1)

A = (1,1)

B = (1.5,2.0)

C = (3.0,4.0)

D = (5.0,7.0)

E = (3.5,5.0)

F = (4.5,5.0)

G = (3.5,4.5)

K = 2

C1 = (1.5,2.0)

C2 = (3.0,4.0)

Iteration 1:

A:

A is in C1

B:

B = C1, B is in C1

C:

C = C2, C is in C2

D:

D is in C2

E:

E is in C2

F:

F is in C2

G:

G is in C2

A and B are in C1, C, D, E, F, G are in C2.

C1 = 1/2 \* ((1,1) + (1.5, 2.0)) = 1/2 \* ((1 + 1.5), (1 + 2.0))

= (1.25, 1.5)

C2 = 1/5 \* ((3.0,4.0) + (5.0,7.0) + (3.5,5.0) + (4.5,5.0) + (3.5,4.5)) = 1/5 \* ((3+5+3.5+4.5+3.5), (4+7+5+5+4.5)) = 1/5 \* (19.5,25.5)

= (3.9,5.1)

Iteration 2:

A:

A is in C1 – no change

B:

B is in C1 – no change

C:

C is in C2 - no change

D:

D is in C2 - no change

E:

E is in C2 - no change

F:

F is in C2 - no change

G:

G is in C2 – no change

No point changed cluster membership, no update to centers, clustering is completed.

2)

In “Ten simple rules for responsible big data research”, Matthew Zook et. al. propose a list of ten guidelines for researchers to follow when working with large data sets to use the data responsibly and avoid inadvertently inflicting harm to any individuals or groups or causing public mistrust of the research field as a whole as well as the research organization individually.

Researchers must consider the data in scope and scale and understand that at one level, data may be benign but at larger scale and especially when combined with other also benign data, the new level of complexity can produce damaging results and if such potential outcomes are given consideration at all stages of research then the accidental use of the data to cause harm can be minimized.

The rules they propose are:

1. “Acknowledge that data are people and can do harm”. Researchers must remind themselves that the data they are working with is representative of real people and has real potential to do harm. As data research grows in ability, it’s ability to create harm from datasets that were previously considered harmless also grows. There are already examples of this happening where new uses of data has resulted in influencing criminal justice decisions, credit decisions, limiting access to online retail services, and identifying otherwise anonymous individuals.
2. “Recognize that privacy is more than a binary value”. Designating something as public or private is not straightforward. Data that a user considers public in one context, they would consider private in another. The scope of the use of the data affects this classification as well. For example, viewing a single photo posted to social media by one person is generally considered acceptable but systematically cataloging their entire social media presence, while permissible under current regulations, is likely to be viewed as improper by the user. If researchers maintain awareness of the context of their use of the data, such situations can be avoided.
3. “Guard against the reidentification of your data”. Many data collection systems utilize only the bare minimum of anonymization of data that makes it easily re-identifiable when combined with other data sets. Researchers should acknowledge that no data point is irrelevant simply because it can’t currently be used to identify an anonymized source and that it is likely that at some point in the near future, any data point can be used for reidentification and must employ adequate safeguards against that.
4. “Practice ethical data sharing”. Some of the data researchers work with are gathered from sources that require users to agree to mandatory terms of service that make it impossible for the researchers to obtain informed consent to use the data in their specific study. The researcher must still use the data ethically, according to their principles, regardless of any such terms of service.
5. “Consider the strengths and limitations of your data; big does not automatically mean better”. Researchers must maintain the proper scope of the data that they are using and ensure it stays within the scope of the research being conducted. Proper analysis of the findings is also important and must be clearly specified for anyone that may use the data later in another study.
6. “Debate the tough, ethical choices”. Researchers must comply with review board regulations but that should not be the only set of guidelines they follow. If researchers participate in active debate about the potential uses and misuses of the data at each stage of research, they can be prepared for when ambiguous situations arise. There are existing models for researchers to use to develop their debates to construct practical policies that will protect both the researchers and the research subjects.
7. “Develop a code of conduct for your organization, research community, or industry”. Ethics should be regarded as highly as the integrity of any other part of the study. Researchers should develop “codes of conduct” to avoid any unethical uses of data. This can serve as a preemptive tactic that can avoid any damaging ethical issues that could otherwise arise. There are already policies being considered by governing agencies.
8. “Design your data and systems for auditability”. Developing an internal auditing process will allow researchers to organize their processes to be more methodical and easier to review. Organized processes will also be easier to replicate and allow for easier comparisons of the data between studies.
9. “Engage with the broader consequences of data and analysis practices”. Researchers should consider the consequences of their work beyond the academic. They should consider their potential environmental impact and try to take steps to mitigate it. They must also consider any potential social, economic, and political effects and try to find ways to use their work to improve the world and not to harm it.
10. “Know when to break these rules”. There are some situations that may merit violating any of these rules “for the greater good” such as public health emergencies or disaster relief but, research is still subject to strict regulatory policy so before violating any rule, researchers must carefully consider several important elements and use good judgement to weigh any gains with any losses.

These rules are intended to serve as guidelines to help researchers evaluate any potential harm their work can cause and to provide them with tools to analyze how to minimize any potential damage. The overall goal is to act responsibly and to be ethical by making sure that those thoughts are constantly being considered.

In the article “Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment”, Chelsea Barabas et. al describe the evolution of risk assessment tools employed by the criminal justice system to make recommendations and their benefits and pitfalls.

Although the criminal justice system has employed data-driven, machine learning practices, the underlying principle that the system is built on is biased and focuses on predicting future offenders although it has the potential to be used to identify key areas where the system could intervene to prevent criminals from reoffending.

The system dates back to the 1920’s and began using skilled professionals to make risk assessments based on “semi-structured” clinical evaluations. These were judged as too subjective and were replaced in the 1970’s with regression modeling that focused on unchangeable historical factors that are implicitly biased and unactionable as they are factors that offenders cannot change. This resulted in higher incarceration rates which have been proven ineffective at lowering crime rates.

The next two iterations of the system incorporated dynamic factors but are still employed to predict future criminals rather than to determine possible intervention methods that could prevent future crime thus achieving the overall goal of lowering crime.

There has been much discussion over incorporating machine learning into this system but the underlying system still does not provide a means for identifying important causal links which are considered key to successfully using the system in an unbiased and effective manner.

Without new and extensive research to analyze the data to determine and effectively employ causal relationships, the use of machine learning in this area will continue to make the same mistakes that critics of the previous systems have highlighted and will not offer any advancements that the technology is capable of.

Q3:

In her talk “The Trouble with Bias”, Kate Crawford speaks about the social implications of large scale data, machine learning, and AI. The fields are expanding into every aspect of life and they are bringing humans inherent biases with them. Industry leaders are acknowledging the problem and considering its scale, it has the potential to affect up to two billion people per day.

The problem will not be solved easily. It is more than just a technical issue, a large part of the problem is the inherent bias in the training data. There is a history of structural bias in humans that is reflected in the data.

She discusses five major themes related to this issue. The first theme is, “What is bias”? It is prejudice. Speaking from statistics, it is the “systematic difference between sample and population” and that some data may have a tendency to be selected more frequently than other data. The legal definition is “Judgement based on preconceived notions”. The bias in computer systems comes largely from the data it’s trained on.

The second theme is “Harms of Allocation” where a system allocates or withholds a resource but it also includes representational harm where someone is harmed whether or not a resource is allocated, by problems such as stereotyping or word embedding where gender stereotypical language is used in certain contexts. There is an example from one website that was showing advertisements for criminal background checks more often when African American names appeared. Such ads could have negative impact if the reason that name is being searched is related to employment because it can unfairly bias a potential employer against a candidate. There are also harms from denigration such as culturally disparaging terms that can show up in things like Google autocomplete providing negative or racially biased search suggestions to certain cultural groups. Another problem is underrepresentation and she gives an example of a Google image search for CEO that produces a large number of results of white men.

The question of how to address these issues has some suggestions of breaking the negative associations or “scrubbing” the data to neutral. The question then becomes whose definition of neutral? Who should get to decide what the baseline is for the data? What is a fair way to represent the data?

This is an issue of classification and it is more than a simple technical issue, it is a social issue that begs the question, “What if this is always going to be a problem?”. Kate states that there are two themes of classification, one is that “it is always a product of its time” and two, “We are in the biggest experiment of classification in history”. Essentially, every attempt at classification will be influenced by the issues of our time, social, political, cultural.

The things we need to do as a community to address these issues are one, to fully test systems using multi-disciplinary groups from multiple perspectives and let people raise the flag whenever they find an issue. Two, focus more on the ethics of the problem of classification. Ask ourselves seriously “Are there some things we shouldn’t build?” and act responsibly.

Unfortunately, there are no easy answers to any of these questions. This is a challenge that is facing data scientists now and we need to actively participate in discussions to work to improve the systems as much as we can.

Empirical Analysis:

I received full points for the empirical analysis on Homework 4 so I have not done the empirical analysis for Homework 5.