**Syllabus**

**Fundamentals of Data Science**   
Fall 2017

590-05 and 590-12 (distance)

Thursdays, 3:05-5:50 p.m. Schiciano A

*(v. 8.28.17 – subject to modification)*

**Daniel Egger**

[Daniel.egger@duke.edu](mailto:Daniel.egger@duke.edu)

Office Hours

Monday 2:30-4:30 (“walk in” - no appointment needed)

3575 CIEMAS

Other meeting times by appointment

Teaching Assistants:

Chinmay Adjnadkar

[chinmay.ajnadkar@duke.edu](mailto:chinmay.ajnadkar@duke.edu)

Kevin Yu

[kefeng.yu@duke.edu](mailto:kefeng.yu@duke.edu)

Problem Sets to be uploaded to Sakai Dropbox before start of class when they are due. Because Problem Set solutions are discussed immediately in class, late work cannot be graded.

David J MacKay, *Information Theory, Inference, and Learning Algorithms*

Cambridge University Press (2003). Textbook is available for free online at:

<http://www.inference.phy.cam.ac.uk/itprnn/book.pdf>

\*Topics marked with an asterisk below indicate that Coursera video lectures and practice quizzes prepared by Egger are available to supplement in-class learning.

Course Structure: Seven weeks on Data Science Math, Methods and Concepts

Five Weeks on Practical Data Science Projects

Grading: Attendance, Class Participation and Problem Sets: 30%

Exam: 35%

Project (written, poster, and in-class presentation): 35%

Attendance is required. There will be some additional in-class assignments to be submitted by end of day on class days. Discuss any necessary absences with Professor well ahead of time to avoid embarrassment.

August 31 Course Overview

Review of Basic Probability Theory (Part 1)

*Definitions, Sum Rule, Joint Probability Rules, Venn Diagrams, Factorial, Permutations, Combinations and “M choose N,” and Dependence and Independence\**

September 7 Review of Basic Probability Theory (Part 2)

*Product Rule, Bayes’ Theorem, Inverse Probability, Binomial Theorem\**

*Common Probability Distributions*

*Bayes’ Theorem for Maximum Likelihood Estimation*

*Problem Set 1 Assigned*

*Recommended further reading in MacKay for basic and more advanced probability concepts:*

*Intro. pp. 1-2; sections 2.1-2.4 pp. 21-32; sections 3.1-3.6 pp. 47-64; section 21.1 pp. 293-295; sections 23.2-23.3 pp. 311-315; sections 28.1-28.2 pp. 343-351; Sections 37.1-37.5 pp. 457-466.*

September 14 New Concepts in Probability

*The Entropy of a Probability Distribution, Relative Entropy, Mutual Information and a New Definition of Dependence and Independence\**

*Recommended further reading in MacKay for information measures (entropy):*

*sections 2.4-2.9 pp. 32-40; sections 4.1-4.2 pp. 67-74;*

*sections 8.1-8.4 pp. 138-144; sections 18.3-18.5 pp. 265-268.*

September 21 Evaluating Binary Classification Models

*Performance Metrics, Cost Functions and the ROC Curve\**

*Problem Set 1 Due*

*Problem Set 2 Assigned*

September 28 Working with Gaussian Distributions

*The Central Limit Theorem, Linear Regression,*

*Root Mean Square Error in Parametric Models\**

October 5 Common Mistakes in Research & Modeling

*Understanding Sample Selection, P-values and Power, and Over-fitting.*

*Problem Set 2 Due*

*Problem Set 3 Assigned*

October 12 Directed Graph Models

*Networks, Page Rank, and Contagion*

October 19 Course Review and Practice Problems

*Problem Set 3 Due*

October 26 Exam

November 2 Introduction to Data Science Project Management

Credit Rating Practice Project\*

November 9 Python Week

November 16 Best Practices for Research Poster Design

*No class TH November 23 Thanksgiving Holiday*

November 30 Written Project Presentations Due

Individual meetings with faculty to review

December 7 Final In-Class Presentations

*END – Have a Happy Winter Holiday*