Washington State University
School of Electrical Engineering and Computer Science
CptS 315 – Introduction to Data Mining
Online

Ananth Jillepalli

Homework 5

Name: Nam Jun Lee

Student Number: 11606459

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); 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.

[K-Means clustering]

Input: 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)

Num of clusters k = 2

Initial two cluster centers: C1 = B = (1.5, 2.0)

C2 = C = (3.0, 4.0)

Euclidean Distance = $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

Iteration 1:

Find the Euclidean Distance between each point to C1 and C2.

Euclidean Distance between A and C1: $\sqrt{(1.5-1)^2 + (2.0-1)^2} = 1.118$

Euclidean Distance between A and C2: $\sqrt{(3.0-1)^2 + (4.0-1)^2} = 3.61$

Euclidean Distance between B and C1: $\sqrt{(1.5-1.5)^2 + (2.0-2.0)^2} = 0$

Euclidean Distance between B and C2: $\sqrt{(3.0-1.5)^2 + (4.0-2.0)^2} = 2.5$

Euclidean Distance between C and C1: $\sqrt{(1.5-3.0)^2 + (2.0-4.0)^2} = 2.5$

Euclidean Distance between C and C2: $\sqrt{(3.0-3.0)^2 + (4.0-4.0)^2} = 0$

Euclidean Distance between D and C1: $\sqrt{(1.5-5.0)^2 + (2.0-7.0)^2} = 6.103$

Euclidean Distance between D and C2: $\sqrt{(3.0-5.0)^2 + (4.0-7.0)^2} = 3.61$

Euclidean Distance between E and C1: $\sqrt{(1.5-3.5)^2 + (2.0-5.0)^2} = 3.61$

Euclidean Distance between E and C2: $\sqrt{(3.0-3.5)^2 + (4.0-5.0)^2} = 1.118$

Euclidean Distance between F and C1: $\sqrt{(1.5-4.5)^2 + (2.0-5.0)^2} = 4.243$

Euclidean Distance between F and C2: $\sqrt{(3.0-4.5)^2+(4.0-5.0)^2}$ = 1.803

Euclidean Distance between G and C1: $\sqrt{(1.5-3.5)^2+(2.0-4.5)^2}$ = 3.202

Euclidean Distance between G and C2: $\sqrt{(3.0-3.5)^2+(4.0-4.5)^2}$ = 0.707

Make a table:

Point	Euclidean Distance to B	Euclidean Distance to C	Cluster
A	1.118	3.61	C1
B	0	2.5	C1
C	2.5	0	C2
D	6.103	3.61	C2
E	3.61	1.118	C2
F	4.243	1.803	C2
G	3.202	0.707	C2

Hence,

C1 = A, B = (1,1), (1.5,2)

C2 = C, D, E, F, G = (3.0,4.0), (5.0,7.0), (3.5,5.0), (4.5,5.0), (3.5, 4.5)

Find Centroid:

C1 = (1+1.5) / 2 = 1.25; (1+2) / 2 = 1.5

= (1.25, 1.5)

C2 = (3.0+5.0+3.5+4.5+3.5) / 5 = 3.9; (4.0 + 7.0 + 5.0 + 5.0 + 4.5) /5 = 5.1

= (3.9, 5.1)

Iteration 2:

Using the centroid of C1 and C2 obtained from iteration 1, find the Euclidean distance.

Euclidean Distance between A and C1: $\sqrt{(1.25-1)^2+(1.5-1)^2}$ = 0.559

Euclidean Distance between A and C2: $\sqrt{(3.9-1)^2+(5.1-1)^2}$ = 5.022

Euclidean Distance between B and C1: $\sqrt{(1.25-1.5)^2+(1.5-2)^2}$ = 0.559

Euclidean Distance between B and C2: $\sqrt{(3.9-1.5)^2+(5.1-2)^2}$ = 3.920

Euclidean Distance between C and C1: $\sqrt{(1.25-3.0)^2+(1.5-4.0)^2}$ = 3.052

Euclidean Distance between C and C2: $\sqrt{(3.9-3.0)^2+(5.1-4.0)^2}$ = 1.421

Euclidean Distance between D and C1: $\sqrt{(1.25-5.0)^2+(1.5-7.0)^2}$= 6.657

Euclidean Distance between D and C2: $\sqrt{(3.9-5.0)^2+(5.1-7.0)^2}$ = 2.195

Euclidean Distance between E and C1: $\sqrt{(1.25-3.5)^2+(1.5-5.0)^2}$ = 4.161

Euclidean Distance between E and C2: $\sqrt{(3.9-3.5)^2+(5.1-5.0)^2}$ = 0.412

Euclidean Distance between F and C1: $\sqrt{(1.25-4.5)^2+(1.5-5.0)^2}$ = 4.776

Euclidean Distance between F and C2: $\sqrt{(3.9-4.5)^2+(5.1-5.0)^2}$ = 0.608

Euclidean Distance between G and C1: $\sqrt{(1.25-3.5)^2+(1.5-4.5)^2}$ = 3.75

Euclidean Distance between G and C2: $\sqrt{(3.9-3.5)^2+(5.1-4.5)^2}$ = 0.721

Make a table:

Point	Euclidean Distance to C1	Euclidean Distance to C2	Cluster
A	0.559	5.022	C1
B	0.559	3.920	C1
C	3.052	1.421	C2
D	6.657	2.195	C2
E	4.161	0.412	C2
F	4.776	0.608	C2
G	3.75	0.721	C2

Hence,

C_1 = {A, B}

C_2 = {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.

Article 1: Ten simple rules for responsible big data research

This article explains 10 rules to solve ethical problems in big data research. Currently, as society is digitized, many companies and governments are using big data . As the size and complexity of the dataset increase, this presents many ethical proble ms.

The first rule that can solve this problem is that researchers should be aware that data can harm people. The second rule is that the most common ethical problem in big data research is that individual privacy infringement should be recognized. For example, many people are now releasing their personal profiles on social networks. N o matter how much shared data is used, using such data does not mean that it allow s individuals to use their data. Therefore, it is important to understand the situation o f the data to minimize this damage, as conducting research using such shared data ca n constitute privacy infringement. The third rule is to prevent data re-identification. If unexpected re-identification occurs while combining data, this can be a problem, so it is necessary to minimize it by finding a re-identifiable vector in the data in advance. The fourth rule is that ethical data sharing should be done. These ethical data are m ainly used in disease research, and the burden of ethical use and sharing in generatin g and studying these human-targeted data lies with researchers. Therefore, ethical conc erns arising from the sharing of informally collected human data should be informed t o shared institutions in advance. The fifth rule is that the limitations of the data must be considered. It is very important for people who study through big data to properl y ground their data sets. If research results are derived, including unimportant data, th is can invalidate the potential multiple meanings of the data, so researchers should co llect data considering the organic nature of many datasets. The sixth rule is to discus

s ethical choices. As researchers conduct research using big data, they can often face a crisis that differs from IRB's obligations or faces external situations. Therefore, it is important for researchers to discuss these problems within their peer groups. The seventh rule is to develop a code of conduct for the industry. Since big data research is increasingly likely to develop, developing a code of conduct can bring long-term benefits, researchers should establish appropriate ethical behavior norms within the community by actively developing rules for big data research. The eighth rule is to design a system for auditability. These designs can help researchers make decisions and improve their understanding of the data, making their research a better study. The ninth rule is that it should be involved in the wide-ranging consequences of practice. Since big data research is currently digitized, researchers should focus on the fundamental goal of research as a way to improve the world. The last rule is to recognize when to deviate from these rules appropriately. As mentioned for this purpose, using big data creates many ethical problems, so we must try to conduct ethical research, but sometimes these ethical studies are not the perfect answer. For example, we are currently using disagreeable personal data to find solutions to pandemic diseases such as coronaviruses. This is unethical, but researchers should sometimes know when to deviate from these rules, as such research can rather improve current society.

In conclusion, using big data can cause many ethical problems, so researchers should use the 10 rules mentioned above well in their studies to minimize these ethical problems. It is also important to recognize that researchers may violate individual privacy by conducting research in an unethical manner, but that such unethical behavior has a great impact on the development of our society.

Article 2: Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment

This article describes the competitive concept of fairness and prediction accuracy of actuarial risk assessment tools. Risk assessment tools have been adopted to support decision points throughout the criminal justice system. The criminal justice system says this tool plays an important role in understanding and intervening in the lives of criminals to lower the crime rate in the future. This risk assessment is a prediction technology that has more accurate decision-making power than human decision-making. However, there is an ethical debate that this risk assessment shows low accuracy in a significant part.

The first-generation risk assessment tool appeared in the 1920s and was criticized for being too subjective and low in prediction accuracy. As a result, a second-generation evaluation tool that applied regression modeling, a new statistical method, emerged in the 1970s. This regression modeling tool was suitable for prediction-oriented evaluation, so it began to be utilized a lot for many static historical factors. However, some scholars argued that these prediction-oriented risk assessment tools do not take into account the criminal justice system's own characteristics. Later, in the 1980s, risk assessment tools were reorganized to lower the risk of recidivism and imprisonment rate. The tool included intervention elements that are considered to influence re-invasion risk and dynamic elements that can intervene in the risk. After that, in the 1990s, it reappeared as a fourth-generation tool by adding responsive factors such as intelligence level and psychological disorders to further improve response results.

In conclusion, risk assessment is being used today as a prediction-oriented and reduction-oriented approach. Predictive-oriented approaches are being used to facilitate accurate and efficient predictions of future recidivism, and reduction-oriented tools are being

used to inform treatment and supervision plans. Although many machine learning and artificial intelligence techniques currently allow predictive data analysis, most of these risk assessment tools are based on regression models, the same way as the second-generation assessment tools. Here, the author argues that it is necessary to understand the effects of crime-causing and switch to a diagnostic method that helps evaluate the effects of interventions designed to interfere with the cycle of crime. Also, argue that instead of regression models that simply cling to variables to make the predictive models currently used for covariates, empirical-based tools should be used to help understand and respond to the underlying drivers of crime.

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.

Video: The Trouble with Bias – NIPS 2017 Keynote – Kate Crawford

This video shows sociologist Kate Crawford talking about the problem of bias in machine learning. Currently, the power and scope of machine learning are rapidly expanding in our society. Huge new ecosystems of technology and infrastructure are emerging, and these technologies are creating stereotypes and unfair decisions. Kate Crawford has been working on these problems for seven years, and although bias is important, machine learning systems are affecting millions of people. As such, machine learning is now part of a huge business, so many people are consuming the cost of machine learning. Such machine learning has many problems with structural biases such as fairness and transparency, as we collect data only in a world with a long history of discrimination.

Structural bias is a social and technical issue that will cause difficulties in the field of machine learning. Kate Crawford insists there is still no definite solution to this prejudice. In the sense of machine learning, it is natural that real problems arise in cooperating between disciplines on this topic, as biased systems can produce biased results. For example, if we ask for an interpretation of the results learned through two experts, the main points we receive may be different. To solve these problems, we must develop our ability to break academic boundaries.

She also mentions the issue of classification problems, saying that we need to understand the culture and history of understanding these classification problems. We must continue to try to see the results as fair or to have the distribution we think of. This is a very difficult task, but it actually has a lot of political implications. As such, the technical response

of this classification is very important, but the harm that arises here is that we should consider the bigger problem underlying fairness and prejudice.

There are two things we need to know when conducting classification experiments. The first is to remember that classification is always a product of its time, and the second is to be aware that it is now one of the largest classification experiments in human history. Kate Crawford cited Aristotle as an example. Aristotle was very revolutionary at the time through the study of natural classification, but in this day and age, it has become a basic scientific common sense. As such, it shows that the history of classification will always reflect all attempts socially.

Such machine learning is a job that can develop arbitrarily or culturally specific classifications. Basically, it can be used socially, politically, and legally, which leads to more accurate results by showing neutral determination superior to human judgment. It can earn money for many kinds of events as a means of modern times to predict the future. Therefore, it is time for us to take interdisciplinary convergence seriously.

In conclusion, Kate Crawford noted that the ethics of classification need to be considered more deeply. She says there are many ethical problems with these potential technologies now. Therefore, in this issue of fairness and machine learning, it is important for people in this field to think about who is to blame for the system and who is to be harmed.