BLOG@CACM

The NIPS Experiment

By John Langford
January 8, 2015
Comments (1)

PRINT

Corinna Cortes (http://research.google.com/pubs/author121.html) and Neil Lawrence (http://inverseprobability.com) ran the NIPS experiment (http://inverseprobability.com/2014/12/16/the-nips-experiment/) , where 1/10th of papers submitted to the Neural Information Processing Systems Foundation (NIPS (http://nips.cc)) went through the NIPS review process twice, and then the accept/reject decision was compared. This was a great experiment, so kudos to NIPS for being willing to do it and to Corinna & Neil for doing it.

The 26% disagreement rate presented at the NIPS conference (http://nips.cc/Conferences/2014/) understates the meaning in my opinion, given the 22% acceptance rate. The immediate implication is that between half and two-thirds of papers accepted at NIPS would have been rejected if reviewed a second time. For analysis details and discussion about that, see here (http://mrtz.org/blog/the-nips-experiment/) .

Let's give P (reject in 2nd review | accept 1st review) a name: arbitrariness. For NIPS 2014, arbitrariness was ~60%. Given such a stark number, the primary question is "what does it mean?"

Does it mean there is no signal in the accept/reject decision? Clearly not—a purely random decision would have arbitrariness of ~78%. It is, however, quite notable that 60% is much closer to 78% than 0%.

Does it mean that the NIPS accept/reject decision is unfair? Not necessarily. If a pure random number generator made the accept/reject decision, it would be 'fair' in the same sense that a lottery is fair, and have an arbitrariness of ~78%.

Does it mean that the NIPS accept/reject decision could be unfair? The numbers give no judgement here. It is, however, a natural fallacy to imagine that random judgements derived from people imply unfairness, so I would encourage people to withhold judgement on this question for now.

Is an arbitrariness of 0% the goal? Achieving 0% arbitrariness is easy: just choose all papers with an md5sum that ends in 00 (in binary). Clearly, there is something more to be desired from a reviewing process.

Perhaps this means we should decrease the acceptance rate? Maybe, but this makes sense only if you believe that arbitrariness is good, as it will almost surely increase the arbitrariness. In the extreme case where only one paper is accepted, the odds of it being the rejected on re-review are near 100%.

Perhaps this means we should increase the acceptance rate? If all papers submitted were accepted, the arbitrariness would be 0, but as mentioned above arbitrariness 0 is not the goal.

Perhaps this means that NIPS is a very broad conference with substantial disagreement by reviewers (and attendees) about what is important? Maybe. This even seems plausible to me, given anecdotal personal experience. Perhaps small, highly-focused conferences have a smaller arbitrariness?

Perhaps this means that researchers submit themselves to an arbitrary process for historical reasons? The arbitrariness is clear, but the reason less so. A mostly-arbitrary review process may be helpful in the sense that it gives authors a painful-but-useful opportunity to debug the easy ways to misinterpret their work. It may also be helpful in that it perfectly rejects the bottom 20% of papers which are actively wrong, and hence harmful to the process of developing knowledge. None of these reasons are confirmed, of course.

Is it possible to do better? I believe the answer is "yes," but it should be understood as a fundamentally difficult problem. Every program chair who cares tries to tweak the reviewing process to be better, and there have been many smart program chairs that tried hard. Why isn't it better? There are strong nonvisible constraints on the reviewers time and attention.

What does it mean? In the end, I think it means two things of real importance.

1. The result of the process is mostly arbitrary. As an author, I found rejects of good papers very hard to swallow, especially when the reviews were nonsensical. Learning to accept that the process has a strong element of arbitrariness helped me deal with that. Now there is proof, so new authors need not be so discouraged.

2. The Conference Management Toolkit (CMT (http://cmt.research.microsoft.com/cmt/)) now has a tool for measuring arbitrariness that can be widely used by other conferences. Joelle and I changed ICML 2012 (http://hunch.net/?p=2517) in various ways. Many of these appeared beneficial and some stuck, but others did not. In the long run, it's the things which stick that matter. Being able to measure the review process in a more powerful way might be beneficial in getting good review practices to stick.

Other commentary from Lance (http://blog.computationalcomplexity.org/2014/12/the-nips-experiment.html) , Bert, and Yisong (http://yyue.blogspot.com/) .

Comments

Cassidy Alan

March 09, 2015 10:11

As long as government interference doesn't drive up the cost of the Internet more...

The Internet offers an excellent review process. Items all timestamped for at least a localized priority, and open for all eyeballs. Even throw it open for all eyeballs to see. There can be some way of "officializing" it, like filters for who can vote up or down... And more, of course....

Displaying **1** comment