Probability and statistical inference for data science

Instructor:

Name: Ying Hung

Email: yhung@stat.rutgers.edu

Office: 463 Hill Center

Office hours: Wednesday 5:00 ~ 6:00 and by arrangement

TA: Linglin He, Email: lh511@rutgers.edu

TA Office hours: Thursday 10:30~11:30 and by arrangement, Hill Center 553

Lectures: Wednesday 6:40-9:30, SEC 203 BUS

Text/Reference:

1. *All of Statistics: A Concise Course in Statistical Inference*, by Larry A. Wasserman;

2. *Pattern Recognition and Machine Learning*, by Christopher M. Bishop.

Grading: Homeworks (35%), Midterm exam (30%), and Final exam (35%)

Emphasis and topics:

Probabilistic and inferential tools important for applications in data science. Topics covered in this course include:

a. Probability distributions. Important continuous and discrete distributions (Bernoulli, binomial, Poisson, Gaussian, chi-squared, t). Independence, joint, and conditional distributions. Expectation and variance. (2 weeks)

b. Decision theory. Minimizing expected loss, loss functions, classification, prediction. Bayes rules and Bayesian inference. (2 week)

c. Probability inequalites: Chebyshev's inequality, Chernoff inequality and the law of large numbers. (2 week)

d. Moment generating functions and the central limit theorem. Relationship to statistical inference. (2 weeks)

e. Point and interval estimation. Methods for constructing estimators: Method-of-moments and maximum likelihood. Asymptotic results: Consistency and asymptotic normality. (2 weeks)

f. Hypothesis testing, multiple testing and false discovery rates. Type I and type II error, power. Familywise error rate and false discovery rate. Benjamini-Hochberg procedure. (2 Weeks)

g. Linear methods for regression and classification. Regression: Least-squares, bias-variance decomposition, Bayesian linear regression. Classification: Fisher's linear discriminant. (1 weeks)

h. Non-linear methods for regression and classification. Kernel density estimators and nearest-neighbor classifiers. (1 weeks)