Teaching Statement

Samuel Thomas

Data Science

Data Science is an emerging field that requires expertise in statistical methodology, machine learning, and technology. As an applied discipline, Data Science also requires a substantial understanding of the relevant domain, whether in business or in academia. Finally, Data Scientists must be able to communicate effectively to audiences of all levels of technical sophistication.

In addition to these core competencies, data scientists must develop an intuition on when to switch between these different roles. For some situations, a data scientist must play the role of applied statistician. For example, A/B testing of a website requires an understanding of study design. Other situations require strong computational skills. Data scientists must be able to fit statistical or machine learning models using software such as **R** or Python. However, they must also be prepared for more mundane data preparation tasks that may require querying large databases using SQL. Finally, data scientists must be able to execute a final product in a way that benefits their organization.

In my experience, data scientists tend to migrate to the particular area of expertise where they feel the greatest degree of comfort and satisfaction. Some data scientists become experts at SQL and data bases, but develop only a cursory understanding of statistics and machine learning. Others enjoy the challenge of building models, but eschew data preparation and dislike communicating to those outside of their field. Still others communicate well with stakeholders, but fail to appreciate the time, skill, and detailed work necessary to implement a data science project. I believe that Data Science education must push students to develop competencies in all areas of the field, particularly those aspects that come less naturally.

Professional Experience

I began my professional work in this field in 2008 as a statistical contractor for a large financial services firm, long before Data Science became a well-known discipline. The first data science model that I developed was an employee attrition model for a large service center based using cox regression. This model used employment tenure as time with a substantial portion of the data that was right censored. Independent variables included the service center site, overtime worked, and the U3 unemployment rate. I developed the model in **R** and communicated the results to management. A significant finding of the analysis was the impact of employee benefit vesting on attrition. Retirement vesting incremented 20% each year after the first year, with full vesting occurring at the sixth year of employment. At each year anniversary, the projected attrition rate spiked, with a local maximum occurring at the sixth anniversary.

After communicating the results of the model, I implemented the model using Excel which, as of 2008, was the only practical option available. This task required manually programming the many formulas required in survival analysis. For this implementation task, I relied on my understanding of statistical theory as well as my computational skill. This model was used by the business for eight years until a fundamental restructuring of the business.

Data Science technology has evolved since 2008, but the essential elements of practical data science remain the same. For the implementation of my first model, I used Excel, the only tool available to me at the time. Data scientists must be prepared to use whatever technology they have available to them as well. Today better tools such as **R** Shiny can be used to directly push models to users via the web. However, data scientists must have the depth of education to manually program a solution from scratch, should the need arise. Even today many businesses have limited budgets to use data science. Data scientists must demonstrate their value before some businesses will make substantial investments.

Teaching Experience

I was employed as an adjunct professor in Mathematics at Ivy Tech University for 8 years. In that time, I primarily taught College Algebra. Algebra was a required course for the vast majority of majors. While a few students enjoyed math, many students viewed Algebra as a "gatekeeper" class to obtain their degree.

The variance in ability of my students was substantial. In each class, a few of my students enjoyed math and felt comfortable with the material. The majority possessed some aptitude in math, but merely tried to get through the class. Still others had little background in math, or had been out of school for many years and forgot much of what they had learned.

I structured my class time to be about 2/3 lecture and 1/3 practice. I spent the first portion of the class lecturing on the material essential to the course. In the second portion of the class, I gave quizzes to students that were open book and open notes. I encouraged the students to work in groups, but they had to turn in their own individual quizzes. As students worked through the quizzes, I walked around the room and assisted students when they had questions. In my experience, this structure worked well. Since quizzes represented a significant portion of the students' grade, students needed to stay through the entire class period before they could take their quiz.

I view the time students spent working math problems as the most productive portion of classtime. Students discovered their weak spots when working through problems, and gained confidence with practice. I would have preferred a 50/50 split of lecture and practical application. However, for this class there was simply too much material to cover in half of the available class time.

Data Science also requires a significant investment of practical work. I anticipate dividing class time for data science between lecture and practice, somewhat similar to my teaching experience at Ivy Tech. Some more theoretical classes may require more lecture time, while others may benefit from having most of the class focused on practical application.

Potential Coursework

I believe the combination of my academic research, teaching, and professional background would be valuable in providing an educational experience for new data scientists. New data scientists are likely to have a variety of interests. Some may want to work directly in industry after obtaining a Bachelor's degree. Others may pursue an advanced degree, but with a desire for industry. Still others may be drawn to data science research in academia. With my background in a variety of areas of data science, I believe that I can help provide an educational experience that will prepare new data scientists for the type of career that they desire.

Examples of coursework that I would feel comfortable developing and teaching include:

- Introduction to Statistics
- Introduction to Probability
- Statistical Inference
- Quality Control Statistics
- Time Series and Forecasting
- Bayesian Statistics
- Statistical Computation using **R** and Python
- Data Science Consulting
- Machine Learning
- Business Communication