RIT - GCCIS
Syllabus

ISTE.780.01
Data-Driven Knowledge Discovery

Summer 2018 (Term 2178)
Draft of May 26, 2018

Details

Important note: The information presented in this syllabus is subject to expansion, contraction, change, or stasis during the semester. In case of conflict between versions, the copy on myCourses takes precedence.

Course Number. 80818

Prerequisites.

- ISTE-600 (Analytical Thinking) or equivalent
- PSYC-640 (Statistics) or equivalent

Time. Online

Place. Online

Dates. 17 May 2018–8 August 2018

Final Exam. Online, Aug 10th–14th. Time to be announced.

Instructor. Mick McQuaid

Email. mjmics@rit.edu

Office. 70-2675

Office Hours. 1300–1400 M,W,F or by appointment

Description

Rapidly expanding collections of data from all areas of society are becoming available in digital form. Computer-based methods are available to facilitate discovering new information and knowledge that is embedded in these collections of data. This course provides students with an introduction to the use of these data analytic methods, especially statistical learning approaches, within the context of the data-driven knowledge discovery process.

Materials

Each student will need a computing environment capable of running the R programming language. A minimal Linux, Mac, or Windows computer should suffice. R will be taught as part of this course. Students should have proficiency with some full-featured text editor. A network connection capable of streaming video is essential for this online course.

The textbook for this course is *An Introduction to Statistical Learning with Applications in R*, James et al. (2013). This book is available free online and in Wallace Library. We will also review three award-winning papers from SIGKDD 2017: Hill et al. (2017), Hou et al. (2017), and Vandal et al. (2017).

Schedule

Arabic numerals refer to days. The course runs for twenty-two days, Tuesday and Thursday for twelve weeks (Roman numerals), except the first week when the course begins on Thursday and the last week which ends on Tuesday.
Week I

1. The textbook — Chapter 1 (pp. 1–14) — Supervised and unsupervised learning — Prediction, inference, and description — Linear and non-linear models — Linear algebra — Vectors and matrices — n and p — The R language — Object orientation — The virtual machine — Examples — Grading — Individual work

Week II

2. Chapter 2 (pp. 15–28) — Input and output, predictor and response, independent and dependent variable, X_i and Y — A basic model $Y = f(X) + \epsilon$ — Prediction, inference, and description — Irreducible and reducible error — Notation for expected value of error $E(Y - \widehat{Y})^2$ and variance of error — Examples — Estimating f as $\hat{f}$ — Training data — Parametric methods — Estimating parameters — Non-parametric methods — Accuracy, flexibility, and interpretability — Supervised and unsupervised learning — Regression and classification — Exercises 2.4.1, 2.4.2

3. Chapter 2 (pp. 29–42) — Measuring the quality of fit — Mean squared error — Training and test data — Flexibility or degrees of freedom — Overfitting — Bias, error and variance — Classification — Indicator variables — Training error rate — Test error rate — Bayes classifier — Conditional probability — Bayes decision boundary — Bayes error rate — K-nearest neighbors or KNN — Exercise 2.4.9

Week III

4. Chapter 2 (pp. 42–57) — Introduction to R — Basic R commands — R graphics — Indexing data — Loading data — Graphical summaries of data — Numerical summaries of data — To turn in I: Exercise 2.4.8

5. Chapter 3 (pp. 59–101) — Linear Regression — Estimating coefficients — Assessing the accuracy of the coefficients — Assessing the accuracy of the model — Multiple linear regression — Estimating multiple regression coefficients — Questions about multiple linear regression — Qualitative predictors — Extensions of the linear model — Potential problems with the linear model — Exercise 3.7.8

Week IV

6. Chapter 3 (pp. 102–126) — An extended example of multiple linear regression — Lab on linear regression — Libraries in R for linear regression — Simple linear regression in R — Multiple linear regression in R — Interaction terms in R — Nonlinear transformations of predictors — Qualitative predictors — Writing functions — To turn in II: Exercise 3.7.9

7. Chapter 4 (pp. 128–154) — Classification — Overview — Alternative to linear regression — Logistic regression — Logistic model — Estimating logistic regression coefficients — Making predictions — Multiple logistic regression — Logistic regression with more than two response classes — Linear discriminant analysis — Using Bayes' theorem for classification — Linear discriminant analysis in one dimension — Linear discriminant analysis in more than one dimension — Quadratic discriminant analysis — Comparing classification methods — Exercise 4.7.1

Week V

8. Chapter 4 (pp. 154–174) — Lab on logistic regression, LDA, QDA, and KNN in R — Stock market data — Logistic regression in R — Linear discriminant analysis in R — Quadratic discriminant analysis in R — K-nearest neighbors in R — Application to caravan insurance data — To turn in III: Exercise 4.7.11

9. Chapter 5 (pp. 175–190) — Resampling — Cross-validation — Validation set approach — Leave-one-out cross-validation — k-fold cross-validation — Bias variance trade-off — Cross-validation on classification problems — Bootstrap — Exercise 5.4.5

Week VI

10. Chapter 5 (pp. 190–202) — Lab: Cross-validation and bootstrap in R — Validation set approach in R — Leave-one-out cross-validation in R — k-fold cross-validation in R — Bootstrap in R — To turn in IV: Exercise 5.4.6

11. Chapter 6 (pp. 203–244) — Linear model selection — Subset selection — Stepwise selection — Choosing the optimal model — Shrinkage methods — Ridge regression — Lasso — Selecting the tuning parameter — Dimension reduction methods — Principal components regression — Partial least squares — Considerations in high dimensions — High-dimensional data — What can go wrong in high dimensions — Regression in high dimensions — Interpreting results in high dimensions — Exercise 6.8.8

Week VII

12. Chapter 6 (pp. 244–264) — Lab: Subset selection in methods in R — Forward and backward stepwise selection in R — Choosing among models using the validation set approach and cross-validation in R — Lab: Ridge regression and the lasso — Ridge regression in R — Lasso in R — Lab: Principal components regression and partial least squares in R — Principal components regression in R — Partial least squares in R — To turn in V: Exercise 6.8.10

13. Chapter 7 (pp. 265–280) — Beyond linearity — Polynomial regression — Step functions — Basis functions — Regression splines — Piecewise polynomials — Constraints and splines — Spline basis representation — Choosing the number and locations of knots — Comparison between regression splines and polynomial regression — Smoothing splines — Overview of smoothing splines — Choosing the smoothing parameter λ — Exercise 7.9.6

Week VIII

14. Chapter 7 (pp. 280–301) — Local regression — Generalized additive models — GAMs for regression problems — GAMs for classification problems — Lab: non-linear modeling — Polynomial regression and step functions in R — Splines in R — GAMs in R — To turn in VI: Exercise 7.9.7

15. Chapter 8 (pp. 303–323) — Tree-based methods — Basics of decision trees — Regression trees — Classification trees — Trees versus linear models — Advantages and disadvantages of trees — Bagging, random forests, boosting — Bagging — Random forests — Boosting — Exercise 8.4.3

Week IX

16. Chapter 8 (pp. 323–336) — Lab: Decision trees — Fitting classification trees in R — Fitting regression trees in R — Bagging and random forests in R — Boosting in R — To turn in VII: Exercise 8.4.8

17. Chapter 9 (pp. 337–354) — Support vector machines — Maximal margin classifier overview — Definition of a hyperplane — Classification using a separating hyperplane — Maximal margin classifier — Constructing the maximal margin classifier — Non-separable case — Support vector classifiers — Overview of the support vector classifier — Details of the support vector classifier — Support vector machine overview — Classification with non-linear decision boundaries — Support vector machines — Application to the heart disease data — Exercise 9.7.4

Week X

18. Chapter 9 (pp. 355–372) — Support vector machines with more than two classes — One-versus-one classification — One-versus-all classification — Lab: Support vector machines — Support vector classifier in R — Support vector machine

in R — ROC curves in R — SVM with multiple classes in R — Application to gene expression data — To turn in VIII: Exercise 9.7.5

19. Chapter 10 (pp. 373–400) — Unsupervised learning — Challenge of unsupervised learning — Principal components analysis — Defining principal components — Alternative interpretation of principal components — Details of principal component analysis — Other uses for principal components — Clustering methods — K-means clustering — Hierarchical clustering — Practical issues in clustering — Exercise 10.7.9

Week XI

20. Chapter 10 (pp. 401–413) — Lab: Principal components analysis — Lab: Clustering — K-means clustering in R — Hierarchical clustering in R — Lab: NCI60 data example — Principal components analysis on the NCI60 data — Clustering the observations of the NCI60 data — To turn in IX: Exercise 10.7.10

21. Reading from SIGKDD 2017: Vandal et al. (2017)

Week XII

22. Reading from SIGKDD 2017: Hou et al. (2017) — To turn in X: Online Discussion Contribution from reading Vandal et al. (2017)

23. Reading from SIGKDD 2017: Hill et al. (2017)

Week XIII

24. Summary / Review

Grading

The grading scale used along with the grade components follow.

- A >= 90.0%
- B >= 80.0% & < 90%
- C >= 70.0% & < 80.0%
- D >= 60.0% & < 70.0%
- F < 60.0%

The course grade is composed of 6 percent for each of the ten homeworks and 40 percent for the final exam, for a total of 100 percent. Homeworks will be graded strictly and with an attention to detail. Instructions will be provided that must be followed carefully to expect a passing grade.

Policies

The following are brief statements of policy that are, in many places, expanded at the URLs provided. You are bound by these policies and any protest that you did not read the extended versions at the provided links will fall on deaf ears. Your familiarity with the following policies, dates, and and parameters will be assumed in this course.

Last day of 7-day add/drop period. Thursday 24 May 2018

Last day to withdraw with W. Wednesday 25 July 2018

myCourses. All project assignments, lecture notes, and other distributable course materials will be available via myCourses. Except where otherwise indicated, all student project assignments will be submitted via myCourses dropboxes.

Grade Challenges. IST department policy states that a student has one semester to challenge any grade. After that, grades cannot be challenged.

Late Work. Any work not submitted by the final due date receives a grade of zero unless arrangements are made previous to the initial due date.

Extra Credit. No extra credit is available in this course.

Accommodations. If you have a "Notice of Accommodation", you must provide your instructor with a copy of it within 1 week of starting this course. You must follow all the rules of the relevant office.

Academic Dishonesty. The policy on dishonesty is simple: Anyone caught cheating receives an "F" as a course grade, is removed from the section and a letter detailing the incident is placed into his or her folder. Any student accused of cheating should realize that the evidence has already been verified by other faculty members and will withstand an appeal. Additionally, please review the institute policy at http://www.rit.edu/studentaffairs/studentconduct/rr_academicdishonesty.php

Acceptable Use. We are bound by the following Acceptable Computer Use policy at http://www.rit.edu/academicaffairs/policiesmanual/sectionC/C82.html

Student Responsibilities. Please review the general student responsibilities as outlined at http://www.rit.edu/~301www/rr.php3

Policy on Reporting Incidents of Discrimination and Harassment. RIT is committed to providing a safe learning environment, free of harassment and discrimination as articulated in our university policies located on our governance website. RIT's policies ***require faculty to share information*** about incidents of gender based discrimination and harassment with RIT's Title IX coordinator or deputy coordinators, regardless whether the incidents are stated to them in person or shared by students as part of their coursework. RIT Governance website: https://www.rit.edu/academicaffairs/policiesmanual/policies/governance

If you have a concern related to gender-based discrimination and/or harassment and prefer to have a ***confidential*** discussion, assistance is available from one of RIT's confidential resources on campus:

1. The Center for Women & Gender: Campus Center Room

1760; 585-475-7464; CARES (available 24 hours/7 days a week) Call or text 585-295-3533.

2. RIT Student Health Center – August Health Center/1st floor; 585-475-2255.
3. RIT Counseling Center - August Health Center /2nd floor - 2100; 585-475-2261.
4. The Ombuds Office – Student Auxiliary Union/Room 1114; 585-475-7200 or 585-475-2876.
5. The Center for Religious Life – Schmitt Interfaith Center / Rm 1400; 585-475-2137.
6. NTID Counseling & Academic Advising Services – 2nd Floor Lyndon B. Johnson; 585-475-6468 (v), 585-286-4070 (vp).

References

Hill, Daniel N., Houssam Nassif, Yi Liu, Anand Iyer, and S.V.N. Vishwanathan. 2017. "An Efficient Bandit Algorithm for Realtime Multivariate Optimization." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17*. ACM Press. https://doi.org/10.1145/3097983.3098184.

Hou, Shifu, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. "HinDroid." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17*. ACM Press. https://doi.org/10.1145/3097983.3098026.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Vandal, Thomas, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R. Ganguly. 2017. "DeepSD." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17*. ACM Press. https://doi.org/10.1145/3097983.3098004.