Creating an undergraduate course for Principles of Data Science

Sungkyu Jung

Feb 2, 2018

Data Science

- Data science is an emerging interdisciplinary field stemming from statistics, mathematics and computer science.
- Using automated methods to analyze massive amounts of data and to extract knowledge from them
- Incorporating inferential and computational thinking, and data ethics in practice of data science

Data Science at Pitt

- Existing programs
  - Former School of Information Sciences (now SCI): post-bachelor's and post-master's certificate
  - College of Business Administration: Certificate in Business Analytics
- Efforts for graduate and undergraduate programs data science major

Proposal: A new course STAT1261: Principles of Data Science

- Fundamental pipeline of data science, ranging from data acquisition, data clean-up, data exploration and visualization, modeling and inference, ethics in dealing with data, to professional reporting.
- Designed to be the first half of a data-science sequence (with 1361: Statistical Learning), offered from Department of Statistics
- Offered in Fall 2017
- 50 undergraduate students enrolled (mostly in their 3-4th year)
- Several graduate students audited

Course Transformation

Proposed activities

1. Reinforced learning by recitation/lab activities
2. Incorporating new technologies
3. Longitudinal survey
4. Teaching material for new instructors

Assessment

1. End-of-semester survey
2. Focus group
3. A year-long study on the students' performance through STAT1361

1. Reinforced learning by recitation/lab activities

- Hands-on experience is central in learning data science
- A typical week consists of
  - two lectures (Monday-Wednesday)
  - One-hour lab (Friday)
  - Homework extending lab activities

Lab activity example (Lab 4)

Lectures

- Dissecting data graphics
- Principles of building graphics

Lab

Lab4.html

2. Incorporating new technologies

Goals:

- Lynda.com videos to supplement learning
- Online forum for feedback to students and communications among students

Piazza forum

- Piazza (URL https://piazza.com), a Q&A web service
  - Anonymous and instantaneous communications
  - Incentives to promote participation

https://piazza.com/class/j6pvkicfewy20n?cid=27

See an example post and statistics.

3. Longitudinal survey

Initially planned

- to do using typeform.com software,
- to use their api to retrieve data,
- and to analyze the data

I gave up after first survey

- longitudinal analysis too advanced for students
- data type through api too complex

4. Teaching material for new instructors

Most material on the web:
http://www.stat.pitt.edu/sungkyu/course/pds/

Solutions, quizzes, etc, locally stored.

General course structure

12 Labs

10 Lab homework + 3 homework

3 Quizzes

2 Final Project (1 individual + 1 group work)

No exam

Responses and reflections

Student learning experience

- Many said the course was **useful** for their career
- Almost all students became proficient in R data wrangling / visualization
- Learned basic computational thinking in statistics and machine learning
- Liked the individual projects / did NOT like group project
- Posting all material online made some students skip classes
- Need more time for a more seasoned statistical analysis (shorten other parts/conversion to 4 credit?)
- Will students be better prepared for advanced courses? (No definite answer yet)

From survey results

- This class made a valuable contribution to my professional development (4.32)
- This course helped to develop my ability to solve real problems in this field. (4.26)
- This course helped me learn to apply concepts from this course to new situations. (4.21)

2. Lab activities

Overall, students feel that lab activities were the best part of the course.

- Pre-lab and post-lab difference was not tested.
- Better retention: students better understood materials covered in lab and were able to use those weeks later. [E.g. Quiz 3: Good performance on #2; low performance on #1]
- Synthesis: Labs/projects in the latter half of the semester require synthesis of older material

From survey results

- Lab experiences helped to clarify the lecture material. (4.63)

3. Piazza forum / technology

- Lynda.com videos too long
- Students generally liked Piazza forum
- Many students were at both TA and instructor's office hours (intimidated by R in the beginning) but the conversation moved to online in the middle of semester
- Online forum could be more useful for larger classes

From survey results

- The instructor facilitated a sufficient level of online interaction (4.37)
- The online interactions contributed to my understanding of the course content (3.74)
- Overall, I was satisfied with my online course experience (3.89)

Summary

The good: Useful class. Real-world applications. Lab activities and online material helpful. Right amount of effort.

The bad: Group project. Not enough statistics. Classroom too large.

Thank you!