17-645 Midterm, Summer 2019

Christian Kaestner

Name:

AndrewID:

Instructions:

- Fill in answers in this document or write in a separate document. Ideally, start each question on a new page. Upload the solution as a PDF to Gradescope, mapping questions to pages.
- All questions in this midterm refer to the scenario on the first page. Answers are graded in the context of the scenario; generic answers that do not relate to the scenario will not receive full credit.
- The exam has a maximum score of 56 points. The point value of each problem is indicated. We allocated approximately one point per minute.
- We give an amount of space commensurate with what we expect you to need for each question. Write concise, careful answers; short and specific is much better than long, vague, or rambling.
- The exam is open book. You may use notes, books, and the internet, but do not interact to other humans.

Good luck!

Scenario

During your last year of school at CMU, you have been taking a machine learning class from a professor who also does innovative research on using cheap sensors in smart phones and smart watches for medical diagnoses, such as asthma and heart problems. As the COVID-Pandemic hits the research team consisting of two PhD students has hastily tried to repurpose their research to detect COVID-19 symptoms from audio recordings and gyro sensors: users lie on their back and place the smartphone on their chest and a deep neural network will classify audio and movement as COVID-19 symptoms or not.

Working with UPMC, they managed to get recordings from 108 diagnosed COVID-19 patients and they collected samples from 284 more people that had similar symptoms but did not have COVID-19 and 121 test users (mostly friends and colleagues) without any symptoms – from each of these people, they have about 100 samples for a total of nearly half a million labeled samples. After some innovate modeling and transfer learning from prior research, model accuracy seems fairly promising, with recall around 93% and precision around 65%. They expect that this can be improved with more training data.

Hopeful about their results, they want to release their work. They received a grant from a charitable foundation for extending their research and building a production system that can be used for free. You decide to help with the effort and defer your lucrative start at a big tech company for half a year to help full time with developing and deploying the software as an Android and iOS app. Everybody is excited about the work, but neither the professor nor the two PhD students have any experience with building production systems. Several motivated master and undergraduate students want to help in their spare time.

The project is rewarding but also stressful with problems everywhere. Given the potential, you want to release it soon. The foundation provides sufficient funding to cover infrastructure cost and will deal with all bureaucracy around approval of medical devices (i.e., legal and certification aspects are not in the scope of this midterm), but funding is not enough to hire more full-time developers. Training the model takes several days and the resulting model is over 300mb. Evaluating the model on a smartphone hardware to make predictions is consuming significant energy (several phones get warm) and takes several seconds. The microphones and gyro sensors of different smartphones don’t always have the same quality. You hope that the project will get successful with millions of downloads and users after some press coverage.

Question 1: Goals [8p]

For the project of the scenario, identify a goal of the system and a corresponding measure (that could be realistically assessed in the context of the scenario) at each of the following 4 levels:

(1) **Organizational objective:**

Measure:

(b) **Leading indicator:**

Measure:

(c) **User outcome:**

Measure:

(d) **Model property:**

Measure:

Question 2: Model Quality Offline [8p]

(a) [4p] The precision and recall numbers seem promising at a first glance but you want to get a better sense of how good this really is. Within the realism of the scenario, suggest two baseline heuristics that you could compare against as part of the offline evaluation in a continuous integration system.

*

*

(b) [4p] One of the students helping out in their spare time ask why the dataset is split three ways into training, validation, and test set, rather than just into training and validation. Explain in the context of the scenario why this three-way split may be needed.

Question 3: Model Quality in Production [16p]

(a) [10p] Once deployed as a smartphone app to thousands or even millions of users, you want to see how well you are doing in production. Describe what measure(s) you would collect that relates to *precision* of the predictions (not recall) and how it relates (directly or indirectly) to precision. Then describe what telemetry data you would collect, and how you operationalize the measure with the telemetry data.

Measure:

How the measure relates to precision of the model's predictions:

Telemetry collected:

Operationalization (how the measure is computed from telemetry data):

(c) [6p] There are many different ways to collect telemetry, some more accurate, some more annoying to users, some more privacy invading than others. Briefly argue why your metric design is suitable in the context of the scenario, if necessary by contrasting it to other designs. Your argument should cover at least (1) how much it relies on human feedback, (2) accuracy and precision of your measure, and (3) amount of data collected.

Question 4: Deployment Architecture [16p]

The research team asked you initially just to “put the model into an app and ship it,” but you are wondering whether you should not rather deploy the model as a microservice in some cloud infrastructure.

(a) [4p] What are advantages of deploying the model as part of the app? Explicitly refer to relevant qualities and how they are better achieved in this deployment strategy.

(b) [4p] What are advantages of deploying the model in the cloud? Explicitly refer to relevant qualities and how they are better achieved in this deployment strategy.

(c) [4p] Which deployment architecture would you adopt? Explain your judgement call and the involved tradeoffs (if any) by explicitly referring to the relative importance of the qualities. If you are missing important information for that decision make assumptions (informed guesses) and state them.

(d) [4p] Is there a scenario where you would adopt a different deployment architecture if the research team could come up with a different model (e.g., smaller, interpretable, faster)? Explain what model qualities the researchers could prioritize to change your deployment decision or explain why you always would make the same decision independent of model properties.

Question 5: Risk [8p]

Since you are dealing with medical information, you want to be careful and anticipate risks of the work upfront. Below are two of several risks you identify. For each suggest a mitigation strategy and briefly explain how it reduces the risk.

Risk 1: Users may receive a diagnosis of COVID-19 and may panic taking risky actions, such as trying to get access to unproven medication with severe side effects.

Mitigation and explanation:

Risk 2: Predictions may have high rates of false negatives on cheaper smart phone models, leading to poor health outcomes and increased community spread among people with low income.

Mitigation and explanation:
