Ethics in AI, Machine Learning, and Data Science

December 4, 2020

Preliminaries

This assignments' focus is on ethical and societal aspects of AI, Data Science, and Machine Learning. The aim of this assignment is to gauge your understanding of data science methods, in the context of applications that may involve sensitive private data and with societal implications, and their ethical pitfalls. There will be particular focus on your ability to reflect on the methods inherent limitations and potential misuse.

The assignment (10 pts(=points)) is comprised of three parts:

1. Reading material and multiple choice questions (3.5 pts)
2. Case 1: Wearable health-technologies and private health insurance. (3.5 pts)
3. Case 2: AI-enabled human personality prediction (3.0 pts)

Part 1 is **10** multiple choice questions with **5 choices for each**. Each question has **atleast** one correct choice. Each question can give up to 0.35 pts and down to -0.15 pts, the points for each question is determined by the number of correct and incorrect choices made. Blank questions give -0.1 pts per question – so, all blank is -1pt. Consequently, the highest score is 3.5 pts and the lowest is -1.5 pts. To receive a passing grade with the lowest score in **Part 1**, will require a perfect scores for **Parts 2** and **3**, *and* a timely submission.

Parts 2 and **3** are composed of multiple essay questions where the answers should be made:

1. As succinctly as possible (for example, 1 paragraph, 5-6 sentences)
2. Using clear and logical argumentation, based on technical, statistical, and other scholarly merits. Use these observations and arguments to discuss ethical problems.

Overly long answers with lengthy discussions of irrelevant aspects will be deducted a maximum of -0.2 pts per question.

We are aware that some of the topics covered in this assignment may invoke valid emotional reactions. However, the point of this assignment is to make arguments on a technical and scientific basis, and to illustrate how poorly designed statistical and data science models can serve malicious ends – either intentionally or unintentionally. The best way to combat these ethical pitfalls is by systematic and scholarly argumentation.

When building arguments keep in mind the on the aims outline above and the learning outcomes from the course canvas page.

Part 1: Privacy and history and limitations of AI

The following questions are based upon the reading material and lecture slides available on the Canvas page. The questions are designed to test whether the texts are read and understood, but also to test basic understanding of concepts and conclusions made in the texts. **Grading rules are outlined in the preliminaries above.**

1. Statistical disclosure limitation (SDL) is a collection of methods designed to protect the privacy of individuals whose data has been used in analyses that will be released to the public. Which of the following specifically compromise the usefulness of SDL to protect the privacy of individuals

   ☐ Increasing computational power

   ☐ A critical bug in broadly used SDL softwares

   ☐ Fast internet connectivity to mobile devices including LTE and 5G.

   ☐ Availability of personal data on the internet and other places

   ☐ Increased public concern about privacy following revelations by Edward Snowden

2. Which of the following are ethical arguments *for* protecting subjects' private data in commercial and research applications?

   ☐ If subjects do not want their data to be disclosed publicly they should not share it.

   ☐ Protecting the privacy of subjects is a matter of safeguarding their dignity and welfare.

   ☐ Private data should only be protected insofar that it does not interfere with a company's ability to maximize economic growth, because human well-being is associated with increasing economic propesperity.

   ☐ Protecting private data is only important when it concerns highly private information such as sexual orientation, religion, or political beliefs.

   ☐ A failure to protect the privacy of subjects could render them vulnerable to identity theft and extortion.

3. The following analysis techniques or data acquisition techniques are *not* differentially private

   ☐ In a movie rating data-set, removing the names of subjects and randomizing some of their ratings.

   ☐ In personal health records including hospital admission, removing names, blood type, annual income, information about HIV status, and social security numbers of subjects.

   ☐ In a financial transaction data-set, removing the name and address of subjects excluding postal code, and randomly exchanging labels of credit card transactions within subjects transactions, such that individual payment amounts cannot be tracked to payment of particular services or goods.

   ☐ In an online survey measuring adolescents and young adults recreational use of illicit drugs using a yes or no question and registering their age. The subjects answers to the yes or no question are stored with a probability of 50%. If a subjects answer is not stored it is sampled randomly with a 50-50 chance of yes or no. Similarly, the age is either registered as their correct age with a probability of 50%, else a fake age is sampled from a uniform distribution of integers from 14 to 29 years.

- ☐ A series of street-level cameras in down-town Gothenburg are set up to disincentivize crime and simultaneously to collect data to reduce crowding and improve citizens shopping experiences. The cameras use a facial recognition algorithm to track pedestrians faces and count how often subjects pass by a given camera. No images are stored, each camera generates a coded representation of each face using the same algorithm on each camera. The coded representation of a face, along with the GPS coordinates of the camera, date, and time are stored.

4. Fisher made important contributions to statistical significance testing. He saw it as a way of communicating 'statistical findings that was as unassailable as a logical proof'. Statistical significance testing can *not:*

   - ☐ Tell us whether a sample is likely to have occurred under the null-hypothesis.
   - ☐ Tell us whether what we are measuring is important with regards to the conclusion we wish to draw.
   - ☐ Tell us whether a sample is unlikely, up to some probability threshold, to have occurred under the null-hypothesis.
   - ☐ Infer the underlying cause of the measured data.
   - ☐ Objectively relate humans ethnicity's relationship to their personality traits.

5. What were important early drivers of the statistical research of Galton, Fisher, and Pearson?

   - ☐ The study of coin-flips
   - ☐ Eugenics
   - ☐ Association of genetic variants observed in large human cohorts to disease phenotypes from next-generation sequencing technologies.
   - ☐ Biological evolution.
   - ☐ Fault detection of machines in factories powering the industrial revolution.

6. In a study from 1925 published in the scientific journal which is now known as "Annals of Human Genetics," Pearson claimed to have shown that Jewish children, in particular girls, were less intelligent on average than their non-Jewish counterparts. He further claimed that intelligence was not significantly correlated with any environmental factor that could be improved, such as health, cleanliness, or nutrition. What aspects of the study may influence these conclusions?

   - ☐ Pearson did not use the right statistical significance test.
   - ☐ Pearson did not use a certified intelligence test.
   - ☐ Pearson made use of self-reported data (survey data) of the conditions of students home lives.
   - ☐ Pearson made use of teachers assessment of students intelligence.
   - ☐ Pearson did not account for environmental factors that could not be improved, e.g. the students' refugee status.

7. In the Fragile Families Challenge researchers from 160 teams were tasked to predict a number of social outcomes — child grade point average, child grit, household eviction, household material hardship, primary caregiver layoff, and primary caregiver participation in job training — using extensive longitudinal data. Is this a difficult prediction task? Why? Why not?

- ☐ Yes, because the challenge includes predicting the future ("wave 6"), and only selected data is available for the time-point ("wave 6") to be predicted.
- ☐ Yes, because the challenge includes predicting the future ("wave 6"), and no data is available for the time-point ("wave 6") to be predicted.
- ☐ Yes, because social outcomes depends on many parameters. Many of which may not be captured well by the dataset.
- ☐ No, because the researchers can use any machine learning method and they have access to a big data set.
- ☐ No, because social outcomes are known to be easy to predict even with simple models.

8. The authors of the Fragile Families Challenge study argue that *'predictibility'* is poor measure for understanding the mechanisms of social outcomes. What alternatives do the authors propose?

   - ☐ Descriptions.
   - ☐ Using statistical significance testing.
   - ☐ Using Foucault's analysis of power.
   - ☐ The current understanding is correct but incomplete because it lacks theories that explain why outcomes are difficult to predict even with high-quality data.
   - ☐ Causal inference.

9. What important *ethical* implications may a poor *'predictibility'* of social outcomes have?

   - ☐ It will be more difficult for employers to use AI to screen job applicants, leading to an increased work load on Human Resource employees.
   - ☐ Using predictive modeling in the criminal legal system may lead to wrongful convictions and arrests.
   - ☐ It will put legal strain on the police and military because the most performant models for identifying criminals and terrorists use deep learning that are difficult to interpret.
   - ☐ Using predictive modeling in the child-protective services may lead to wrongful removal of children from their families.
   - ☐ It will require larger machine learning models to improve predictions, which in turn require more compute power and electricity to run, putting strain on the environment.

10. What reasons can lead to unfair and unethical AI and Data Science models and algorithms?

    - ☐ Sampling bias.
    - ☐ Systemic bias in society.
    - ☐ Poorly designed surveys.
    - ☐ Low memory capacity of embedded devices.
    - ☐ Overly rigorous testing and use of the common task method.

Part 2: Wearable health-technologies and private health insurance

A vibrant start-up – Vivaksa – is developing an app for health and stress monitoring. The team is three young caucasian men in their late 20s early 30s, with degrees in Machine Learning, Software Engineering, and Business. The product aims to collect physical activity data and use this to predict health and stress states.

The business model of Vivaksa is to sell the data-analytics software, predicting the health status and stress of costumers in an online manner to private health insurance companies. The company leverages the broad use among the public of smart and fitness watches which include motion and health tracking sensors.

Vivaksa intends that the insights provided by their app will help health insurance providers to fairly tailor insurance premiums to customers dependent on their health statuses. Vivaksa wants to deploy its product ethically and responsibly, so they ask health insurance companies to allow their customers to opt-out of using their data for price adjustments.

1. At Vivaksa, they have collected training data generated by Vivaksa employees using the employees personal Apple Watches over one full month. Their big dataset covers physically active and inactive times of the day, along with daily blood and saliva tests indicating levels of important biological markers for health and stress. The machine learning model uses the sensory data from smart watches and health trackers – any combination of input from photoelectric pulse wave sensors, barometers, accelerometers, gyroscopes, and orientation sensors – to train a machine learning model to predict the biological markers of health and stress.
   (1.5 pts) How can the training pipeline affect the predictions of Vivaksa's health status prediction algorithm when deployed on the general public?
   (1 pt) Can Vivaksa improve their predictions? if so how? and if not, why?

2. A health insurance company has agreed to use Vivaksa app in a pilot program offering customers with good health predictions a lower health insurance premium. To motivate their customers to opt-in on the program, the health insurance company increases the premium for customers who opt-out of the program. A core motivation for the health insurance company to adopt this technology is to increase fairness. Costumers who are healthier and better at maintaining an active and healthy lifestyle have fewer sick-days, fewer hospitalization, and lower costs for medicine and healthcare for the insurance company.
   (1 pt) Can the roll-out of this pilot program undermine the insurance company's own value of fairness? – explain why or why not.

Part 3: AI-enabled human personality prediction

The new start-up – Selfie2Personality – is developing a mobile phone app that let users give active feedback to an AI system that applies visual filters to their smartphones front-facing ("selfie") camera. They allow their users to use the app for free and to share customized 'personality-matched' filtered selfies of themselves on social media.

The founders of Selfie2Personality were inspired by two scientific papers (see reading material) that claim to be able to provide accurate predictions of criminality and trustworthiness based on photo evidence alone. The Selfie2Personality app augments these approaches to allow for the integration of the information about filter settings. The idea of Selfie2Personality is to generate fingerprints of the users' personality traits, including preferences to certain products, political stance, education and income level, trustworthiness, and criminality. Selfie2Personality claim to be able to generate such fingerprints by using photo and filter settings of users combined with their engagement with posts and products on social media (likes or comments) and news articles they share. This information is valuable for several industrial and public partners including mortgage lending companies or law enforcement.

1. (1pt) What are the limitations of Selfie2Personalitys technology - if any?

2. Following protests in a city that caused significant damages to public and private property law-enforcement is looking for suspects. Analysing social media posts prior to the protests, the law-enforcement identify a pattern of certain political leanings, however without being able to pin-point specific suspects as to who caused the damages. To narrow their possibilities, law-enforcement reach out to Selfie2Personality who agree to cooporate, as they see it as their civic duty to help bring justice in society.
   (2pt) Selfie2Personality share the social media accounts and a psychological profile of users from the city metropolitan area who fit the political stance (as predicted from their app) that law-enforcement found to be sympathetic to the protests. Are there any ethical and privacy concerns with this collaboration and how may they manifest?