25/09/2020

Team management

Alg-e: platform + first models (not multiclass), supervised (lamaciarno)

multiclass learning by another project (antoine, ba student)

Alg-e map: temporally clusters, unsupervised (Zeineb now quentin)

Alg-e impute: which features are most relevant to a task (jonathan)

Ebola models baseline (Reda)

Cumulator carbon footprint (Tristan)

Overall (lia)

Unsupervised maps (kuan = quentin? lol)

Build aims and objectives, follow objectives and throw away those that “suck”

Subject thesis after 1st week “Interpretability and validation” (something something)

Interpretability:

* Allowing people to understand what you are doing (e.g. adding question marks to explain anything, plus visualizing what the algorithm is doing, what space is it in)
* What the approach is about
* Communicating with people outside the field (1st layer)
* 2nd layer: probe the system (confidence intervals on certain deep learning tasks)
  + Explore a way to express confidence intervals
  + Descriptive statistics
  + Confidence intervals (look into)
  + Make the models more interpretable
  + Improve visualization
  + Having statistical ways of breaking open black boxes
    - Get insides on uncertainty
* Both shallow and deep

Platform management:

* Communicate, speaking with IT team
  + What do you want when you have a model, etc…
  + How to communicate gradients
  + Advise

Validation:

* Taking Ebola dataset and applying to Alg-e
* Ebola decentralized learning

Master thesis:

* Piece of scientific writing
* Systematic approach
* Not being biased
* Balance on each side
* Review of existing tools and approaches
* Background: argument why you need to exist (what has been done before and why it sucks)

Aim:

* “I want to save the world” using the key terms in the thesis
* 1 sentence

Objectives:

* Specific and detailed
* SMART
  + Specific
  + Measurable (tangible results… figures, etc)
  + A
  + Real
  + Time-based
* 5 or 6 objectives
* Objective 1: Review literature in interpretable methods in ML in medicine, and available tools. Then give critical reviews why they suck!
  + From where, which key terms, number of results, time range, etc…
* Objective 2: Manage alg-e platform and ensure integration in the alg-e team for the development. Have a flow diagram of how the information and the communication is organized, as a project manager (flow of information)
* Objective 3: improve interpretability of alg-e models through inter-disciplinary visualization. Allowing people to explore probabilistic space. Intelligent descriptive statistics (percentages, or what clusters are in your unsupervised learning model, how does your model appear in the probabilistic space, how are groups separated if they are and what does that mean)
  + Explore interpretability through visualization: people should understand the data they’ve uploaded, understand the model they are using and understand the results (confidence interval)
  + Take a look at lamarciano’s paper
* Objective 4: validation.
  + Validate using the ebola dataset, what works and don’t/ Make models and report what you found on alg-e using algorithms. Insights from my model on important questions.
  + Validate again using someone’s results (Reda) that has been doing that specifically.
    - Validate platform with the dataset, validate the results using Reda’s results
    - Eventually, if data is ready then validate with the real (and real time) dataset
    - Validate it by giving it to clinicians and asking them for criticism

Interpretability part:

* Confidence intervals

Calendar: make timeline and moving goals

02/10/2020

* Be more precise in how you find paper
* General movement of the approach of integrating ML into clinical field
* Thesis: 1 week clean up 1 week finishing methods + generate plots / else + 1 week write results and discussion (intro and methods need to be done prior to that)
* New data incoming: images from lungs and sounds from lungs
  + Would need new models that are very specific to the type of data coming
  + NLP (sentence, tokenize words, feed into classification model / deep learning model – agnostic to length of sentence) on images of different sites on the anatomic body.
    - Why NLP? Because you don’t need “same length” inputs! Makes it more robust
    - Performing very well
    - No way of people uploading images
    - Chest exam – specific model
  + Images
  + Sounds
* How to include this into Alg-E?
* Don’t make too many apps! They are polluting the healthcare system (too numerous and heterogeneous)! Make it as generic as possible
  + Alg-E should be as generic as possible! Do not fall into this trap. It’s a model TRAINING platform rather than just a generator of results.
  + **Is Alg-E the right place to add these models + new type of data?**
    - Who are the users?
    - Rethink what Alg-E is about and allow it to mutate into something else?
    - Originally; key features; allows clinicians to make their own machine learning models, understanding data / model / results and making it friendly (create trust and promote use), gives control over the black box type model = communication tool. Give responsibility to the person USING the model not the person creating the model. Then people upload their datasets. We want to learn from dataset without sharing them (collaborative models) 🡪 model training exchange (in the future, we’re not here yet) sort of like DISTRIBUTED LEARNING.
    - Need same nomenclature across data for now
    - Alg-E = a place where you can train models. Should it be a model-exchange place? E.g. we offer a model trained on Covid cases in Geneva, would you like to use it and train it to recognize pneumonia, or use it for Covid in Tanzania?
    - Is it worth it to have a platform that is ONLY for chest images?
    - If we decide this is a great idea to merge everything into a “big monster’: e-POCT: point of care test, the point of them is to be positive/negative or numerical🡪 tabular. Only two POCT require deeper understanding: lung exam (ultrasound – plugged into phone and oscultation – also plug into phone)
      * POCUS – point of care ultrasound: can be used ANYWHERE on the body –> ultrasound is a worthwhile point of care tool!
        + **Alg-E is a platform to train tools form data that is coming from point of care?**
    - “Butterfly” tool for ultrasound
    - **Alg-E: making representative models for your region for your data!**
* Also existing as of now: Xray model: but not really POCT

For the thesis:

* Steps:
  + How to include images in Alg-E? How to upload? Check butterfly API.
  + How to include new models, in a tab?
* Reviewed thesis objectives + aims
* Create slack channel for team member communication

09/10/2020

Alone:

* Slack team check out what I had in mind
* Done with current lit rev, only need to get some stuff written now I think
  + Any additional papers you have (not that you need to find – anything you have right now)
  + Should probably read some more stuff about state of the art so I can write the background
* Alge platform:
  + should be able to choose different metrics at once.
  + Maybe have a entire left tab for basic explanation? Plus little questions marks that have pop ups, but pop ups take too much space 🡪 redirect to that tab.
  + Should have a “fill this first” before you can fill next fields explanation
  + Possible new interpretation results
    - Shap: computing takes too much time? LIME, partial dependence plots? Remains the best insight I know.
    - Model internals, hyperparameter explanation and limitations
    - CI
    - Result visualization: grouping of results?
* Including other stuff on alge:
  + Would need more specifications about what type of data is acceptable for images and sounds…
  + Should probably go for it
* Hand in thesis on : 22.01.2021: all good on calendar

With Annie:

* Antoine ba multi class model :
  + Contact, he ll get to know the code on a precise level so I can ask him so questions about what is doing what
* Jonathan donz on Mondays (impute)
  + Contact: he knows the code so I can ask
* Another person :Satellite data + infection disease connections. But maybe he should transfer to reinforcement learning and improving the maps we already have.
  + Unsupervised learning cluster detection: these people have the same stuff
  + Go test them
  + Supervised learning: cluster = labeled feature; clinicians look at what is inside the cluster and label it as interesting or not.
  + Epidemiology: person – time- place: if you have identified clusters in space and time, you can say if it is interesting or not and then boom
  + Surveillance team / site: don’t want to test randomly or arbitrarily: want to represent interest not population (bias it on location)
* Ebola dataset: “switch” account and check out databases on alg-e to get started
* FOR sound and images: specific data, specific pipeline, specific use cases: your data needs to be super specific and what you would want to do is super specific (although models are extremely performant they are specific and only for e.g. asthma, covid, pneumonia)
* Having a guide of key features of the model + how to use it
  + **Make that fit in the thesis: what was missing, what I have done to improve that**
  + **add basic information about Ml models, which models are good or not**

**where is the platform located? Amazon cloud, google cloud?**

* **About optimization**: just do it and we’ll see later if it can be optimized, or if it’s worth it.

So far:

* **WRITE background**
* **Get SOME MORE insight on the code take a look at all of it.**
  + INTEGER ROUNDING ISSUE: somewhere in the code you’re forcing int values 0 to 1 so obviously that can be an issue if you have values in between
* **Have a date :** where / when we can talk to clinicians so we can show them what’s up and what needs to be done.

Objective:

* I reviewed the platform: this is good, this needs to be added
* Solicited advice from clinicians and saw if they agreed, which issues they brought up
  + Possible survey of what they think
  + They probably want to be able to say: how many people do I have to test to get certain results (limited resources…)
    - We have to give them the cutoff… we adjust the sensitivity of how many patients we are capturing. Then yield accuracy or specificity.
    - Needed feature: those most at risk (score): take the 10 most at risk / highest score, and only those would be tested (resource allocation)
    - Displaying risks of individuals: predicting ebola severity: a clinical prioritization score for ebola virus disease figure with score + two curves
      * Include image here
* After every step: address what can be addressed
* **Priorities about the project**

**New**

Comm / slack stuff

* Anything I didn’t address?
* People missing who should be in there?

Alg-E Project:

EPFL MLO Alg-E lab team:

·       <https://github.com/epfl-iglobalhealth>

·        YOU:<https://github.com/epfl-iglobalhealth/Alg-E-sandbox>

·        Antoine (multiclass):<https://github.com/epfl-iglobalhealth/AlgE-multiclass>

·        Lia (overall):<https://github.com/epfl-iglobalhealth/Alg-E>

·        Kuan (unsupervised maps): coming...

·        Zeineb (unsupervised maps):<https://github.com/epfl-iglobalhealth/AlgE-Map>

·        Tristan (cumulator carbon footprint):<https://github.com/epfl-iglobalhealth/cumulator>

Goals of the slack channel + collaboration:

-        Be aware of the project as a whole

o   Where the project is going:

§  Data collection as of now

§  New inputs: should we integrate… images and ultrasounds (lung)?

§  Different platforms for different POCTs? Or 1 giant platform?

o   What other team members are currently doing: see above

o   Feedback / ideas are greatly appreciated: where do you think the project should go? What definitely needs to be done?

-        Be aware of the ongoing development, major discoveries and possible hick-ups: basically anything you think people should know

o   Any major news to the project →  Annie / I  will share

o   Any major news to your project / change your subject →  share

o   Something that can help the team / gives insight on the project as a whole or on someone else’s project / anything you think people should know →  share

-        Need advice / any resources concerning someone else’s subject →  ask

-        This is not a substitute for the numerous slack channels already there

o   Try to keep stuff related to this project only

What I’m asking you to do:

-        A two line description of what your project is / your thesis title

-        Mine: Interpretability of Machine Learning algorithms using e-POCT data for clinicians and validation of the Alg-E platform models

o   1 part interpretation of ML models

o   1 part validation of models on the platform

OTHER stuff

-        Slack team check out what I had in mind

-     num num data: access to dataset (need to sign CDA probably) so I can get working on the code

-        Done with current lit rev, only need to get some stuff written now I think

o   Any additional papers you have (not that you need to find – anything you have right now)

o   Should probably read some more stuff about state of the art so I can write the background

-        Alge platform:

o   should be able to choose different metrics at once.

o   Maybe have a entire left tab for basic explanation? Plus little questions marks that have pop ups, but pop ups take too much space à redirect to that tab.

o   Should have a “fill this first” before you can fill next fields explanation

o   Possible new interpretation results

§  Shap: computing takes too much time? LIME, partial dependence plots? Remains the best insight I know.

§  Model internals, hyperparameter explanation and limitations

§  CI

§  Result visualization: grouping of results?

§ integer rounding issue?

-        Including other stuff on alge:

o   Would need more specifications about what type of data is acceptable for images and sounds…

o   Should probably go for it

-        Hand in thesis on : 22.01.2021: all good on calendar

Todo:

* Write background
* Continue reading stuff
* Start coding
* Look at the alg-e platform code some more
  + Also take notes on model hyperparameters / etc for comparison