Class Information

Course Title: Applied Data Mining (index#22052)
Economics 422/Section 01

Instructor: Dr. I-Ming Chiu
Office: ARMITAGE 328
Phone (856) 225 6012
E-mail address: ichiu@camden.rutgers.edu

Class Meeting: BSB 336. 9:30 AM-10:50 PM (Tuesday & Thursday)
Office Hours: 3:30-4:30 pm, Tuesday & Thursday or by appointment

Course Description: We are living in an era where new data are being fast produced from all kinds of human activities that include our daily emails, social network postings, and transactions on S&P 500 etc. How do we utilize these high-volume data? The answer is “Data Mining”. Data mining is a computing process of using mathematical and statistical algorithms to uncover patterns from huge data and convert them into decision-making information and new knowledge. A variety of data mining algorithms, supervised and unsupervised, will be introduced and the topics include linear regression model, logistic regression model, KNN, Support Vector Machine, Naïve Bayes, Association Rules, and K-Means Clustering, etc. Each topic covered is accompanied with a case study, so students will gain many hands-on learning experiences. The ultimate goal in this course is to equip students with modern data analytical tools, which has a high demand in the job market. Please click the links below to read the article titled “The Sexiest Job of 21st Century” and find out the potential reward for data science related jobs:
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm


(Electronic edition can be purchased at Amazon.com)
The course material will be drawn from various sources. The above two books cover most of the essential mining tools; the former focuses on the theory and the latter is more practical.

**Other References:**
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning/with Applications in R*, Springer 2013.


**Computing:**
All the computations will be done using an open source statistical software R. It can be downloaded at [http://www.r-project.org](http://www.r-project.org). You're encouraged to download and use RStudio at the following site, which is a very user-friendly IDE (integrated development environment) for R. [https://www.rstudio.com/products/rstudio/](https://www.rstudio.com/products/rstudio/)

**R Installation:** [https://www.youtube.com/watch?v=Icawuhf0Ygo](https://www.youtube.com/watch?v=Icawuhf0Ygo) (for Mac)
[https://www.youtube.com/watch?v=hxj0UG4boGU](https://www.youtube.com/watch?v=hxj0UG4boGU) (for PC)

**Class Material:** Data, handouts, assignments, and additional readings will be posted on Sakai website: [https://sakai.rutgers.edu/portal](https://sakai.rutgers.edu/portal).

**Online Learning:** [https://www.datacamp.com/](https://www.datacamp.com/) (Learn Data Science online)

**Useful Websites:** [http://www.statmethods.net/](http://www.statmethods.net/) (Computing using R web site)
[http://www.ats.ucla.edu/stat/](http://www.ats.ucla.edu/stat/) (Computing learning at UCLA)
[http://socserv.mcmaster.ca/jfox/](http://socserv.mcmaster.ca/jfox/) (Dr. Fox’s statistics site)
[https://www.analyticsvidhya.com/](https://www.analyticsvidhya.com/) (Machine Learning)

**Spring ’19 Calendar:** [https://registrar.camden.rutgers.edu/academic-calendar-2018-2019](https://registrar.camden.rutgers.edu/academic-calendar-2018-2019)
Grading: Contributions to Final Grade

- Attendance 5%
- Take-home problems 40%
- Midterm Exam 25%
- Final Project 30%
- Participation (extra credit) 5%

Grading Policy: Term grades will be based on the final distribution of the above grading weights.

Exam Preparation: The exam questions will be drawn from three sources: (i) homework assignments, (ii) course lectures, and (iii) reading material.

Class Participation: Class attendance is essential for learning achievement. When missing a class, it would cost you more time to learn on your own. I strongly recommend the following steps for your successful learning: (1) attend every class and take notes; (2) review everything you learn from the class immediately, never put it off; (3) ask questions and participate in class discussions.

Academic conduct: Make up exams will be given only upon prior notice. I request prior knowledge of any expected absence from an exam. If this is not feasible, you can document a valid reason for missing the exam. Unexcused absence on any exam will result in a grade of zero. Dishonesty in seeking an excused absence or in the examination process will result in a grade of zero on the exam involved and in university discipline. More detailed information can be found at the following site: https://academicintegrity.rutgers.edu.

Disability Services: Students with disabilities should contact the Rutgers-Camden Office of Disability Services (ODS). For more information, visit http://learn.camden.rutgers.edu/disability-services. Accommodation will be made in accordance with Rutgers University Policy.
Course Outline:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Mathematical and Statistical Fundamentals</td>
</tr>
<tr>
<td>Topic 2</td>
<td>Introduction to Computing Using R</td>
</tr>
<tr>
<td>Topic 3</td>
<td>What is Data Mining? The Familiar Linear Regression Model</td>
</tr>
<tr>
<td>Topic 4</td>
<td>K Nearest Neighbors (KNN)</td>
</tr>
<tr>
<td>Topic 5</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Topic 6</td>
<td>Decision Trees</td>
</tr>
<tr>
<td>Midterm Exam</td>
<td>Date: TBA in the class</td>
</tr>
<tr>
<td>Topic 7</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>Topic 8</td>
<td>Logistic Regression Model</td>
</tr>
<tr>
<td>Topic 9</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>Topic 10</td>
<td>Association Rules</td>
</tr>
<tr>
<td>Topic 11</td>
<td>Social Network Analysis</td>
</tr>
<tr>
<td>Topic 12</td>
<td>K-Means Clustering</td>
</tr>
<tr>
<td>Final Project</td>
<td>Details will be explained in the class meeting</td>
</tr>
</tbody>
</table>