Morgan Baccus
CptS 315
Homework 4

Q1.

The equation can be understood as the squared distance between one-dimensional projections of points x_i and z_i is always less than or equal to the distance between the points in a d-dimensional space. For the nearest neighbors calculation in a d-dimensional space, the distance between points in a d-dimensional vectors which are later sorted to find the nearest neighbor. While using this inequality, we can infer that a similar sorting order can also be obtained by calculating mean of feature values to project each point on to a real number line (by a simple addition of d points and then a scaler division) and then sorting those values, which is obviously less complex operation overall.

Q2.

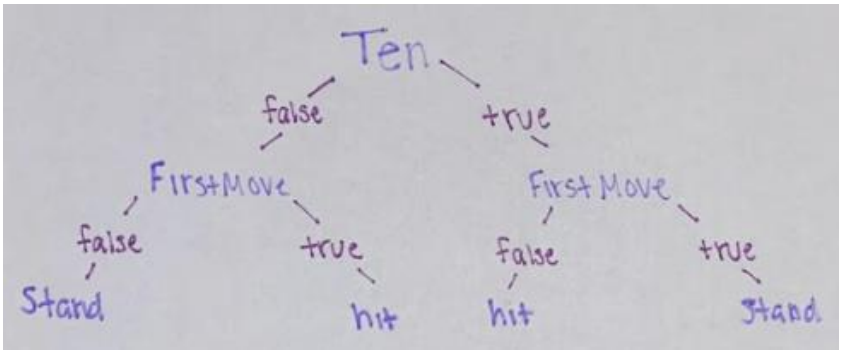
Yes, it is possible to convert the set of rules R = {r_1, r_2, …, r_k} into an equivalent decision tree. For example, suppose we have a decision tree where the class Play is determined based on the values of three input features Ace, Ten, and FirstMove. Each input feature can be true or false and the decision for Play can be either stand or hit. Given the below dataset, we can create if-then rules and a decision tree.

Ace	Ten	FirstMove	Play
F	F	F	Stand
T	F	T	Hit
T	T	F	Hit
T	T	T	Stand

Rules:

1. If Ten = false and FirstMove = true, the Play = Hit.
2. If Ten = true and FirstMove = false, then Play = Hit.
3. If Ten = false and FirstMove = false, the Play = Stand.
4. If Ten = true and FirstMove = true, then Play = Stand.

Decision Tree:



Q3.

Suppose we have the following 7 data points 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) and that the center of cluster C_1 is (1.5, 2.0) and the center of C_2 is (3.0, 4.0).

First Iteration:

Point	**Distance to C_1**	**Distance to C_2**	**Cluster**
A	1.118	3.605	C_1
D	6.103	3.605	C_2
E	3.605	1.118	C_2
F	4.243	1.803	C_2
G	3.202	0.707	C_2

Now we have two clusters: C_1 = (A, B) and C_2 = (C, D, E, F, G).

Second Iteration:

Now the centers of the clusters are X = (1.25, 1.5) and Y = (3.9, 5.1).

Point	**Distance to C_1**	**Distance to C_2**	**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

Q4.

Ten Simple Rules for Responsible Big Data Research

The article “Ten Simple Rules for Responsible Big Data Research” explains the ethical ramifications that big data research can have, as well as different ways that researchers can minimize these negative effects. These ten rules are derived from a two-year National Science Foundation (NSF) funded project that led to the creation of the Council for Big Data, Ethics, and Society. The Council is comprised of 20 scholars from all different disciplines. The main task of The Council is to set the standards for ethical scientific and engineering research by informing the NSF what these standards are and how to encourage their use.

The article continues to list and explain each of the ten rules that they believe are required for ethical big data research. The first rule is to acknowledge that data are people and can do harm. This rule is meant to force researchers to draw the association between the data to the people that the data can effect. For example, researching data that is collected from individuals such as types of social media posts can contain private or sensitive data. It is crucial to acknowledge this and protect the information as best as possible.

The second rule is to recognize that privacy is more than a binary value. This is further explained by saying that privacy cannot always just be public or private. The privacy of data must be decided on the context of the research and the data. It is important to contextualize your data in a way that will prevent privacy breaches and minimize harm to those whose data was collected.

The third rule is to guard against the reidentification of your data. This simply means to anonymize the data as best as possible. More measures must be taken to do this than only removing the obvious such as name. Just because a factor would be impossible to identify someone with now, it could be used in the future to further reduce the possible identities of who supplied the data. It is the researcher's responsibility to identify the possible vectors of reidentification in the data and reduce them as much as possible in the results when published.

The fourth rule, practice ethical data sharing, seems a little bit obvious. This rule is a summarization of a couple different rules listed in this article. To practice ethical data sharing, the researcher needs to consider the wellbeing of the people whose data they are collecting. If sharing data that is collected could result in harm for the participants, then the researchers are responsible for not sharing that data.

The fifth rule is to consider the strengths and limitations of your data. Just because the data is called "big data", does not mean that the data fit every need or show exactly what you want it to. It is important to remember that the data can show many things and sometimes won't be as clear as you want it to be.

The sixth rule is to debate the tough, ethical choices. The purpose of this rule is to require researchers to gather insight from multiple other researchers and colleagues to discuss the choice and come to the most ethical decision. If only one person or a few people make the decision, their own biases could affect the result. This rule ensures that bias will be removed from the decision-making process as best as possible.

The seventh rule is to develop a code of conduct for your entity. This could be your organization, research community, or industry. A code of conduct is essentially a collection of ethical standards that a group all agrees to abide by. By having the group all be aware of and agree to the ethical standards, it creates a system where they all hold each other to those standards. This helps to prevent ethical issues as everyone knows what their standards are.

The eight rule is to design your data and systems for auditability. The key reason for this is so that if a question of ethical research is raised about your data or process, it will be easy for others to go through it all. Designing the research with auditability in mind also holds the researcher to a high degree of responsibility since they know they could be audited and found unethical at any time.

The ninth rule is to engage with the broader consequences of data and analysis practice. This purpose of this rule is to get researchers to understand that the negative effects of their research can extend beyond harming the individuals that the data was taken from. Big data research can affect society both positively and negatively. It is the researcher's responsibility to minimize the negative effects and maximize the positive effects.

The tenth rule, while somewhat counterintuitive makes perfect sense. The rule is to know when to break these rules. Rules are not meant to be broken, but in a society that is constantly changing and where the effect and use of technology is rapidly increasing, these rules might become outdated.

Eventually, these rules might need to change, and some may even need to be broken given the right situation.

This article, while not very specific, clearly lays out what the NSF views as ethical research. Big data can have wide effects in places that are not predicted. It is important to minimize the negative effects while maximizing the positive effects with sound and accurate research.

Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment

The article "Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment" explains how risk assessment tools have evolved to contain biases towards certain groups and how casual inference could be a better method. Risk assessments used in court that are based on older actuarial methods are not as useful as ML or AI tools. The ML and AI tools can be used to intervene in risk rather than only predict them. Additionally, as the actuarial decision-making processes have incrementally been altered to reach a social or institutional agenda. ML and AI, while can be programmed to have bias, begin with a pure interpretation of the data. This kind of algorithm begins with no biases or agendas unless a human alters the program to have one. Casual inference would be better suited for this role than the outdated actuarial methods.

Q5.

The main points of Kate Crawford's talk given at NIPS-2017 are that Big Data and similar research can lead to unknown biases. These biases can cause negative social effects that potentially harm one group while benefitting another. Companies/organizations from Amazon to the local library can have these bias issues in their algorithms. Big data can cause an over generalization by group which can lead to certain algorithms producing stereotypical or negative results.

Q6.

Hidden Technical Debt in Machine Learning Systems

This article discusses the dangers of using the results from ML algorithms without carefully considering the cost. While these algorithms can produce results very quickly while running properly, they produce no results when there is an issue. ML algorithms have extremely high maintenance requirements which must be performed by a person. That person needs to get paid and thus, the required maintenance to keep the algorithms running is very expensive. While all ML algorithms will need to be updated, how the programmers design the system can reduce or increase the amount of maintenance it will need in the future. Factors such as boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system level anti patterns will all play a role in how much maintenance is needed.