**Program Cover Document - STA 404: Computational and Bayesian Statistics**

**I. Basic Course Information**

Computational and Bayesian Statistics is primarily a junior/senior level course. The prerequisites are Probability (MAT 316), Analysis of Algorithms (CSC335), or Artificial Intelligence (CSC380); at least one 300-level Applied Statistics course (presently STA303, 304, 305, 306, 307, 314, 318), 300-level Math/Applied Math course, or 300-level Computer Science course, and a programming course (CRI215 or CSC220 or above). The course is designed to introduce students to modern computational methods of parameter estimation in multidimensional spaces and to Bayesian methods.

The study of statistics has been greatly impacted by the development of computational power over the last 20 years. The course will focus in two main areas. First, several key methods developed in classical statistics that rely on computational power will be introduced. Second, computational Bayesian statistical methods will be introduced. In both cases, students will learn the methods by solving problems from the point of view of classical statistics (confidence intervals, classification) and within scientific studies (spectral analysis, imaging, model selection).

The development of modern computing has led to the ability to utilize the previously intractable approach introduced by Bayes and Laplace and updated by Jaynes in the 20th century. The use of these methods however requires the combination of the concepts of conditional and marginal probability of distributions learned in MAT316 and substantial programming skills gathered in introductory programming courses and refined to the application in statistics or mathematics in 300-level courses.

*Course prerequisite*: STA215

 MAT316 or CSC335 or CSC380

CRI215 or CSC220 or CSC250

MAT3xx or STA3xx or CSC3xx

**II. Course Description**

This course will present an introduction to computational and Bayesian statistics. Topics include Bayesian computational methods, including the effect of prior knowledge in estimation and setting of confidence intervals, and the impact of computational methods on non-Bayesian methods. Substantial statistical programming will be involved.

**III. Learning Goals**

This course will expose students to several computational and Bayesian statistical methods. By the end of the course, the students will be able to:

1. Convert a physical model or observational data set to a computational statistical model
2. Estimate sampling and null distributions for any univariate statistic using bootstrap and permutation methods
3. Utilize techniques developed in the course to estimate parameters in low and high dimensional parameter spaces
4. Perform variable transformations to convert a physical model to a statistical model for the observations generated
5. Develop and code Markov chain Monte Carlo estimation methods for high dimensional parameter spaces
6. Understand and derive Cox’s rules of inference and distinguish Bayesian and LaPlacian methods of inference from those of Fisher and Pearson

Given that this is a 400-level course, students will improve their ability to work independently on multi-step problems, and students will read a number of the fundamental papers and reviews.

**IV. Learning Activities**

Learning activities may consist of a combination of lectures, group work, and significant review of programs. The specific choice will depend on the individual instructor. Outside of class, students are expected to do a significant amount of individual and group homework to achieve the learning goals, including creation of complex code for parameter estimation and visualization.

**V. Student Assessment**

Students will receive feedback on their work through either homework assignments, projects, and/or examinations.

**VI. List of Major Course Topics**

The following list of topics will be covered in the course. Items in the right column are optional, however it is expected that the instructor will choose at least two of these topics for inclusion in the course. Each topic involves both theoretical derivation and computational instantiation of the methods.

**Required topics Optional topics**

1. Estimation of Sampling Distributions

 a. Classical statistical examples and their limits

 b. The Bootstrap

 c. Empirical null distributions

2. Expectation-Maximization

 a. Two-Class estimation

 b. Sensitivity and Robustness in EM

 c. Bayesian EM

3. Rules of Inference

 a. Cox’s rules of inference

 b. Bayesian vs Frequentist concepts of inference

4. Bayes’ Equation

 a. Joint Probability to Bayes’ Equation

 b. Interpretation of Prior Probabilities

 c. The Log-Posterior Space and the Evidence

 d. Conjugate priors

 e. Philosophical motivation

5. Point Estimation in One and Two Dimensions

 a. Maximization of the Log-Posterior

 b. The Effects of Priors

 c. Taylor Series Approximation of the Log-Posterior

 d. Newton-Raphson method

6. Posterior Estimation

 a. Absolute Probability and Posterior Shape

 b. Monte Carlo Estimation

 c. Acceptance-Rejection Sampling

 d. Sampling Importance Resampling

 e. Integration with conjugates

7. Markov Chain Monte Carlo

 a. Discrete Markov Chains

 b. Steady States and Classification of State

 c. The Ergodic Theorem

 d. The Metropolis Algorithm

 e. The Metropolis-Hastings Algorithm

 f. Continuous Markov Chains

 g. Gibbs Sampling

 h. Variational Bayes

 i. Maximum Entropy

 j. Massive Inference

 k. Nested Sampling

 l. Reverse-Jump MCMC

**STA 404: Computational and Bayesian Statistics**

Course Syllabus, Fall 2017

**Textbook:** D. S. Sivia, Data Analysis: A Bayesian Tutorial, 2nd Edition

**Instructor:** Professor Michael Ochs

 Office: P246 Phone: 771-2189

 Email (preferred): ochsm@tcnj.edu

**Office Hours:** Wed/Fri 10:30 AM – 12:00 PM and Additional Hours as Requested

**Course Description:** This course will present an introduction to computational and Bayesian statistics. Bayesian methods have surged in recent years, as the advent of high-powered computational engines has enabled previously intractable calculation of posterior probability distributions. Bayesian computational methods are widely used in many fields of applied statistics and data mining due to their ability to model complex systems and allow incorporation of diverse knowledge in a comprehensive statistical framework.

**Course Philosophy:** The recovery of knowledge from data has been the focus of statistical study since the early days of the scientific revolution, demonstrated by the work of Bayes and Laplace. This course will focus on the revolution in statistics driven by the rediscovery of the Bayesian paradigm and the tremendous computational resources available to everyone. The focus will be on application of computational methods, including major developments in traditional statistics (Bootstrap, Expectation-Maximization) and Bayesian methods (Markov Chain Monte Carlo, Variational Bayes).

Computers have reduced the burden of “running the numbers” on analysts in all disciplines. We will use R in this course for problem sets and the final project, and students will be expected to integrate output into coherent summaries of analyses. However, the focus of the course will be on statistical concepts and all exams will focus on problems in theory and setting up calculations using programming. The application of statistical principles and the correct formulation of the problem mathematically and computationally are the critical aspects students should focus on.

**Evaluation:** Evaluation will be based on psets (30%), two exams during the semester (40%), and the final project (30%). **Students may work together on psets, however all psets must be written individually without copying each other’s work.** Also note that it will be statistically impossible to pass this course without doing sufficiently well on exams and the final project, so students should insure that they understand the problems when working in groups. **Each pset is due at the end of the day noted in the syllabus and must be uploaded to Canvas.** The final project will involve use of R in the development and analysis of a Bayesian statistical problem, which is to be summarized in a paper describing the analysis and conclusions in detail. Each student will work alone of this project. The final paper must be uploaded by the end of the final exam at TCNJ. Late psets will incur a 25% penalty and no credit will be given if psets are not uploaded by the day prior to the next class meeting.

**Schedule (Subject to Non-Random Variation)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Topic** | **Note** | **Book** | **Date** | **Topic** | **Note** | **Book** |
| 29-Aug | Statistics Review with R |  |  | 20-Oct | Marginalization | PSET Due | Ch 3 |
| 1-Sep | Statistics Review with R |  |  | 24-Oct | Error Estimation |  | Ch 3 |
| 5-Sep | MONDAY SCHEDULE - NO CLASS | PSET Due |  | 27-Oct | Error Propagation | PSET Due | Ch 3 |
| 8-Sep | The Bootstrap |  | Paper | 31-Oct | Model Selection |  | Ch 4 |
| 12-Sep | Empirical Null Distributions |  |  | 3-Nov | Model Selection | PEST Due | Ch 4 |
| 15-Sep | Outlier Statistics | PSET Due | Paper | 7-Nov | NO CLASS  |  |  |
| 19-Sep | Expectation Maximization (EM) |  | Paper | 10-Nov | Maximum Entropy | PSET Due | Ch 5 |
| 22-Sep | Expectation Maximization (EM) | PSET Due |  | 14-Nov | Maximum Entropy |  | Ch 5 |
| 26-Sep | Cox and the Rules for Inference |  | Ch 1 | 17-Nov | Markov Chains |  |  |
| 29-Sep | Hume, Bayes, and Kant | PSET Due | Paper | 21-Nov | EXAM 2 |  |  |
| 3-Oct | Prior Belief and Prior Distributions |  | Ch 2 | 24-Nov | Markov Chains |  |  |
| 6-Oct | Inference with Priors | PSET Due | Ch 2 | 28-Nov | THANKSGIVING BREAK |  |  |
| 10-Oct | FALL BREAK |  |  | 1-Dec | MCMC |  |  |
| 13-Oct | EXAM 1 |  |  | 5-Dec | Metropolis-Hastings | PSET Due | Ch 9 |
| 17-Oct | History: Conjugate Priors |  |  | 8-Dec | Variational Bayes |  | Ch 9 |

**Classroom Policies**

In this class, the deep learning outcomes associated with TCNJ’s 4th hour are accomplished by a series of rigorous educational assignments that extend beyond the typical scheduled class time. This includes development of skills in R programming and readings of the primary statistical literature, which students should expect to be quite demanding.

**Attendance**: All students are expected to attend all classes and are responsible for all information provided. A student who is absent for a test will not be permitted to make up the test unless prior arrangements with the instructor have been made. Approval for missing a test will only be permitted in exceptional circumstances. In the case of illness, a doctor’s note will be required. Please view TCNJ’s attendance policy at http://policies.tcnj.edu/policies/digest.php?docId=9134

**Academic Honesty**: Please make sure you are familiar with TCNJ’s academic integrity policy. Any suspected violation of this policy will be confronted in the strict accordance with the policy: http://policies.tcnj.edu/policies/digest.php?docId=7642

**Americans with Disability Act Policy:**

http://policies.tcnj.edu/policies/digest.php?docId=8082

**Final Exam-Evaluation-Reading Days Policy:**

http://policies.tcnj.edu/policies/digest.php?docId=9136

**LEARNING GOALS**

Upon completion of this course, students will have a thorough understanding of and ability to code

 The Bayesian Paradigm

 Calculational methods for estimating sampling distributions

 Permutation tests

 Monte Carlo methods

 Expectation Maximization

 Markov chains and Markov chain Monte Carlo

 Metropolis, Metropolis-Hastings, and Gibbs sampling methods

R will be introduced and used throughout this course. Students will also use R to generate figures and summary statistics to include in required reports on their analyses. As such, written presentation skills suitable to statistical work will be developed