W201 Student Portfolio

Nishant Velagapudi

nishray@berkeley.edu

Fall 2017

Contents

Discussion Question Notes
- Week 1: Prospective Biography
- Week 2: Tactical and Strategic Decisions
- Week 4: Bias
- Week 5: Kuhn’s Data Science Revolution?
- Week 6: Research Design
- Week 7: Fallacies
- Week 10: Visualization
- Week 11: Persuasion

Individual Assignments
- Week 3: Impact of Data Science and Policy Implications
- Week 9: Visualization as Story
- Week 13: Featuring… You!

Group Presentation Notes and Statement of Individual Contribution
- Week 8: Case Studies
- Week 12: Ethics

Final Project Notes and Statement of Individual Contribution
- Week 14: Final Project Proposal
- Week 15: Final

Discussion Question Notes

Week 1: Prospective Biography

I'd like to be an entrepreneur in the data science space. Specifically, I'd like to be working in applied machine learning. I think that the amount of data that is currently available is yet to be fully exploited for insights. I earned a degree in Bioengineering - I'd like to get back to my roots in healthcare if possible and I know there are a lot of opportunities to leverage data to reduce costs/increase quality of care.

Week 2: Tactical and Strategic Decisions

I'll use Microsoft: my current employer. I would define strategic decisions as pre-planned, high impact decisions. I would define tactical decisions as scoped down pivots in planned work/ideas that occur in the pursuit of an overarching strategic goal. One strategic decision my organization (Office 365) is constantly faced with is the balance between customer satisfaction and cost savings. Along those lines, a tactical decision we were faced with was whether or not a new feature was "worth it" as it had increased satisfaction and cost. One road block to success is organizational buy in: the team behind the feature itself is much more inclined to slant reporting that satisfaction is highly valuable. Another road block is being able to personalize our data approach on a tenant-by-tenant basis and not performance analyses that miss too much granularity.

Week 4: Bias

Regression to the Mean – a Need to Assign Blame

I experienced an example of the regression to the mean bias in annual review meetings regarding a certain KPI. We observed the KPI had significantly dipped in June, and a full investigation was underway to understand the root cause behind the assumed regression. There were several key stakeholders from different project groups and backgrounds in these meetings: each of course had their own agenda as to what could cause the worsened metrics.

The bias at hand is an immediate assumption of causality of a regression to the mean. What was ultimately missed prior to this meeting was that the extremely high KPI numbers reported in April and May (which suggested a steep trend upwards) were outliers. The discussion immediately went towards identifying the cause behind the regression: calling out differences (and the associated month-to-month statistical significances) in various API time-out data, the click-through rates on various portions of the website, the kinds of questions being asked, and even the demographics of the visiting users. Kahnemann diagnoses the situation perfectly, arguing that "causal explanations will be evoked when regression is detected, but they will be wrong because the truth is that regression to the mean has an explanation but does not have a cause" (Kahnemann, 2011). If I had the knowledge and experience I do now, I would have pushed for a step back in the discussion. Broadening our perspective and looking at historical trends before diving in would have helped us avoid this bias. When reviewing KPI data from the 15 months prior to April and May, it was easy to see that the metric is stable and June dip was just a return to normal from an unusual high. Like Kahnemann's drill instructors, we were so focused on individual datapoints that we missed the pattern obvious from the entire logs (Kahnemann, 2011). In the rush to explain the April/May to June dip, analyses were focused on just those two months of data and missed greater context.

This experience runs both ways – having seen this regression to the mean effect, I am far less effusive about improvements quantified by this KPI. Instead of immediately giving credit to whatever feature is flavor of the month, I am at first skeptical of any causal inference. Morton characterizes regression to the mean in the field of healthcare – pointing out that abnormal singleton test results often lead to unnecessary treatments (Morton, 2003). She suggests using more data or sequential testing (Morton, 2003). Similarly, we can avoid unnecessary and expensive actions by taking a top-down approach to investigating changes in key metrics rather than being biased towards assuming somebody is "at fault."

Bibliography

1. Kahneman, D. (2015). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
2. Morton, V., & Torgerson, D. (2003). Effect of regression to the mean on decision making in health care. *The BMJ, 326*, 1083-1084. doi:10.1136/bmj.326.7398.1083

Week 5: Kuhn's Data Science Revolution?

Paradigm shift - necessity the mother of openminded-ness. Memory grows faultier the longer the time period. Subconscious factors can play a part in stakeholder decisions.

I would consider the medical field's upcoming transformation. While advances are built upon existing literature in the use of data in healthcare, the status-quo of doctor intuition is hard to shift. Eventually, we will likely look back on the amount of pain caused by inefficient systems, but for now, data scientists and computer engineers are not exactly the most welcome people in care centers. "Inquiry" is currently ongoing - folks are finding correlations and historical trends that could have been used in diagnoses/treatment. A pre-paradigmatic movement could be those backing the introduction of AI (IBM's Watson) into the healthcare space. At that point, competing hospitals would take differing approaches and the paradigms would compete.

Week 6: Research Design

Article:
https://fivethirtyeight.com/features/what-a-doctor-calls-a-condition-can-affect-how-we-decide-to-treat-it/

The article addresses a very important issue: each ICD9 diagnosis has an associated emotional response. However, 538 frames this question in a strange way: the hypothesis is directional and implies a degree of quantification that ultimately does not exist in the study. To summarize, their question is: “How does diagnosis affect antibiotic seeking behaviors and parental decision to keep kids at home?” This question could be framed in a slightly more clearcut manner if we acknowledge the qualitative side of this research. I would pose the follow question: “How do the preconceptions about various diagnoses alter patient behavior from what is doctor recommended?” The data presented in the article fit this new question much better, and we are not forcing a qualitative study to pose as a quantitative one. The question is more actionable - we’ve broadened our scope without losing too much focus and the results are more impactful.

I think that the authors made this mistake because they are highly data-driven and qualitative study is likely frowned upon in their space. However, the subject matter here is societal behaviors - which is difficult to quantify.

Week 7: Fallacies

URL: https://www.washingtonpost.com/blogs/post-partisan/wp/2017/08/03/trumps-fall-wont-be-quick-these-new-polls-show-why/?tid=a_inl&utm_term=.d3df5c628bda

The author describes why Trump’s recent poor performance in the polls don’t necessarily indicate that his chances to win the re-election campaign in 2020 are worsening The author has a vested interest in gaining clicks - which entails pushing a narrative that counters the grain. The idea that worsening approval ratings is not indicative of shifting public narrative is a dismissal of sampling statistics. The author even points out that any poll shows the same results - lowering approval due to healthcare and policy failures. His point revolves around the growing political divides and radicalization of beliefs. The stickiness of this fallacy is that it is completely intentional - the author sensationalizes to draw more clicks. However, the point itself is somewhat overstated.

Week 10: Visualization

Awful visualization:
http://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6#i-never-thought-it-was-possible-but-i-actually-understand-soccer-less-after-looking-at-this-chart-3

I wanted to understand how the various team salaries worked and a friend sent me this. Author uses color, area distinction, x/y positioning, and ordinal ranking (?). Four dimensions of visualizations are used to show two dimensions of data - which is salary and team. I have no idea what the visualization is trying to show.

Great visualization:

http://www.mamartino.com/projects/rise_of_partisanship/

This author made a network visualization to show the “rise of partisanship” in congress - nodes represent members of congress, color indicates party affiliation, edges indicate more than “threshold” agreement. The growth of two distinct clusters tells a story.

Week 11: Persuasion

At work, we worked to persuade a team that the seemingly positive results from their feature were not actually representative. The feature inherently gamed one of our KPIs and thus looked very good - but in depth analysis brought in to question how valuable the feature actually was. We had to work to persuade a group that their work was actually not adding the expected value.

Individual Assignments

Week 3: Impact of Data Science and Policy Implications

slides:
Diagnostic_Potential_Watson.pptx

In this presentation, I talked about the potential of Watson, the stakeholders in onboarding data science into the healthcare space, and potential challenges.

Week 9: Visualization as Story

Kodak_VisualStorytelling.pptx

Presentation on the story of Kodak: focus on visual storytelling in the presentation.

Week 13: Featuring… You!

Title: Using Data Science to Improve Safety in Healthcare
Abstract: The cost of medical malpractice in the United States is about $55.6 billion, $45.6 billion of which is spent on practice to avoid lawsuits. 40% of doctors reported that they took on more patients than they felt they could handle. Nurses are often overworked even more than residents. The development of wearable devices allows for constant data collection that can allow for centralized monitoring of employee status in high-stakes situations. Sleep and stress indicators can monitor both nurses and physicians throughout procedures, or even throughout a

day. Avoidable error rates were correlated with wearable device signals. Thresholds were established for “too fatigued” or “too stressed” to continue work. Monitoring employee physiological signals allows for intelligent allocation of break time or shifts to minimize error rates. Expansion of this technology will reduce healthcare error rate, thus bringing down the overall malpractice costs in the United States and ultimately leading to lower healthcare costs for the patients.

Keywords: Malpractice avoidance, fatigue monitoring, wearable devices

About the Author: Nishant Velagapudi received degrees in Informatics and Bioengineering from the University of Washington in 2016: his research included protein dynamics simulations and predicting undergraduate dropout rates. He currently works for Microsoft Office 365 as a Data Scientist and is pursuing a graduate education at the University of California at Berkeley.

Group Presentation Notes and Statement of Individual Contribution

Week 8: Case Studies

Slides: Optimizing_VC_Investiment_Slides.pptx
I worked on the research question, impact, data sources, and “how will this help” slides. I thought of the web crawling idea to add a novel data source to our research. I wrote the body of the report when our research goal was more centered around web crawling as a novel value add to Venture Capitalism. I put together the challenges and obstacles section, though Tiffany tied it back towards the class concepts (fallacies, bias, decision making).

Week 12: Ethics

Slides: Ethics_GeneticsData.pptx

I contributed the introduction, context, and middle ground portions of our presentation. I had some background in biomedical engineering, so I wanted to leverage that to set the stage. My experience in computational biology in specific gave me a few talking points. I proposed the middle ground involving using IRB’s to dictate allowable uses of genetic data as well.

Final Project Notes

Week 15: Final

Slides: ErrorAvoidance_Healthcare.pptx

I owned the data and methods portion of the presentation. I built out the draft data in the domain and took a first stab at the research question: Nick helped quite a bit in refining and ultimately presenting these two sections.