NIPS 2015: Challenges in Machine Learning (CiML) Workshop

Saturday, December 12, 2015 6:07 AM

Welcome & Introduction: Evelyne Viegas

Challenges in Medical Image Analysis: Bram van Ginneken

Michele: One way to exploit the final results of a challenge, would be to have something like ML as a Service. Would you see this a s a next step for challenge organizers? For example, once the challenge was completed, someone could upload the data and get the results - in some sense you could see all the algos as resources, and then data could be uploaded to the site. Perhaps the organizers c ould offer the participants an opportunity to share in a monetary reward via implementing an ensemble approach.

Bram: Folks are working on this idea - there is one company who claims this as there business model.

Yi: Is there value in specifying specific engineering constraints on solutions in order to force people to design things that c an be deployed. Perhaps you can set multiple time horizons via the use of intermediate goals?

Audience: Will the slides be available?

Bram: Yes!

Ben: Launching a new high-profile image recognition challenge.

Michele: For a challenge, you claim the need to set up things in a very neat way. Is there a way to go with low quality values and s till implement a successful challenge?

Bram: The fact that we are dealing with digital data makes this much easier. You could organize screenings in very innovative way s, but the medical industry is often too conservative for this.

Comment: Perhaps John Arnold Foundation (Center for Open Science) may engage the FDA in order to overcome their resistance to innova tion.

Techniques & Technologies for Benchmarks: Martha Larson

Challenges should help connect algorithm development directly to satisfying users. (Challenges should be realistic.)

Plista: Recommender systems; Evento 360: Search Flickr for events. (Example of an intelligent system)

Example MediaEval Tasks: Violent scenes; Camomile collaborative annotation platform; CLEF NewsREEL (news recommendation algo)

Balasz: Howmuch success / engagement do you have in the scientific community?

Martha: Our chief metric is # of publications - almost 600. Another one that we track is our cross-pollination into other communities.

Yi: On the CLEF newsreel - what are the requirements for the targets there?

Martha: Competitors are responsible for maintaining the uptime of the system and then integrating w/ the API of Plista.

Open Innovation: Balasz Kegl & Ben Hamner

Yi: Can you elaborate on the Data Leakage problem?

Ben: We are able to filter out about 98% of that, but occasionally it creeps in. We are very proactive about trying to avoid or ban bad behavior given our large scale.

Isabelle: A good practice is to have prior winners test the challenge before opening it up - this reveals bugs early thus resulting in a higher-quality challenge. Not everyone is trying to cheat - there are some good actors as well.

Yi: If you stipulate "winning requirements", people aren't accepting of that?

Ben: The difficulty there is that it's sometimes very hard to figure out who is cheating and how.

John: You need very simple rules for success - if you have a panel who has to adjudicate, it's very dangerous, because it's easy to introduce bias into the process.

Isabelle: How can we reward people for pointing out data leakage problems? For organizers this is a benefit and a safety measure.

Balasz: This happened recently in a Physics Challenge we were running!

Jaffray Woodriff: How does Kaggle handle interaction w/ external data?

Ben: We're generally agnostic about that and we are driven by what the goals of the competition are.

Michele: How do we assign credit to participants? When participants do well, they know what kind of approaches are meaningful - why not have the participants vote? The rules could be "both"

Gael: as a lead contributor to SciKitLearn, it's always a "soft" decision to decide to whether or not to put something into the p ackage. It's more acceptable to have a Bot evaluate whether or not something makes the cut, than a human review panel - contributors like the Bot better b/c it is percevied to be impartial.

Chris: In the VOC we have annual challenges, so the best evaluation criteria often comes from tweaking last year's results. The c lassic evaluation of AI systems is "failure cases". We tend to compress this into a small set of scores, but we need to recognize that there's more to it than that.

Ben: Small trusted communities function fundamentally different than large communities at scale

Isabelle: We used "number of times your code was downloaded" as an affirming measure that helped contributors to see

Gael: Trust doesn't scale. When you're with 10 people it's OK, but with higher numbers it doesn't work.

Balasz: But shame scales!

Ben: No, there are better places for people to shame one another than on our site!

Martha: I think trust does scale if you have trust brokers, such as in a crowdsourcing platforms like… No one believes it scales, because it's intuitive.

Gael: I'll re-phrase - it's about "Making Trust Scale" which is very hard.

Sebastian: One way to disincentivize cheating is to offer different opportunities for people to contribute and get recognition for doi ng something good, like how many downloads, or how different your code is from a given median, etc. If people can excel in other criteria, that might help.

Ben: We have a good asset to help make that happen if anyone wants to take it on.

Chris: From the VOC experience, looking at the patterns of the results can address Sebastian's point in a measurable way. One way to see this is if you build a method learner on top, you can see these types of things.

Jaffray: About Kaggle's team / capacity - are you overworked? Are you automating? How is that going?

Ben: Any small start-up is defined more by what it chooses not to do, rather than what it does. As far as capacity goes, the thing I need more th an anything is systems engineers - if you know a talented web developer, have them speak to me.

AutoML Challenge: Isabelle Guyon

Who won? Everybody!

AutoML3: Damir Jajetic (Croatia) - won with Naieve Bayes.

Previous Rounds: Frank Hutter: James Loyd; etc.

Successes & Challenges: Frank Hutter

Bayesian Optimization; AutoWEKA; Auto-sklearn

Isabelle: Are there tricks of the trade?

Frank: Yes, there are lots, esp. around hyper parameter tuning, I'll discuss in a sec…

Audience: Other than discontinuities are there other pathological problems?

Frank: If you have non-gaussian noise, that's one issue. There's another meta-problem here, which is how to address the AutoML imperative - you not only need to improve your models, but you need to bring the right tools to the problem.

SMAC: Sequential Mode-Based Algorithm Configuration

Automl.org/hpolib

Gael: Would it be possible to get a distribution over useful parameters - guidelines for default parameters for this work?

Frank: I agree - I think that's a paper that wants to be written. Matthias may tackle it.

Chris: Can you say a bit more about what pre-processing stages you actually handled?

Frank: There was only ever 1 type of pre-processor, like random embedding. We pretty much used everything in SciKitLearn. For that, you need a lot of pre and post co nditions, and that's not easy.

Gael: You want caching

Freeze-thaw Bayesian Optimization (FTBO): James R. Lloyd

Audience: Do you model correlation b/w the different models?

James: Yes, but I felt that failing fast was more critical than managing correlations

Jaffray: Are there 5 phases in general?

James: I'm glad you asked - coming in a few slides. I need to make this modular.

Balasz: Frank mentioned you could do sub-sampling - is that a similar process?

James: Yes.

Bram: You seem to use classifiers that are relatively quick to train?

James: A combination of quick and slow classifiers.

Matthias: Given there were several different metrics, did you stack according to some primary metric?

Isabelle: in the AutoML challenge, we had a primary metric

Scalable Ensemble Learning w/ Stochastic Feature Boosting: Eugene Tuv

Slides: (Example data set from Intel)

1. Modern Data Domain Challenges
2. Flexible & Informative base learner is needed
3. Base Learner: Time series data type handled directly or through a supervised codebook
4. Base Learner Choice: Surrogates are calculated for every primary split
5. Parallel & Sequential Tree Ensembles
6. Variable Importance
7. Variable Masking
8. Challenge preprocessing
9. ACE: Relevant Feature Selection
10. Redundancy Elimination
11. Challenge Learning Engine: Stochastic Gradient Trees & Feature Boosting (SGTFB)

12. Computational Complexity & Implementation
13. IDEAL Demo

I didn't do any tweaking of the data - all my runs were default runs. I don't have time to do tweaking. :)

Michele: If there is a bit of noise in your features, after some steps, the residual will be noise, yes?

Eugene: I never saw this happen - we don't get more features by overfitting garbage. Residuals is like an error - in boosting you are trying to eliminate error.

Pascal VOC Challenges: Chris Williams

Evelyne: Question re. pairing people - is this done automatically, and if not, are there ways to do that?

Chris: We could automate that retrospectively given a baseline data set.

Bram: You mention ImageNet's breakthrough in 2012, was this caused by training on your data set?

Chris: For a year or two, we mentored ImageNet with FeiFei. There wasn't enough data in the Pascal VOC to train AlexNet. Also, w e were quite tired of doing this after 7 years, so we were happy that the ImageNet team came in and picked up where we left off.

Bram: This is a great learning for me, because it suggests that new methods may emerge from multi-generational teams.

Michele: Was it possible to inspect the inside of the neural network to help with segmentation. For example, if an algo is predicti ng for a cat, and it sees a sheep, you could invert the pixels and see which ones are sheep pixels.

Chris: There are a variety of ways to approach such a problem - I'm probably not the best person to answer it.

Audience: For ImageNet, they choose the top 5 predictions, but when you see the error going down it makes sense to change the evaluat ion metric. Do you know why they don't go down to just the top 1 metric instead of the Top 5?

Chris: One issue is how these images are labelled. If they only have 1 label, it becomes problematic. This brings us back to the question of annotation.

Audience: What you're saying is that the evaluation procedures should rely upon the evaluation metric?

Chris: There are practical issues: the way the data is collected.

Balasz: It's been our mission to broaden participation in the AutoML challenge, but it has been tough because the bar is relatively high for researchers to approach this problem.

Chris: Full automation is a very ambitious goal, but simply creating tools to help with this process is, in itself, an interesting question.

Isabelle: I must elaborate - I have to plead guilty that I was only interested in the data mining part. I thought that segmentation was a lower priority t ask, but what I've learned is that there is a lot of good science that can be done by solving the whole chain of the knowledge discovery process.

Evelyne: James Lloyd said "it's not a pipeline, it's a search space", but here I see you're still displaying this as a pipeline.

Chris: Point taken.

Audience: What if we add "Wisdom" after "Knowledge" in the pipeline :)

Isabelle: that's for next year's CiML Workshop!

Evelyne & Isabelle: Coopetitions

Isabelle: Raise your hand if you participated in a competition - why did you join? How did you benefit?

Joseph Paul: Meeting a random assortment of people from varied backgrounds was extremely beneficial and helped me learn more about the problem space.

Jakub: 3.5 years ago, I joined one of Isabelle's competition, and I found it to be much more challenging than the average Kaggle c ompetition. What made me begin was that there was non-trivial code that was available for me to examine. So there was a very challenging task and somehow the barriers to entry were lowered by having a non-trivial benchmark published.

Isabelle: Lots of people shy away when it's not too hard.

Martha: I think J. is saying it was "at his sweet spot" - they felt they could learn something

Isabelle: There is a sweet spot. Thinking of working on the Higgs-Boson Challenge w/ Balasz.

Evelyne: Can we refine who our target audience is? The sweet spot will be different depending on who the target audience is.

Jaffray: Unfortunately Ben's not here, and I'm going to be embarrassed if this already exists: An incredibly simply evergreen compet ition on the Kaggle site or elsewhere, something that you can see all the data on one screen. It's not clear to people who may be very bright but have never done this before would be helpful to them

Balasz: Kaggle has competitions that are open for the full year, and they also have scripts, etc.

Jaffray: So this exists

Isabelle: But this could be made simpler still.

Balasz: The pipeline is pretty complex, even for Titanic. Is this really a problem compared to AutoML? 10000 submitters per month on Kaggle.

Martha: You have observation bias - 10K seems a lot to us, but perhaps we should have millions of school kids working on this.

Isabelle: We should have Toy problems that help novices learn.

Balasz: I hear you - I've been thinking about this. You could break the pipeline down even further.

Jacob Abernathy: I've been excited by Kaggle & competitions, and now I'm faculty at Michigan, and I'm organizing students undergrad/grad an d putting them into teams of 4 and have them do the SpringLeaf Challenge, and working on an internal one using professor evaluations and determining the best professor, there are prizes, e tc. Initially what happened was that the best students got together and everyone else got owned. Now those best team winners said "Don't put us together - spread us out so we can help other teams".

I asked the students if $200 was enough. I asked if $1,000 was better, and all the students were like "This isn't about the money, this is a resume-builder for us".

Balasz: How many participants:

Jacob Abernathy: 20 active at one time. It's extracurricular. It's fun, beer & pizza goes a long way. Some teams become obsessed and get su per-focused. Michigan has a new thing called MIDAS data science institute. I want it to be fun, not a class room thing. I don't want them to be doing anything because it worked. We do Sp ringLeaf challenge from Kaggle and "RateMyProfessor.com" data scrape, etc. I spoke to Ben and he said he thinks we're probably the only school that's done this. Maybe we can do school competitions, like Michiga n vs. Ohio State, etc.

Audience: Great idea. I couldn't agree more about students benefitting from competitions - I personally transformed my own career and began using that as a guideline for hiring people. Assuming all else is equal, strong performance on a competitions makes a candidate a stand-out. This person is not in school just to get a grade, but they are good at what they do.

Balasz: Are challenges HR & Training tools, or are they solving bigger problems?

Amit: A competition is far more objective than a peer-review process. I dream of a future where papers are selected by challenges rather than exclusively from peer review.

Balasz: I tried to make this point this morning - we need to get out of the 18th-century method of evaluation and into something better. Do we have other use cases where we have participants?

Evelyne: There is a 3rd case, which is up-skilling, like at MS where we have a Hackathon Challenge. People do this who want to learn and up-skill. I wonder if this is a new category? It's not HR b/c it isn't tracked. Teams of 3-5 did much better than individuals. But interestingly, many of the experts didn't perform at the very top. What we did was bring them in to coach the particpants - it was about coaching some of the newbies. This was another way of helping with collaboration.

Isabelle: I like this lifecycle of challenges - to convert the winners into organizers - we need to see more of that. We hear 4 use cases: Teaching Tool, Problem Solving, Up-skilling / Education,

Audience: there's another aspect to this which is the difficulty level - maybe that's another dimension or aspect to this. If you analogize to video games as eSports, they have player rankings and leagues - perhaps. How can we make

Martha: If challenges are a sport, it's worth looking at the gamification literature and looking at player types in games. We have to be careful that we don't set up challenges w/ only one system of incentivization:

Jacob: To add - I did research in prediction markets, and you can formalize this as a way to run competitions. You get paid not for your ulti mate performance, but for your overall performance.

Jaffray: It's possible for someone who is 200th on Kaggle who actually developed a major advance, but because they don't know the re st of the process, there contribution is buried.

Isabelle: We started competitions and John Langford came up with the idea of rewarding novelty - how much a contribution improved the overall performance.

Jacob A.: This is called a "Shackley Value"

Balasz: Our problem now is that the code is completely open, which means that people can lose their impact.

Jacob A.: Git can handle this. The idea that an earlier contributor could be tracked over time, it may form its own complex DAG. F or example - if someone submits a good script to Kaggle, they don't get anything but bragging rights.

Isabelle: What incentives can we give to people to collaborate? One possible incentive is to put them in random teams, another is to make an implicit ensemble…

Jakub: This idea would be difficult if you enforce random team members who are geographically distributed.

Martha: Look at the literature about the definition of "work" - the only difference between "work" and "fun" is a small switch inside our heads that says "I don't have to be doing this if I don't want to". We don't want the incentives to become too good such that it "becomes work".

Isabelle: Has anyone used challenges as part of their class and made it work?

Ahmed Elshamli: We have one professor who is doing this

Isabelle: I did this doing a class at UC Berkeley, and that seemed pretty motivating to students.

Joseph Paul: Another way to look at it is "Why don't people want to work in teams?" Eliminating those reasons will make it easier.

Isabelle: I want to share an experience about Russian Summer School. We were concerned that the more advanced students would be disi ncentivized, but the low-skilled teams were able to help one another up-skill.

Balasz: Why must we create teams?

Evelyne: It depends on the challenge you are trying to solve.

Jaffray: I'm looking at these two questions and I think to incentivize more innovation, have people get "paid" more if they are able to move the ensemble up as it goes asymptotic.

Bram: It's a lot of work to put a marketplace together like Isabelle is saying, but it's worth it.

Jakub: What about making cooperation happen implicitly by breaking up the data into different groups and that way everyone needs t o cooperate in order to win.

Michele: This scheme is very similar to distributed resolution. This is more like ensemble, but my understanding is collaboration is a bit different from ensemble.

Jakob: But this creates the conditions that would cause people to have to work together.

Joseph: If you do this with multi-modal data, you would have a situation where one person couldn't be an expert in all those fields, and that would compel part icipants to cooperate.

Michele: That happens a lot in the placing task I discussed - audio people get together with computer vision people, etc. The Hidden Markov Model actually developed this way. They need to collaborate b/c they need something from each other.

Amit: You could set this up in a GitHub kind of manner, where whenever you submit code, you're changing the objective moderately. You keep assigning credits by the jumps people are getting.

Michele: One way to have people cooperate is to use a recommendation engine "You might like using this piece of code", etc.

Academic Torrents

Bram: There are lots of organizations where you can upload your data and get a DOI.

Joseph Paul: We looked at getting a DOI and it was prohibitively expensive.

Bram: Aren't there organizations that make huge amounts of data available for scientific purposes?

Joseph Paul: Yes, like the Harvard Dataverse, but they have not yet permitted us to integrate w/ them

Audience: You should see what Zenodo does - they are hosted at CERN and they are trying to become an OSS Project.

Michele Sabag

Michele: What would you like to have at the end of this discussion and what do you want to see happen?

Michele: One way to think about Education is to think of a MOOC as a Challenge - can we use our Challenge rules to run a MOOC and solve our education problem? Perhaps the lessons from organizing a challenge shares the same principles as organizing a MOOC.

Martha: It's important to think of summative vs. formative measures of success in class.

Martha: At Delft we have a solar energy MOOC, and the person who hosts the MOOC runs a study by interacting w/ his MOOC students. The MOOC becomes a method for conducting his research.

Isabelle: How can we democratize the process of creating course material.

Gael: Something called "Software Carpentry" created by Craig Wilson - trying to teach non computer scientists to use computers for conducting non-CS research. I think there is something to look at here. He was funded by Mozilla. This is one example of successful crowd-sourcing of educational material.

Michele: When there are good ideas, there are usually plenty of nice people who are willing to give money - perhaps I should turn the floor over to Jaffrey.

Jaffray: I'm taking proposals in this area for gifts to pay for Data Science projects & competitions. I've already given over $10M to the Data Science institute in my home town.

Isabelle: We need to make it easier for people in our community to network.

Martha: Yes, but that's part of the fun of it! You don't want to institutionalize it!

Isabelle: You can't deny that Google has improved the experience of the Internet

Martha: I can deny that! :) It's important, but maybe the more important thing is to think about it in more of a peer-to-peer system. We notice with MediaEval, some people think that the community should become as large as possible, but there is tremendous value in small groups coming together.

Audience: We need a Kaggle model that is open source.

Michele: We could create a recommendation engine for Kaggle like Netflix!

Audience: We want an Open Kaggle.

Gael: GitHub is proprietary, but they still made a difference was because they are really good, they worry about marketing, they worry about human factors, etc.

Balasz: Funding is so important b/c there is a lot of engineering cost.

Isabelle: I love Kaggle, we've been using Kaggle for some time, we helped conceive Kaggle - they have a particular niche even though they've been growing. The world of crowdsourcing is much bigger than just Kaggle. There are many things they will likely never address. In principle, there is no reason we should limit ourselves to one type of model. Like Martha said - connecting people so that there are loose connections between like-minded individuals. Martha and I are examples of this in that we met through Sergio. :)

Jacob A: (History of Kaggle, Jeremy Howard left after they pivoted to Oil & Gas, etc. Kaggle scripts are really interesting tech in terms of making this easier)

Michele: If we had "GitHub-Kaggle", would that fit the goal?

Jaffray: I've reached out to Kaggle and found them to be very difficult to work with.

Gael: Can you pinpoint exactly what the problems are?

Jaffray: I had a tough time connecting with Ben.

Martha: For MediaEval, Kaggle is out of our price range.

Audience: If we think thousands of challenges are possible, $10K per challenge is too much.

Jaffray: I want an Open Source competitor to Kaggle!

Balasz: There are 2 sites like this in Europe, one is dying and one is being born… What is CodaLab's status?

Evelyne: (CodaLab Pitch)

Martha: We need a MOOC on how to Organize Challenges!

Isabelle: Organizing a challenge is really about organizing an experiment.

Balasz: What is hardest?

Bram: Choosing the metrics - lots of challenges fail b/c they make it too hard or too easy or select the wrong metrics, etc.

Gael: Seems like we are talking about Education / Communication problems - it seems that we need people with didactic skills to be paid to write the content and market it, (e.g. - Software Carpentry)

Isabelle: Assuming we figure out a way to create this microcosm, with a cohort of people who can create and administer challenge, another missing piece is how to rate the quality of the challenge. Currently there are plebicited by the participants, which is a post hoc metric. How do we know the challenge was well-prepared?

Martha: Maybe when you set up the challenge you have a board / steering committee to certify the challenge.

Isabelle: A random idea is if we thought of having a bidding process to put prizes on challenges…

Bram: We do that now both for sponsors as well as calls for data.

Martha: It's data and reviewers. We need people who understand challenges. This gets back to the point that you need to educate people on how to administer challenges.

Isabelle: Years ago I started a competition program, and now every year they issue a call for competitions that is peer-reviewed.

Michele: Why must we have an exam to be a challenge organizer? Of course, it's good to have experience, but what happens exactly if a challenge is badly formed?

Isabelle: The most important thing was to write the proposal, because it forced us to have discipline.

Michele: What makes organizing a challenge so special? What is the difference between starting a challenge and starting a company, or a spinoff, or an association - these all require huge effort and clarity of thought. I'm thinking of Bram's example of eye exams in India.

Bram: but this is common b/c in the business world you have subcontractors, you evaluate them, etc.

Michele: One outcome of the challenge should be delivering to the public the results of the challenge.

Isabelle: I'm wondering if we have some sort of micro-funding mechanism where organizers could get $10K for developing and administering a successful challenge.

Martha: Currently there is something called ELIAS where the organizer gets 5,000 Euros per year and we can really stretch it.

Evelyne: Peer review seems good from a networking point-of-view, but on the downside, could it slow down innovation?

Balasz: In my experience, Kaggle doesn't filter their challenges, so there are lots of bad ones out there. Unless you pay them mor e for consulting, they don't help you set it up. For example, there was the "Portuguese Restaurant Challenge" which was bad, and there was a Physics challenge that had huge data leakage.

Michele: You have feedback and a fiter

Gael: This reminds me of what goes into "open review" - something that was great in GitHub is starring projects. When you select a challenge for a conference, you are giving them v isibility. Multiple tracks: Reviewed and not Reviewed.

Isabelle: But it doesn't need to be a review board, it can just be reviewed.

Martha: My main problems are educating challenge organizers (a place they can go to organize challenges) and also this ELIAS thing to incentivize challenge organizer excellence.

Michele: Last question - are there domains where you are craving to see a challenge organized?

Gael: In medicine, the #1 problem is access to data. I would love to have challenges for medical challenge, but these are not go ing to happen because the data is not going to be shared.

Henry: Yesterday, at the multi-modal workshop they said what the field needs is a benchmarking data set like ImageNet.

Martha: Tell them about multi-media comments

Audience: We need a technical infrastructure for doing this challenge thing. It can be public or private, doesn't matter - imagine "terrorist models" we need the technical infrasturcture to run this.

Mohammed: We need inverse kick-starter. I have a problem, come up with a solution - vs. "I have a solution, pay me to get it going." If it's a good project, Kickstarter recommends the project - they enhance the marketability of the project.

Isabelle: This idea of sharing data - part of this is hard because the data is not all lin the same format. If we incentivize people, it can happen.

Henry: Data science could benefit from some of the cyber-security competitions that are well-organized, like the NECCD stuff. ACM Programming.