HW 5

Zach Fechko

Q1

Suppose you are given 7 data points as follows

- $A = (1, 1)$
- $B = (1.5, 2)$
- $C = (3, 4)$
- $D = (5, 7)$
- $E = (3.5, 5)$
- $F = (4.5, 5)$
- $G = (3.5, 4.5)$

Manually perform 2 iterations of K-means clusetering on this data. Use Euclidian distance (L2 distance) as the distance/similarity metric. Assume number of clusters $k = 2$ and the initial two cluster centers C_1 and C_2 are B and C respectively

Answer

Iteration 1

Points	Distance to B	Distance to C	Cluster
A	1.118	3.605	B
D	6.103	3.605	C
E	3.605	1.118	C
F	4.243	1.803	C
G	3.202	0.707	C

Iteration 2
$x = (1.25, 1.5)$
$y = (3.9, 4.1)$

Points	Distance to B	Distance to C	Cluster
A	0.559	5.022	x
B	0.559	3.920	x
C	3.0516	1.421	y
D	6.657	2.195	y
E	4.161	0.412	y
F	4.776	0.608	y
G	3.75	0.721	y

Q2

Read the following two papers and write a brief summary of the main points in at most 4 pages

- ten simple rules for responsible big data research
- Proceedings of Machine Learning Research

Ten Simple Rules for Responsible Big Data Research

Big data and research agendas move well beyond those typical of the computational and natural sciences to more directly address sensitive aspects of human behavior, interaction, and health.

The introduction of Machine Learning and big data into the professional world brings situations that will push researchers outside of their expertise and comfort zone.Social scientists now grapple with data structures and cloud computing, while computer scientists must contend with human subject protocols and institutional review boards. This complexity challenges any normative set of rules and makes devising universal guidelines difficult. This paper presents ten simple rules for responsible big data research.

1. Acknowledge that data are people and can do harm
One of the most fundamental rules of responsible big data research is the steadfast recognition that most data represents or impacts people. This logic is readily evident for "risky" datasets, e.g., social media, but even seemingly benign data can contain sensitive and private info. One example of this is Hague et al.'s use or property records and geographic profiling techniques to potentially identify the street artist Banksy.

2. Recognize that privacy is more than a binary value
Privacy is not reducible to a simple public/private binary. Just because something has been shared publicly does not mean subsequent use would be unproblematic. Looking at a single Instagram photo by an individual has different ethical implications than looking at someone's full history of all social media posts. Understanding and contextualize your data to anticipate privacy breaches and minimize harm.

3. Guard against the reidentification of your data
It is problematic to assume that data cannot be reidentified. When datasets are combined with other variables, it may result in unexpected reidentification. Factors discounted today as irrelevant or inherently harmless may prove to be significant vector of personal identification tomorrow. Even aggregate statistics about groups can have serious implications if they reveal that certain communities suffer from stigmatized diseases or social behavior much more than others.

4. Practice ethical data sharing
Asking participants for broad, as opposed to narrowly structured consent for downstream data management makes it easier to share data. Careful research design and guidance from IRBs can help clarify consent processes. Even when broad consent was obtained upfront, researchers should consider the best interests of the human participant, proactively considering the likelihood of privacy breaches and reidentification issues. This is of particular concern for human DNA data, which is uniquely identifiable. In informed consent and right of withdrawal—are increasingly the exception rather than the rule. Informal big data sources are gathered by agents other than the researcher. These data are only accessible to researchers after their creation, making it impossible to gain informed consent a priori. The burden of ethical use

and sharing is placed on the researcher, since the terms of service can often be extremely broad with little protection for privacy.

5. Acknowledge the limitations of your data
In order to do accurate and responsible big data research it is important to ground datasets in their proper context including conflicts of interests. During the step of data acquisition, it is crucial to understand both the source of the data and the rules and regulations with which they were gathered. While it is tempting to interpret findings based on big data as a clear outcome, a key step within scientific research is clearly articulating. The act of interpretation is a human process, and it is essential to clarify both the strengths and shortcomings of the data. Reflecting on the potential multiple meanings of data fosters greater clarity in research hypotheses and makes researchers aware of the other potential uses of their data. Do not overstate clarity; acknowledge messiness and multiple meanings.

6. Debate the tough, ethical questions
The lack of clear-cut solutions and governance protocols should be more appropriately understood as a feature that researchers should embrace within their own work. Discussion and debate of ethical issues is an essential part of professional development and can establish a mature community of responsible practitioners. Bringing these debates into coursework and training can produce peer reviewers who are particularly well placed to raise these ethical questions and spur recognition of the need for such conversations. There are a number of good models for interdisciplinary ethics research, such as the Science and Justice research center at the University of California, Santa Cruz. Some of the better-known "big data" ethical cases—i.e., the Facebook emotional contagion study provide extremely productive venues for cross-disciplinary discussions. In an effort to help these debates along, the Council for Big Data, Ethics, and Society has produced a series of case studies and recommendations.

7. Develop a code of conduct for your organization, research community, or industry
The process of debating tough choices inserts ethics directly into the workflow of research, making "faking ethics" as unacceptable as faking data or results. Internalizing these debates, rather than treating them as an afterthought or a problem to outsource, is key for successful research. Developing codes of conduct can provide guidance in peer review of publications and in funding consideration.

8. Design your data and systems for auditability
The goal of auditability is to document when decisions are made and, if necessary, backtrack to an earlier dataset and address the issue at the root. Designing for auditability also brings direct benefits to researchers by providing a mechanism for double-checking work.

9. Engage with the broader consequences of data and analysis practices
It is important for responsible big data researchers to think beyond the traditional metrics of success in business and the academy. The pursuit of citations, reputation, or money is a key incentive for pushing research forward, but it can also result in unintended and undesirable outcomes. Recognize that doing big data research has societal-wide effects.

Proceedings of Machine Learning Research

Risk assessment tools have been adopted to assist with a number of decision points throughout the criminal justice system. In spite of the tools' growing influence over judicial or administrative decisions, defendants rarely have an opportunity to probe or challenge them. Critics have called for increased transparency surrounding the development and administration of these tools. Some scholars have focused on the more ethical side of the issue. While assessments can be perceived and maybe marketed as a more objective and scientific alternative to human judgment, they are not immune to bias. In fact, the use of risk assessment tools can perpetuate and exacerbate existing biases in the criminal justice system. This paper presents ten simple rules for responsible big data research. These rules are intended to help researchers and practitioners avoid the pitfalls of bias in big data research. The rules are not intended to be exhaustive, but rather to provide a starting point for researchers and practitioners to consider when designing and implementing big data research projects.

Scholars have identified ways where Machine learning can reproduce existing patterns of individual prejudice and institutionalized bias in a variety of ways, such as by identifying the research question and labeling examples within the training set. It can also reproduce double-encoding bias in proxies for protected classes through its selection of feature selection and feature selection.

In the criminal justice system, risk assessments play an integral role in shaping how practitioners understand and intervene in the lives of offenders. Risk assessments have evolved over the last several decades to support a diverse set of institutional goals and processes. We ask whether the statistical methods currently underlying risk assessments are appropriate to use in the service of this.

The vast majority of risk assessment tools do not use new statistical methods frequently associated with "artificial intelligence," such as machine learning. Instead, they are overwhelmingly based on regression models. Regression models aren't well-equipped to answer causal questions about the relationship between a set of predictors and an outcome. They show correlation but they don't prove causality between predictors.

Q3

Go through the talk by Kate Crawford at NIPS 2017 and write a brief summary of the main points in at most four pages. You can find the talk here: The Trouble with Bias

The Trouble with Bias

Machine learning is becoming really mainstream, being put in multiple facets of our everyday lives. Crawford mentions that the conference had grown from just 200 attendants to over 8000 in just a few short years and uses it as a testement to how mainstream machine learning has become. Machine learning is seen as this hot new commodity and she mentions that new hype overshadows the potential for bias and prejudice in machine learning.

Crawford starts off the example of bias in machine learning models with multiple well-known situations like how a natural language model by Google was scoring words like "black" and "gay" negatively, in contrast with it scoring words like "straight" and even "white power" more positively. Another example she shows is that Amazon's same day delivery range tends to omit areas with a high African-American population and comparing it to the descriminative act of redlining in the 1930s. And the most well known issue of COMPAS scores being massively biased against African-Americans.

The issue of bias in machine learning isn't being completely overlooked however, she shows articles by leaders in the Machine Learning industry that acknowledge that bias in machine learning is an issue that must be fixed.

Crawford mentions that the main issue with bias in machine learning comes from the bias in the data that is used to train it. As most data is collected by humans, that data may reflect that person's personal biases, which could then be passed onto the machine learning model.

She then goes on to break her talk into 5 main points:

1. What is bias?
2. Harms of allocation - where we are now
3. Harms of representation - what we're missing
4. Politics of classification - the big picture
5. What can we do?

What is bias?

> Bias means judgement based on preconceived notions or predjudices

She shows a graphic that shows bias through the fitting of a model. An underfit model fails to capture the trend in the data. An overfit model is extremely sensitive to tiny fluctuations in the data. Bias can even happen when a model perfectly fits the data, like if the model perfectly fits a dataset that contains underlying biases, it will still give biased results.

The most common way that a model becomes biased is through the data that is used to train it. The data could potentially have been collected in a way that isn't entirely transparent. An example she uses is the era of "stop and frisk" in New York from 2004 - 2012. She mentions that around 83% of the people who were stopped and frisked were black or Hispanic. Which

could have been interpreted as black and Hispanic people were more likely to commit crimes, but that doesn't prove causation, only the fact that they were stopped more often, which could be attributed to the racial bias in the cops who stopped them.

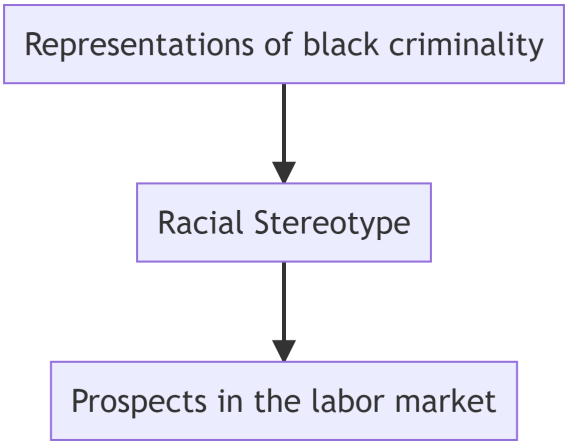
Harms of allocation

> Allocative harm is when a system allocates or withholds certain groups of people from certain resources or opportunities

An example of allocative harm that Crawford uses to contextualize the issue is that if a mortgage support application continually denied certain groups of people from getting a mortgage, that would be allocative harm.

Harms of representation

When certain pieces of information are withheld it can produce to harms of representation. An example she uses is that Google Photos labeled pictures of African Americans as "gorillas". Another being a researcher by the name of Latanya Sweeny discovered a pattern that names that were associated with african americans showed ads for criminal background checks. She shows the following diagram as how racial bias in machine learning represents black people in a negative light which can cause real-world harm.



She then lays out 5 different types of representational harm:

1. Stereotypes
    - Embedded gender stereotypes in words (e.g. "nurse" and "doctor")

2. Recognition
    - Nikon cameras thought that pictures of Asian people were pictures where someone was blinking
    - Image systems not recognizing black faces
3. Denigration
    - Google auto-complete completing queries of "jews should…" with "wiped out"
    - Google Photos labeling pictures of African Americans as "gorillas"
4. Under representation
    - Searching for "CEO" on Google Images brings up pictures of white guys in suits, no women or people of color

Politics of classification

Classification as a practice got its start in the Renaissance with Aristotle grouping things together in order to make imperical conclusions about the world. Classifications of things shine a light onto the social, cultural, and religious belifs of the time (like the renaissance showing a religious bias towards zoological classification).

The faces in the wild dataset features a bunch of white people. Which means that that dataset is better suited for recognizing white people than it is for recognizing people of color. The person who comes up the most is George W. Bush, this is because as a president you are photographed a lot more than the average person.