# Machine Learning I

## STAT 500

## Spring 2018

**TuTh 3:30 - 4:45**

**Location: CAE 204**

**INSTRUCTOR**: Dr. Ousley

**EMAIL**: sousley@mercyhurst.edu

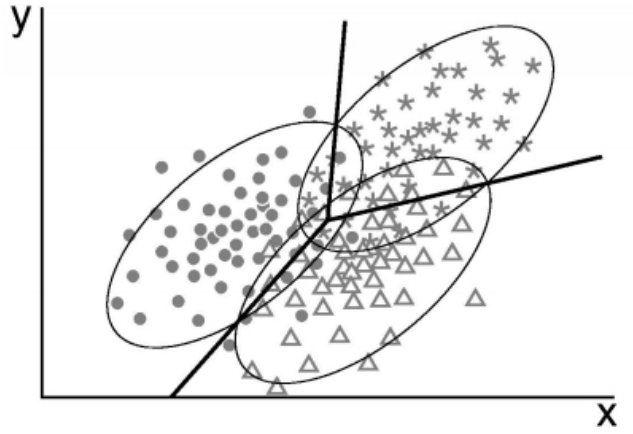
**PHONE**: 814-824-3116

**OFFICE**: Hammermill 413

**OFFICE HOURS**: Mo 5:00 - 6:00 PM

Tu 5:00 - 6:00 PM; We 5:00 - 6:00 PM

Fri 2:00 - 5:00 PM; and by appointment.




---

**COURSE DESCRIPTION:** This course covers Probability and Statistics, including univariate and basic multivariate analysis of data.

**LEARNING OUTCOMES:**
This course is designed to:

- Introduce basic statistical concepts and methods.
- Introduce resampling methods.
- Emphasize descriptive and predictive methods of statistics.
- Show how to use *R* and *R Studio* for data analysis.

**SPECIFIC COURSE OBJECTIVES:**
By the end of the course, students are expected to:

- Be quite familiar with univariate and bivariate data analysis, linear regression, and linear discriminant function analysis.
- Be quite familiar with using *R and RStudio* for statistical analysis.

**REQUIRED TEXTS:**
1. *Practical Statistics for Data Scientists*, by Bruce and Bruce (2017). O'Reilly Media; ISBN-10: 1491952962; ISBN-13: 978-1491952962.
2. Numerous PDFs from various sources

**OPTIONAL TEXTS:**

1. *Statistics: An Introduction Using R* (2015) by Crawley. Good coverage of many topics with some math.
2. *The R Cookbook, 2nd edition.* (2013) by Crawley. Encyclopedic coverage of using R in statistics. Some theory, little math.
3. *R for Dummies, 2nd edition* by de Vries and Meys (2015). Covers the use of R.
4. *R in Action, 2nd edition* by Kabacoff (2015). Many examples of using R for statistical analysis.

**SOFTWARE:**
We will use a software called *R*. *R* is a programming language for statistical computing and graphics. The latest version is 3.4.3 and it can be downloaded for free from http://www.r-project.org. We will use R Studio for regular classroom activities. The latest version of RStudio is 1.1.383 and is available at https://www.rstudio.com/products/rstudio/download/.
R studio is an open source Integrated development Environment (IDE) for R. You should install it on your personal computer and if it is a laptop, bring it to class or use the PCs in the classroom, which will have RStudio already. We may occasionally be using Python.

**COURSE POLICIES:**
***Course Website:*** The course website can be accessed through Blackboard. Course materials, announcements, and evaluations will all be posted on the Blackboard webpage. It is the student's responsibility to notify the instructor if there are any issues accessing the material. **Email will be used for the most important communications outside of class.** Although I will try to circulate all pertinent information via e-mail, you should check the course announcements prior to each class.

***Readings***: In this course there will be readings with machine learning concepts and algorithms. There will also be readings that are interactive involving using R and Python. The latter readings are closer to activities because you will get the most out of them if you read the code examples and immediately try them using R or Python. Ideally, you will read the concepts, then do the exercises related to the concepts, then do the homework. You can actually test ideas and ask "What if..." extensively in this class!

***Homework:*** Homework problems will be assigned during class. For optimal learning you need to master all the homework problems. I strongly encourage you to see me for help if you are unable to solve the assigned problems. The single most important part of this course is doing your homework, because homeworks are the keys to doing well on exams. The best way to do homework well is to keep up with the readings and practice on the class examples first.

***Classroom Accommodations:*** Any student who feels s/he may need an accommodation based on the impact of a disability should contact me privately to discuss your specific needs. Please know that it is the policy of Mercyhurst College that it is the student's responsibility to provide documentation of his/her disability to the director of the Learning Differences Program. Please call the Learning Differences office at 814-824-3017 to coordinate needed accommodations.

***Collaboration versus Cheating:*** Students may discuss homework and assignments in pairs or small groups, but all work must be individually written and all results and figures individually

generated. **You may give each other advice or help point out coding errors, but in the end each of you must do the work individually.** If a student copies text or graphs sent by another student into his or her own homework, both students will be cited for academic honesty violations.

***Academic Honesty:*** All students are expected to comply with Mercyhurst University policy on academic honesty (https://my.mercyhurst.edu/handbook/academic-affairs/ and http://www.mercyhurst.edu/academics/registrar's-office/academic-policies-procedures). **Academic dishonesty will not be tolerated.** This includes, but is not limited to, copying answers or providing others with answers to homeworks or during exams, plagiarizing information, or alteration of graded assignments. Any misconduct will, at the very least, result in a zero for the assignment, and if severe enough, will result in an F for the course. Any cheating on an exam will result in an F for the course.

**If you are behind on a homework it is far better to ask me for an extension and turn it in late, than to copy work from some else!**

## EVALUATION & GRADING:

Grades will be based on the following assessments:

| | | |
|---|---|---|
| Two Exams | 100 points | 200 points total |
| Homeworks | 10-50 points each | 400 points total |
| Class Participation | | 50 points total |

***Grading:*** An average score of at least 90 will result in an A for the class; 80 to 90 will be a B; 70 to 80 will be a C; 60 to 70 will be a D. An average score below 60 will result in an F.

## COURSE SCHEDULE:

**Some readings will be added or changed**!
Check Blackboard and class announcements for any updates!
PSDS: *Practical Statistics for Data Scientists*
Other readings are PDF files on Blackboard.

| Day | Topic | Reading |
|---|---|---|
| Jan 16 | Course Introduction | Ch1-McKillup; |
| | Statistics and Sampling | |
| 18 | Using R and RStudio | Ch1-Verzani-2011; Ch1-Verzani-2014; |
| | | |
| 23 | Data Types | Ch2-Verzani-2014:20-50; |
| | | Ch2-Lantz-2015; |
| 25 | Exploratory Data Analysis (EDA) | Ch2-Verzani-2014:70-80; |
| | Graphical Methods 1 | |

| | | |
|---|---|---|
| 30 | Graphical Methods 2 | Ch2-Verzani-2014:80-87; |
| Feb | | |
| 1 | Numerical Summaries of data | Ch1-PSDS; |
| | | Ch2-Verzani-2014:50-70; |
| 6 | Intro to (Discrete) Probability | CH4-Moore:pgs 231-258; |
| | | |
| 8 | The Normal distribution | CH6-Verzani-2014: p211-236; |
| | Probability distributions in R | CH2-PSDS; |
| 13 | Sampling and continuous probabilities | CH6-Verzani-2014: p227-231; |
| | CLT and Confidence Intervals | CH5-Crawley-2015; |
| 15 | Hypothesis tests | Ch3-PSDS; |
| | t-test, Chi-squared test | |
| 20 | **AAFS - no class** | |
| | | |
| 22 | **AAFS - no class** | |
| | | |
| 27 | Bivariate Data | PSDS:Correlation-end of Ch 1; |
| March | Correlation | Ch4-Moore; |
| 1 | Linear Regression I | CH6-Lantz; Ch7-Crawley 2015; |
| | | Ch4-PSDS; |
| 6 | **Mid-Semester Break** | |
| | | |
| 8 | **Mid-Semester Break** | |
| | | |
| 13 | Linear Regression II | CH5-Conway; |
| | **Midterm take-home Exam** | Osborne-2004; |
| 15 | Linear Regression III | Ch6-Everitt-2010; |
| | | |
| 20 | Multiple Regression | Ch4-PSDS; Ch6-Everitt-2010; |
| | | Ch28-Moore; |
| 22 | Outliers; Leverage; | |
| | | |
| 27 | Power | Ch12-Kabacoff-2010; |
| | Monte Carlo Tests | Ch16-Moore; |
| 29 | **Easter Break** | |
| April | | |
| 3 | Resampling and the bootstrap | Ch3-Rama; |
| | | |
| 5 | Regression and variable selection | CH34-UML; CH7-Sheather-2011; |
| | Regression and resampling | |
| 10 | **Advising Day - no classes** | |

| | | |
|---|---|---|
| | | |
| 12 | **AAPA - no class** | |
| | | |
| 17 | Principal Component Analysis | Ch16-Everitt-2010; Ch14-Kabacoff-2015; |
| | Transformations | CH39-UML; |
| 19 | Statistical Classification: LDFA I | Albrecht PDF; Ch5-PSDS; |
| | | CH14-Kachigan-1986; |
| 24 | LDFA II | Ousley PDFs; |
| | Fordisc | |
| 26 | Multivariate outliers, tests | Penny PDF; |
| May | Classification accuracy I | |
| 1 | Maximum Likelihood Estimation | CH7-Everitt-Hothorn |
| | Logistic Classification | CH13-Kabacoff-2015.pdf; Ch7-Everitt-2010; |
| 3 | Classification accuracy II | Ch7-Rama; CH210-UML; |
| | | |
| 8 | **FINAL take-home Exam** | |