IUPUI
SCHOOL OF INFORMATICS AND COMPUTING
Department of Human-Centered Computing

INFO-H 415
Introduction to Statistical Learning

Fall 2017

Section: *Credit Hours:* 3
Day and Time: Mondays, 6–8:40 pm
Class Location: IT 355
Instructor: William Fadel, Ph.D., Department of Biostatistics
Office Location: HITS 3000
Office Hours: By appointment (Phone: (317) 278-5420, Email: wffadel@iu.edu)

COURSE DESCRIPTION

This course applies statistical learning methods for data mining and inferential and predictive analytics to informatics-related fields. The course also introduces techniques for exploring and visualizing data, assessing model accuracy, and weighing the merits of different methods for a given real-world application. This course provides an essential toolset for transforming large, complex informatics datasets into actionable knowledge.
Prerequisites: ECON E270 or PBHL B300 or PSY B305 or SPEA K300 or STAT 30100 or STAT 35000

EXTENDED COURSE DESCRIPTION

This course applies statistical learning methods for data mining and inferential and predictive analytics to informatics-related fields. Supervised learning approaches include linear regression, logistic regression, linear discriminant analysis, resampling and shrinkage methods, splines and local regression, decision trees, bagging, random forests, boosting, and support vector machines. Unsupervised learning approaches include principal components analysis and *k*-means clustering. The course also covers techniques for exploring and visualizing data, assessing model accuracy, and weighing the merits of different methods for a given real-world application. This course is an essential toolset for transforming large, complex informatics datasets into actionable knowledge.

Required Textbooks:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer. ISBN 978-1-4614-7137-0 (Available for download at no cost from http://www-bcf.usc.edu/~gareth/ISL/.)

James and colleagues apply statistical learning methods to the following datasets:

- Automobile statistics (engineering)
- Housing values (business)
- Caravan insurance (business)

- Car seat sales (business)
- College tuition, demographics (education)
- Credit card default (business)
- Baseball hitters (physical education)
- Gene expression, 4 types of cancer (medicine)
- Gene expression, 64 cancer cell lines (medicine)
- Orange juice sales (business)
- Portfolio allocation (business)
- 5-year S&P 500 returns (business)
- US crime statistics (law)
- Central Atlantic income survey (business)
- 12-years of returns for 1089 stocks (business)

Matloff, N. (2011). *The art of R programming* (1st ed.). San Francisco, CA: No Starch Press. ISBN 978-1-59327-384-2 (Available for download at no cost from Books24x7 http://ulib.iupui.edu/resources/books.) To access, after clicking the link to the left, click E-Books portal, search for The Art of R Programming, and then proceed through either Books24x7 or Ebook Central. You may have to enter your credentials for Books24x7.

Recommended data analytics books using R:

This free book covers approximately same materials as the textbook; however, the explanations are not as clear and detailed:

Ledolter, J. (2013). *Data mining and business analytics with R.* Hoboken, NJ: Wiley. ISBN 978-1-118-44714-7

Ledolter applies statistical learning methods to the following datasets:

- Birth data (healthcare)
- Alumni donations (philanthropy)
- Orange juice sales (business)
- Prostate cancer (healthcare)
- Death penalty (law)
- Delayed airplanes (business)
- Loan acceptance (business)
- German credit data (business)
- Forensic glass (law, criminology)
- Fisher iris data (botany)
- MBA admission data (education)
- Motorcycle acceleration (engineering)
- Online radio (social computing)
- Predicting income (labor studies)

A comprehensive reference book on data analytics:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer. ISBN 978-0387848570

A concise introduction to data mining through the Rattle graphical user interface:

Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. New York: Springer. ISBN 978-1-4419-9889-7

This book focuses on the case studies listed below:

Torgo, L. (2011). *Data mining with R: Learning with case studies.* Boca Raton, FL: Chapman & Hall/CRC. ISBN 978-1-4398-1018-7

- Chapter 2: Predicting algae blooms (biology)
- Chapter 3: Predicting stock market returns (finance)
- Chapter 4: Detecting fraudulent transactions (computer security)
- Chapter 5: Classifying microarray samples (bioinformatics)

Conway, D., & White, J. M. (2012). *Machine learning for hackers.* Sebastopol, CA: O'Reilly. ISBN 978-1449303716

- Chapter 3: Bayesian spam classifier (informatics)
- Chapter 4: Priority inbox (informatics)
- Chapter 5: Predicting webpage views (informatics)
- Chapter 7: Code breaking (mathematics)
- Chapter 8: PCA: Market index (finance)
- Chapter 9: MDS: Senator similarity (political science)
- Chapter 10: kNN: Recommendation systems (informatics)
- Chapter 11: Social graphs (social computing)

Recommended books on statistics and R:

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage. ISBN 978-1-4462-0046-9 http://www.sagepub.com/books/Book236067

Venables, W. N., & Smith, D. N. (2009). *An introduction to R* (2nd ed). Godalming, UK: Network Theory. ISBN 978-0954612085 (Available for download at no cost from http://cran.r-project.org/doc/manuals/R-intro.pdf.)

Required Software:

- R (http://www.r-project.org), a language and environment for statistical computing and graphics. R supports all the data analytics methods covered in the course. R is flexible, fully programmable, and preferred by research statisticians, which means that new algorithms are often implemented in R first. Unlike most other powerful statistical packages, R is free of cost.
- Datasets and R code from http://www-bcf.usc.edu/~gareth/ISL/ used in James, Witten, Hastie and Tibshirani (2013) and "ISLR: Data for An Introduction to Statistical Learning with Applications in R." http://cran.r-project.org/web/packages/ISLR/index.html
- RStudio (http://www.rstudio.org), an integrated development environment (IDE) for R. RStudio is free of cost and operates on Windows, IOS (Mac), and Linux operating systems.

Optional Software:

- Rcommander (Rcmdr - http://www.rcommander.com), a graphical user interface (GUI) for statistics using R. Rcommander is free of cost and operates on Windows, IOS (Mac), and Linux operating systems.

- Rattle (http://rattle.togaware.com or http://cran.r-project.org/web/packages/rattle/), a GUI for data mining using R. Rattle is free of cost and operates on Windows, IOS (Mac), and Linux operating systems.

Required Hardware:
Students must install the required software on their notebook computers and bring their notebook computers to every class.

Principles of Graduate and Professional Learning (PGPL)

Learning outcomes are assessed in the following areas:

1. Knowledge and skills mastery	Major emphasis
2. Critical thinking and good judgment	Moderate emphasis
3. Effective communication	Some emphasis
4. Ethical behavior	

Learning Outcomes:

Upon completion of this course, students will	PLO	PLUS	Assessment
1. analyze datasets with the following supervised learning methods:	C1	P2.3, P3.1	
a. for functional approximation			
i. multiple linear regression,			L3 E3 Q3 M F
ii. splines, and			L7 E7 Q7 F
iii. local regression;			L7 E7 Q7 F
b. for classification			
i. logistic regression,			L4 E4 Q4 M F
ii. linear discriminant analysis,			L4 E4 Q4 M F
iii. decision trees, bagging, random forests, and boosting, and			L8 E8 Q8 F
iv. support vector machines;			L9 E9 Q9 F
2. analyze datasets with the following unsupervised learning methods:	C1	P2.3, P3.1	
a. for dimensionality reduction			
i. principal components analysis;			L6 E6 Q6 M F
b. for grouping			
i. *k*-means clustering and			L10 E10 Q10 F
ii. hierarchical clustering;			L10 E10 Q10 F
3. explore, transform, and visualize large, complex datasets with graphs in R;	C3	P1.4, P3.2	P L3–10 E3–10
4. solve real-world problems by adapting and applying statistical learning methods to large, complex datasets;	A3	P2.3	P L3–10 E3–10
5. identify and select appropriately among statistical learning methods for a particular real-world problem;	C2	P2.3	P F
a. analyze each method with respect to a given dataset or research question in terms of modeling accuracy and the bias-variance trade-off;			L5 E5 Q5 M F
b. perform model assessment (i.e., estimate test error rates) and selection by resampling:			
i. cross-validation and			L5 E5 Q5 M F
ii. bootstrapping;			L5 E5 Q5 M F
c. identify overfitting and underfitting;			L2 E2 Q2 M F
d. perform model selection and regularization by subset selection and shrinkage methods:			
i. ridge regression and			L6 E6 Q6 M F
ii. Lasso;			L6 E6 Q6 M F
e. explain the relative advantages and disadvantages of each statistical learning method for the real-world problem;			P F
6. write programs to perform data analytics on large, complex datasets in R.	B3	P3.2, P2.4	P L_R1–5, 7, 10
7. analyze data from case studies in informatics-related fields (e.g., digital media, human-computer interaction, health informatics, bioinformatics, and business intelligence).	B4	P3.1, P3.4	P E3–10

E: Exercise L (or L_R): Lab Exercise (in R) Q: Quiz M: Midterm F: Final P: Project

WEEKLY SCHEDULE

Lesson	Reading	Assessment
1	James et al. (2013). Chapter 1: Introduction James et al. (2013). Chapter 2: Statistical Learning	Quiz 1 Exercise 1 Lab1 Task: Install R and Rstudio
2	Matloff (2011). Chapter 1: Getting Started Matloff (2011). Chapter 2: Vectors	Lab 2 Quiz 2
3	Matloff (2011). Chapter 3: Matrices and Arrays Matloff (2011). Chapter 4: Lists	Lab 3
4	James et al. (2013). Chapter 3: Linear Regression	Lab 4 Exercise 2 Quiz 3
5	Matloff (2011). Chapter 5: Data Frames Matloff (2011). Chapter 10: Input/Output	Lab 5
6	James et al. (2013). Chapter 4: Classification	Lab 6 Exercise 3 Quiz 4
7	Matloff (2011). Chapter 7: R Programming Structures	Lab 7
8	James et al. (2013). Chapter 5: Resampling Methods	Lab 8 Exercise 4 Quiz 5
9	James et al. (2013). Chapter 6: Linear Model Selection and Regularization	Lab 9 Exercise 5 Quiz 6
10	Review James et al. (2013). Chapters 3–6.	Midterm
11	James et al. (2013). Chapter 7: Moving Beyond Linearity	Lab 10 Exercise 6 Quiz 7
12	James et al. (2013). Chapter 8: Tree-Based Methods	Lab 11 Exercise 7 Quiz 8
13	James et al. (2013). Chapter 9: Support Vector Machines	Lab 12 Exercise 8 Quiz 9
14	James et al. (2013). Chapter 10: Unsupervised Learning	Lab 13 Exercise 9 Quiz 10
15	Review James et al. (2013). Chapters 7–10.	Final Project

Assessments

Each student should not only read the assigned material but also arrive at a competent understanding of it prior to assessment. Four measures will be used to assess student-learning outcomes:

1. ***Labs*** provide students the opportunity to gain practical experience programming in R and performing data analytics in the R environment. Labs are assessed by the completion of lab exercises and/or by the creation of short programs that demonstrate skills employed in the lab exercises.

2. ***Exercises*** are selected among conceptual and applied problems at the end of Chapters 3 to 10 of James et al. (2013) and/or provided by the instructor on informatics and related applications.
3. ***Quizzes*** assess comprehension and skill acquisition. Quizzes are available in the learning management system (Canvas), and may quiz you either before or after you encounter material. Students work under a time limit and will be notified of errors after submission of the quiz. Points may be deducted from quizzes completed after the deadline. Each quiz remains open for review for one week after the corresponding lesson. After this time, the quiz will be closed, and submissions will not be possible. You will be given two chances for each quiz, and the highest score will be kept as your grade.
4. The ***midterm*** and ***final*** have a format resembling the exercises and quizzes. The midterm covers lessons 1 to 9. The final focuses on lessons 11 to 14 but includes some material from lessons 1 to 9. It is understood that each chapter builds on all previous chapters; thus, acquired knowledge and skills need to be maintained throughout the course.
5. The ***project*** affords the opportunity for students to apply data analytics methods to a dataset of their choice. Students may select among datasets from the recommended readings or their own dataset. The project includes analyzing, visualizing, and interpreting the dataset and writing up the results in a report format. It may also include a presentation at the instructor's discretion. A project may be completed by an individual or by a group of up to three students. Group projects should include a more comprehensive analysis of the dataset and should document the percentage contribution of each group member to each part of the project.

Grade Calculation:

1.	Labs (13)	10%
2.	Exercises (9)	40%
3.	Quizzes (10)	20%
4.	Midterm (1)	10%
5.	Final (1)	10%
6.	Project (1)	5%

Grading Scale:

A+	97 – 100	Outstanding achievement, given at the instructor's discretion
A	93 – 100	Excellent achievement
A–	90 – 92.99	Very good work
B+	87 – 89.99	Good work
B	83 – 86.99	Marginal work
B–	80 – 82.99	Very marginal work
C+	77 – 79.99	Unacceptable work (Course must be repeated)
C	73 – 76.99	Unacceptable work (Course must be repeated)
C–	70 – 72.99	Unacceptable work (Course must be repeated)
D+	67 – 69.99	Unacceptable work (Course must be repeated)
D	63 – 66.99	Unacceptable work (Course must be repeated)
D–	60 – 62.99	Unacceptable work (Course must be repeated)
F	Below 60	Unacceptable work (Course must be repeated)

EXPECTATIONS, GUIDELINES, AND POLICIES

Attendance:

A basic requirement of this course is that you will participate in all class meetings, whether online or face-to-face, and conscientiously complete all required course activities and assignments. Class attendance is required for classroom-based courses. It entails being present and attentive for the entire class period. Attendance shall be taken in every class. If you do not sign the attendance sheet while in class, you shall be marked absent. Signing the attendance sheet for another student is prohibited. The instructor is required to submit to the Registrar a record of student attendance, and action shall be taken if the record conveys a trend of absenteeism.

Only the following are acceptable excuses for absences: death in the immediate family (e.g. mother, father, spouse, child, or sibling), hospitalization or serious illness; jury duty; court ordered summons; religious holiday; university/school coordinated athletic or scholastic activities; an unanticipated event that would cause attendance to result in substantial hardship to one's self or immediate family. Absences must be explained with the submission of appropriate documentation to the satisfaction of the instructor, who will decide whether missed work may be made up. Absences that do not satisfy the above criteria are considered unexcused. To protect your privacy, doctor's excuses should exclude the nature of the condition and focus instead on how the condition impacts your attendance and academic performance.

Missing class reduces your grade through the following grade reduction policy: You are allowed two excused or unexcused absences. Each additional absence, unless excused, results in a 5% reduction in your final course grade. More than four absences result in an F in the course. Missing class may also reduce your grade by eliminating opportunities for class participation. For all absences, the student is responsible for all covered materials and assignments.

Incomplete:

The instructor may assign an Incomplete (I) grade only if at least 50% of the required coursework has been completed at passing quality and holding you to previously established time limits would result in unjust hardship to you. All unfinished work must be completed by the date set by the instructor. Left unchanged, an Incomplete automatically becomes an F after one year. http://registrar.iupui.edu/incomp.html

Deliverables:

You are responsible for completing each deliverable (e.g., assignment, quiz) by its deadline and submitting it by the specified method. Deadlines are outlined in the syllabus or in supplementary documents accessible through the learning management system (e.g., Canvas). Should you miss a class, you are still responsible for completing the deliverable and for finding out what was covered in class, including any new or modified deliverable. In fairness to the instructor and students who completed their work on time, a grade on a deliverable shall be reduced 10%, if it is submitted late and a further 10% for each 24-hour period it is submitted after the deadline, i.e. 1 day late, 10%, 2 days late, 20%, 3 days late, 30%, etc.

CODE OF CONDUCT

All students should aspire to the highest standards of academic integrity. Using another student's work on an assignment, cheating on a test, not quoting or citing references correctly, or any other form of dishonesty or plagiarism shall result in a grade of zero on the item and possibly an F in the course. Incidences of academic misconduct shall be referred to the Department Chair and repeated violations shall result in dismissal from the program.

All students are responsible for reading, understanding, and applying the *Code of Student Rights,*

Responsibilities and Conduct and in particular the section on academic misconduct. Refer to *The Code > Responsibilities > Academic Misconduct* at http://www.indiana.edu/~code/. All students must also successfully complete the Indiana University Department of Education "How to Recognize Plagiarism" Tutorial and Test. https://www.indiana.edu/~istd You must document the difference between your writing and that of others. Use quotation marks in addition to a citation, page number, and reference whenever writing someone else's words (e.g., following the *Publication Manual of the American Psychological Association*). To detect plagiarism, instructors apply a range of methods, including Turnitin.com. https://citl.indiana.edu/teaching-resources/academic-integrity/turnitin/

Academic Misconduct:

1. **Cheating:** Cheating is considered to be an attempt to use or provide unauthorized assistance, materials, information, or study aids in any form and in any academic exercise or environment.
   a. A student must not use external assistance on any "in-class" or "take-home" examination, unless the instructor specifically has authorized external assistance. This prohibition includes, but is not limited to, the use of tutors, books, notes, calculators, computers, and wireless communication devices.
   b. A student must not use another person as a substitute in the taking of an examination or quiz, nor allow other persons to conduct research or to prepare work, without advanced authorization from the instructor to whom the work is being submitted.
   c. A student must not use materials from a commercial term paper company, files of papers prepared by other persons, or submit documents found on the Internet.
   d. A student must not collaborate with other persons on a particular project and submit a copy of a written report that is represented explicitly or implicitly as the student's individual work.
   e. A student must not use any unauthorized assistance in a laboratory, at a computer terminal, or on fieldwork.
   f. A student must not steal examinations or other course materials, including but not limited to, physical copies and photographic or electronic images.
   g. A student must not submit substantial portions of the same academic work for credit or honors more than once without permission of the instructor or program to whom the work is being submitted.
   h. A student must not, without authorization, alter a grade or score in any way, nor alter answers on a returned exam or assignment for credit.
2. **Fabrication:** A student must not falsify or invent any information or data in an academic exercise including, but not limited to, records or reports, laboratory results, and citation to the sources of information.
3. **Plagiarism:** Plagiarism is defined as presenting someone else's work, including the work of other students, as one's own. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged, unless the information is common knowledge. What is considered "common knowledge" may differ from course to course.
   a. A student must not adopt or reproduce ideas, opinions, theories, formulas, graphics, or pictures of another person without acknowledgment.
   b. A student must give credit to the originality of others and acknowledge indebtedness whenever:

1. directly quoting another person's actual words, whether oral or written;
2. using another person's ideas, opinions, or theories;
3. paraphrasing the words, ideas, opinions, or theories of others, whether oral or written;
4. borrowing facts, statistics, or illustrative material; or
5. offering materials assembled or collected by others in the form of projects or collections without acknowledgment

4. **Interference:** A student must not steal, change, destroy, or impede another student's work, nor should the student unjustly attempt, through a bribe, a promise of favors or threats, to affect any student's grade or the evaluation of academic performance. Impeding another student's work includes, but is not limited to, the theft, defacement, or mutilation of resources to deprive others of the information they contain.
5. **Violation of Course Rules:** A student must not violate course rules established by a department, the course syllabus, verbal or written instructions, or the course materials that are rationally related to the content of the course or to the enhancement of the learning process in the course.
6. **Facilitating Academic Dishonesty:** A student must not intentionally or knowingly help or attempt to help another student to commit an act of academic misconduct, nor allow another student to use his or her work or resources to commit an act of misconduct.

OTHER POLICIES

1. **Right to revise:** The instructor reserves the right to make changes to this syllabus as necessary and, in such an event, will notify students of the changes immediately.
2. **IUPUI course policies:** Several campus policies governing IUPUI courses may be found at the following link: http://registrar.iupui.edu/course_policies.html
3. **Classroom civility**: To maintain an effective and inclusive learning environment, it is important to be an attentive and respectful participant in lectures, discussions, group work, and other classroom exercises. Thus, unnecessary disruptions should be avoided, such as ringing cell phones engagement in private conversations and other unrelated activities. Cell phones, media players, or any noisy devices should be turned off during a class. Texting, surfing the Internet, and posting to Facebook or Twitter during class are generally not permitted. Laptop use may be permitted if it is used for taking notes or conducting class activities. Students should check with the instructor about permissible devices in class. IUPUI nurtures and promotes "a campus climate that seeks, values, and cultivates diversity in all of its forms and that provides conditions necessary for all campus community members to feel welcomed, supported, included, and valued" (IUPUI Strategic Initiative 9). IUPUI prohibits "discrimination against anyone for reasons of race, color, religion, national origin, sex, sexual orientation, marital status, age, disability, or [veteran] status" (Office of Equal Opportunity). Profanity or derogatory comments about the instructor, fellow students, invited speakers or other classroom visitors, or any members of the campus community shall not be tolerated. A violation of this rule shall result in a warning and, if the offense continues, possible disciplinary action.
4. **Bringing children to class:** To ensure an effective learning environment, children are not permitted to attend class with their parents, guardians, or childcare providers.
5. **Email:** Indiana University uses your IU email account as an official means of communication, and students should check it daily for pertinent information. Although you may have your IU

email forwarded to an outside email account, please email faculty and staff from your IU email account.

6. **Disabilities Policy:** In compliance with the Americans with Disabilities Act (ADA), all qualified students enrolled in this course are entitled to reasonable accommodations. Please notify the instructor during the first week of class of accommodations needed for the course. Students requiring accommodations because of a disability must register with Adaptive Educational Services (AES) and complete the appropriate AES-issued before receiving accommodations. The AES office is located at UC 100, Taylor Hall (Email: aes@iupui.edu, Tel. 317 274-3241). Visit http://aes.iupui.edu for more information.
7. **Administrative Withdrawal:** A basic requirement of this course is that students participate in all class discussions and conscientiously complete all required course activities and/or assignments. If a student is unable to attend, participate in, or complete an assignment on time, it is the student's responsibility to inform the instructor. If a student misses more than half of the required activities (such as exercises/quizzes/labs) within the first 25% of the course without contacting the instructor, the student may be administratively withdrawn from this course. Administrative withdrawal may have academic, financial, and financial aid implications. Administrative withdrawal will take place after the full refund period, and a student who has been administratively withdrawn from a course is ineligible for a tuition refund. Contact the instructor with questions concerning administrative withdrawal.
8. **Emergency Preparedness:** Safety on campus is everyone's responsibility. Know what to do in an emergency so that you can protect yourself and others. For specific information, visit the emergency management website. http://protect.iu.edu/emergency

MISSION STATEMENT

The Mission of IUPUI is to provide for its constituents excellence in

- Teaching and Learning;
- Research, Scholarship, and Creative Activity; and
- Civic Engagement.

With each of these core activities characterized by

- Collaboration within and across disciplines and with the community;
- A commitment to ensuring diversity; and
- Pursuit of best practices.

IUPUI's mission is derived from and aligned with the principal components—Communities of Learning, Responsibilities of Excellence, Accountability and Best Practices—of Indiana University's Strategic Directions Charter.

STATEMENT OF VALUES

IUPUI values the commitment of students to learning; of faculty to the highest standards of teaching, scholarship, and service; and of staff to the highest standards of service. IUPUI recognizes students as partners in learning. IUPUI values the opportunities afforded by its location in Indiana's capital city and is committed to serving the needs of its community. Thus, IUPUI students, faculty, and staff are involved in the community, both to provide educational programs and patient care and to apply learning to community needs through service. As a leader in fostering collaborative relationships, IUPUI values collegiality, cooperation, creativity, innovation, and entrepreneurship as well as honesty, integrity, and support for open inquiry and dissemination of

findings. IUPUI is committed to the personal and professional development of its students, faculty, and staff and to continuous improvement of its programs and services.