Stat 140B: Statistical Learning

California State University, Sacramento · Department of Mathematics & Statistics

CATALOG DESCRIPTION

Practical statistical modeling of data using professional-grade software. Principles, methodologies, and applications of supervised techniques such as classification, resampling, model selection, nonlinear regression, and tree-based methods; unsupervised techniques including clustering and dimension reduction. **3.0 Units; Letter Graded**

PREREQUISITES

Stat 140A

LEARNING OUTCOMES

Students will be able to:

1. Apply machine learning methods by selecting parameters and fitting models.
2. Create and interpret prediction intervals for new data using existing models.
3. Select appropriate supervised and unsupervised machine learning techniques given a dataset and research goal.
4. Evaluate model performance using standard methods such as cross validation and simulation.
5. Communicate results and justify methodology to a general audience.
6. Apply methodology to real world data using a high-level programming language such as R, Python, or Julia.

SAMPLE TEXT AND MATERIALS

Introduction to Statistical Learning (2nd edition) book by James, Witten, Hastie, and Tibshirani
https://www.statlearning.com

Technology

- Students will need a laptop or desktop computer to access the course materials online.
- The course will make extensive use of a high-level programming language such as R, Python, or Julia.

METHODS OF EVALUATION (GRADING SYSTEM, EXAMINATIONS, ETC)

The course grade will be determined by assignments, projects, midterms, and a comprehensive final examination.

Specific grading breakdown is instructor dependent. A sample breakdown is as follows:

- Homework 15%
- Project(s) 40%
- Midterm Exams 25%
- Final Exam 20%

COURSE OUTLINE

1. Regression models (1 week)
   a. Linear regression
   b. Logistic regression
2. Classification (2 weeks)
   a. Generative models
   b. Comparison of classification methods
   c. Generalized linear models
3. Resampling (2 weeks)
   a. Cross-validation
   b. Bootstrap
4. Linear model selection and regularization (3 weeks)
   a. Subset selection techniques
   b. Shrinkage methods
   c. Dimension reduction methods
5. Nonlinear models (2 weeks)
   a. Polynomial regression
   b. Step functions
   c. Basis functions
   d. Regression splines
6. Tree-based methods (2 weeks)
   a. Decision trees
   b. Random forests
   c. Bagging, boosting, and BART
7. Unsupervised learning (3 weeks)
   a. Principal components analysis
   b. Missing values
   c. Clustering methods