Homework 5

Q1. Suppose you are given 7 data points as follows:

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)

Manually perform 2 iterations of K-Means clustering algorithm (slide 22 on clustering) on this data. You need to show all the steps. Use Euclidean distance (L2 distance) as the distance/similarity metric. Assume number of clusters k=2 and the initial two cluster centers C1 and C2 are B and C respectively.

Euclidean Distance (k = 2)

Initial Clusters ( = B and = C)

|  |  |  |  |
| --- | --- | --- | --- |
| Data Point | ED from B | ED from 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} |

Table

Description automatically generated

= {A,B}

= {C,D,E,F,G}

Q2. Please read the following two papers and write a brief summary of the main points in at most FOUR pages.

[https://www.microsoft.com/en-us/research/wp-content/uploads/2017/10/journal. pcbi\_.1005399.pdf](https://www.microsoft.com/en-us/research/wp-content/uploads/2017/10/journal.%20pcbi_.1005399.pdf)

Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, Jonathan Zittrain: Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. Proceedings of Machine Learning Research (PMLR), 81:62-76, 2018 <http://proceedings.mlr.press/v81/barabas18a/barabas18a.pdf>

*Ten simple rules for responsible big data research. PLoS Computational Biology*

Being able to learn and apply big data techniques can be a powerful tool when done correctly. Knowing its limits to the outside world and being ethically conscious are just a few ways why data analytics has become very controversial. The amount of ethical questions and responsibilities raised by big data researchers have become very evident in the data science world. From this paper, there are ten simple rules that a data analyst should apply when addressing complex ethical issues.

Acknowledge that data are people and can-do harm. It is important to assume that data are people, and people should always be treated with the respect they deserve. Even data that seemingly have nothing to do with people might impact individuals lives in unexpected ways. The value data can hold can be so significant if treated correctly or very detrimental if not. No one ever gets an automatic pass on ethics so make sure the assumption that data are people is always the first case, until proven otherwise.

Recognize that privacy is more than a binary value. Breaches of privacy are key means by which big data research can do harm, and it is important to recognize that privacy is contextual and situational, not reducible to a simple public/private binary. Using a public data set could run counter to the original privacy intents of those who created it and raise questions about whether it represents responsible big data research. It can also become a major issue for the reputation of the users dealing with the data. In order to minimize harm and anticipate privacy breaches, situate and contextualize your data first.

Guard against the reidentification of your data. You cannot assume that data cannot be reidentified. Unexpected reidentification can happen when datasets are thought to be anonymized are combined with other variables. Work to minimize them in your published results to the greatest extent possible when identifying possible vectors of reidentification in your data.

Always practice ethical data sharing. Careful research design and guidance from IRBs can help clarify consent processes. Be cautious that 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. Researchers must balance the requirements from funding agencies to share data with their responsibilities to the human beings behind the data they acquired. Overall, looking for potential harm from informally collected big data and share data as specified in research protocols.

Consider the strengths and limitations of your data; big does not automatically mean better. Context, acquisition, and clearly articulating what data represents are important stages in the research. Given the almost organic nature of many datasets derived from social actions, it is fundamental that researchers be sensitive to the potential multiple meanings of data. Document the provenance and evolution of your data. Do not overstate clarity; acknowledge messiness and multiple meanings.

Debate the tough, ethical choices. Have those open-ended debates. It is an essential part of professional development as it can establish a mature community of responsible practitioners. Do not be afraid to ask the hard questions. Why might one set of scholars see this as a relatively benign approach while other groups see significant ethical shortcomings? Where do researchers differ in drawing the line between responsible and irresponsible research and why?

Develop a code of conduct for your organization, research community, or industry. Make industry researchers and representatives of affected communities’ active contributors to this process**.** A good approach to this is to develop codes of conduct for use in your organization or research community and for inclusion in formal education and ongoing training.

Design your data and systems for auditability. Responsible internal auditing processes flow easily into audit systems and also keep track of factors that might contribute to problematic outcomes. The goal of this is to know when and what decisions are being made, and if needed, being able to backtrack when mistakes were made. Audits should be part of all systems in big data practices.

Engage with the broader consequences of data and analysis practices. Doing big data research always has societal-wide effects, no matter what. Big data researchers collectively represent an interest group that could rally behind such a call for change. There is no time like right now to be part of the change and revolutionize the system for good.

Lastly, know when to break these rules. A pandemic is a perfect situation when rules may need to be broken. But be aware that an “emergency” is not a justification why these rules need to be broken. Be able to backup your reasoning first. Understand that responsible big data research depends on more than meeting checklists.

*Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*

It is a very common conception that the U.S justice system is radically biased. In many cases, this is true. The use of “big data” or “artificial intelligence,” and their methods and tools in question are often incremental iterations on much older actuarial decision-making practices. The use of regression in risk assessments is one of purpose rather than one of bias or accuracy. The service of predicting future crime can have serious affects and can become difficult to break the cycle once it is started.

The current debate on the fairness of risk assessments is that these tools have been adopted to assist with a number of decision points throughout the criminal justice system, from pretrial release to post-conviction sentencing, probation and parole. Sometimes risk assessment can be perceived as an objective means of overcoming human bias, it is not always the case. Human discretion enters into the formation and interpretation of risk scores. Statistical methods underlying risks assessments like regression and machine learning are appropriate methods to use when dealing with unbiased goals in the justice system.

The question we need to ask ourselves is if risk assessments is predictive or diagnostic tool. There have been generations of risk assessments tool ranging from the 1920s until now. The earliest generation were based largely on semi-structured clinical evaluations that were carried out by skilled professionals as part of an effort to identify rehabilitative treatment options for offenders. The second generation optimized and validated for predictive accuracy using a new statistical method called regression modeling that enabled researchers to identify variables that are predictive of an outcome of interest, without necessarily having to understand why that factor is significant. The third generation of actuarial tools integrated factors that were conceived as “criminogenic needs,” or intervenable factors that are believed to impact one’s risk of recidivating. In today’s world, risk assessments are used for two primary purposes, “prediction-oriented” and “reduction-oriented” approaches.

We need understand regression and machine learning and their purposes in order to spot out the differences between them. Regression analysis is widely used for purposes of forecasting future events. Its purpose is to identify a set of variables that are predictive of a given outcome variable. Machine learning has similar principles to regression for predicting an outcome variable, but it expands the set of covariates for prediction. This is achieved via kernel transformations and parameter sweeping for the purpose of increased predictive accuracy.

Causal Inference is another type of risk assessment that needs to be discussed. Unlike regression and machine learning, causal inference is a framework that is used to establish causal relationships between covariates and the outcome variable of interest. It is achieved through the design of experimental conditions in which covariates are altered systematically to see if the alteration produces effect changes in the outcome variable. It produces new data to be analyzed for the purpose of establishing causality and randomly assignment units to both the treatment and control variables.

Q3. Please go through the excellent talk given by Kate Crawford at NIPS-2017 Conference on the topic of “Bias in Data Analysis” and write a brief summary of the main points in at most FOUR pages.

Kate Crawford: The Trouble with Bias. Invited Talk at the NIPS Conference, 2017. Video: <https://www.youtube.com/watch?v=fMym_BKWQzk>

*The Trouble with Bias. Invited Talk at the NIPS Conference*

Implications of machine learning are rapidly expanding into many areas of everyday life. This is seen in healthcare, education, and the criminal justice system. Vast new techniques of infrastructure are emerging, and we are just on the brink of their full capabilities. With that being said, we have to be extremely cautious about the bias assumptions being made in our predictions.

It is no secret that history always repeats itself in many ways. All around the world, there will always be a history of discrimination and injustice between groups and religions. AI and machine learning can be just as at fault to these bias assumptions. For example, the map of zip codes Amazon was adhering to the same day delivery was almost picture perfect to the FHA map where banks were denying mortgages to black area zip codes.

The term bias according to Kate Crawford can be an overlapping meaning that causes confusion. It is systematic differences that cause errors and estimation of a sample of a population. Bias can come from the data it was trained on and it can be constructed in a nontransparent way. There are harms of allocation that derive from resources and there are harms of representation that derive from identity. Both harms should always be recognized at any level and fixed if needed.

There is also the issue of classification. Social issues for people are being classified and this can lead to uprises like we are seeing today. Machine learning can be a problem that will fundamentally divide the world into parts if it is not worked on and fixed now. It should always be noted that the data sets are just as good as the people who researched them. They reflect the culture and the hierarchy of the world they were made in.

Bias is a core problem to the machine learning field. What we need to do in order to make the machine learning a better, more productive system is to address the important points and ask the right questions. Start working on fairness forensics. Ask how we track the life cycle of a training set. Take in the disciplinary and work across different fields in your organization. Always think about the ethics of classification and who will benefit from the system and who will be harmed.