CptS 315: Introduction to Data Mining Homework 5 (HW5)

Instructions

- Please use a word processing software (e.g. Microsoft word) to write your answers.
- You will need to submit your answers as a pdf file on Blackboard.
- This assignment is due by the date stated on Blackboard.
- All homeworks should be done individually.

Q1. (40 points) 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); and 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 C_1 and C_2 are B and C respectively.

Iteration 1

A = (1, 1); $C1^d$ = 1.118 $C2^d$ = 3.6 C = C1

B = (1.5, 2.0); $C1^d$ = 0 $C2^d$ = 2.5 C = C1

C = (3.0, 4.0); $C1^d$ = 2.5 $C2^d$ = 0 C = C2

D = (5.0, 7.0); $C1^d$ = 6.1 $C2^d$ = 3.6 C = C2

E = (3.5, 5.0); $C1^d$ = 3.6 $C2^d$ = 1.118 C = C2

F = (4.5, 5.0); $C1^d$ = 4.24, $C2^d$ = 1.8 C = C2

G = (3.5, 4.5) $C1^d$ = 3.2, $C2^d$ = 0.707 C = C2

C1 = 1.25, 1.5

C2 = 3.9, 5.1

Iteration 2

A = (1, 1); $C1^d$ = 0.559 $C2^d$ = 5.02 C = C1

B = (1.5, 2.0); $C1^d$ = 0.559 $C2^d$ = 3.9 C = C1

C = (3.0, 4.0); $C1^d$ = 3.05 $C2^d$ = 1.42 C = C2

D = (5.0, 7.0); $C1^d$ = 6.65 $C2^d$ = 2.19 C = C2

E = (3.5, 5.0); $C1^d$ = 4.16 $C2^d$ = 0.41 C = C2

F = (4.5, 5.0); $C1^d$ = 4.77, $C2^d$ = 0.6 C = C2

G = (3.5, 4.5) $C1^d$ =3.75 , $C2^d$ = 0.72 C = C2

C1 = 1.25, 1.5

C2 = 3.9, 5.1

Q2. (30 points) Please read the following two papers and write a brief summary of the main points in at most FOUR pages.

Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Pea Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara Knig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, Frank Pasquale: Ten simple rules for responsible big data research. PLoS Computational Biology 13(3) (2017) https://www.microsoft.com/en-us/research/wp-content/uploads/2017/10/journal. pcbi_.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

Over the last 5 years the use of Big Data has grown tremendously. With the recent growth in the usage of big data the rules and responsibilities of using it are due to be revisited. One of the first things that needs to be acknowledged when using big data is that most data that is used will represent or impact people in some way. This can happen in several different ways such as the use of population wide datasets and non-anonymous data. The next rule that needs to be visited is the usage of privacy. This rule means that privacy is dependent upon context and no single privacy rule can be applied in all situations. The 3rd rule of big data is that data needs to be incapable of being reidentified once it has been anonymized. An example of this can be metadata that is left in datasets or other innocuous data that is forgotten about when users are removing any identifying data. The 4th rule is data needs to be shared ethically. Not all data

should be available publicly and any data that is needs to be properly anonymized before it is made available. The 5th rule of big data is to recognize the strengths and limitations of the data being used. Just because a data set has a large sample size does not mean that I will necessarily be better than a smaller data set that contains more relevant data. The 6th rule is all applications of big data should be open to ethical debate. If one person has concerns about the usage of a certain data set, then there is a responsibility for all people involved to discuss the ethical concerns of using said data. The 7th rule is ensure that there is a code of conduct for the organization that is in control of the data. If there are no rules of conduct for the big data then there is a chance that a team might develop a fake code of ethics that allows them to use the data without actually caring about who it could harm. The 8th rule is designing the data and systems with auditability in mind. If the application is built with the capability to audit then it makes it much easier to double check work and trace the root cause of issues when they occur. The 9th rule is to engage with the broader consequences of data and analysis practices. This rule means that users need to think about the impact of their system or application beyond the traditional metrics that are normally used to measure the impact of the data. This could include things such as energy demands for their applications or other metrics that show the negative impact of big data on the environment. The 10th rule is to know when to break the rules. This rule means that there are times where there is behavior that could be undefined by the rules and independent thought is needed.

Q3. (30 points) 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

Bias is a massive issue in AI that could surpass any other danger that is introduced by this field of study. This danger lies in the bias that comes with data and how technology interprets this bias when it is exposed to a data set. As technology spreads through critical areas of society such as medicine and low and as data is fed into AI systems, the results could be dangerous. If a biased data set is handed to an AI system, then the system will end up biased itself. In many industries there are already warning from experts that warn of algorithmic bias that no one is trying to identify or correct. When training a system, it is especially important to look for hidden biases in the data training it. For example, if a black box system that is used in a critical area such as medicine or law that helps assist with decisions, there could be a bias from the system because it was trained by data that was biased against certain individuals or groups. If the consumer has no way of knowing how the AI works or what trained it then how can any consumer ever trust that the results that it provides are truly unbiased and completely objective. This issue with bias goes deeper than just analysis of training data though. Many of

the techniques and algorithms utilized in the AI field have grown so complex that is exceedingly complex and difficult to examine how bias is impacting the internal workings of the AI. One way that this issue needs to be addressed is by exploring ways that AI systems can give simplified approximations of their internal workings to consumers and end users, so they have a way of easily verifying the validity of the decisions made by the system without having to audit all the intermediate work that the AI performed. Another issue around the bias that needs to be kept in mind is consumer grade AI is only offered by a few large companies. That means that any 1 type of bias that is offered in a single companies AI could have a large impact on society because of the large market share that it covers. With that in mind its easy to see that AI wont be making a large jump in capabilities where it goes from a tool used by humans to a system that tries to get rid of us.