Join the discussion…

AS • 4 years ago

Do you have any references for examples of how models with poorly calibrated probabilities fail in practice?

^ | v • Reply • Share ›

Frank Harrell Mod ➚ AS • 4 years ago

Not specifically, but any clinician who is relying on a risk estimate of 0.1 to mean 0.1 when it's really 0.25 is in for trouble.

^ | v • Reply • Share ›

Vladimir Rainish • 4 years ago

Dr. Farrell,

I don't have any background in ML/AI so my question might be somewhat naive (or even more then that :-). I've recently attended a presentation about applications of ML for the cancer genomics using a specific tool.
Presenter mentioned that he used ML approach since it was very difficult to describe what happens in term of "regular/conventional" statistics. At the end of presentation I asked if , potentially/theoreticaly his or other tool can generate not just numeric answer but a symbolic statistical description of what's happened, so it can be applied for new sets of incoming data. His answer was that it can't be done , because of complexity of correlations etc. (BTW I mentioned , prior to my question , that I have no background in ML. since I wasn't sure it my question can be considered as reasonable).

I just thought that if there is an input data which follows simple normal distribution, ML can actually "discover it", so to speak (without checking a predefined hypothesis of normal distribution). May be ML can potentially derive statistics . What do you think ?

^ | v • Reply • Share ›

Frank Harrell Mod ➚ Vladimir Rainish • 4 years ago

Good questions. ML is hard to interpret because of all the connections between predictors that are allowed (interactions). Statistical models are often chosen because of their easier interpretation. And it is not always hard to describe the biological relationships with statistical models, if the features have mainly an additive effect. We need a global measure of additivity that will help us decide between SMs and ML. ML will sometimes give valid predictions on new data, if the original dataset was representative and the ML did not overfit. But the way you get the predictions requires a complex program because the predictive patterns are allowed to be complex.

^ | v • Reply • Share ›

Vladimir Rainish ➚ Frank Harrell • 4 years ago

Prof. Harrell , thanks a lot for such a quick response : I posted my question Friday evening and certainly didn't expect to see your answer Saturday morning.
If I understand you correctly, in case interactions between predictors are of completely or overwhelmingly of linear nature (additive) , then there is no advantage of ML over SM, moreover SM is even preferable because it actually provides a model, rather then a "only" answer.
I am also wondering if the "global measure of additivity" you mentioned in your response can by itself (as a hypothesis) be tested/discovered by ML. It's not a full model, but even knowing how much "addititive" the interactions are is valuable by itself (at least that's my interpretation)

^ | v • Reply • Share ›

Frank Harrell Mod ➚ Vladimir Rainish • 4 years ago

More good questions. Just to clarify terminology I only use 'interactions' to indicate non-additive effects. So I would say that ML is primarily advantageous when you expect complex non-additive (interactive) effects and you can't think of how to pre-specify the interactions. SM are primarily advantageous when the effects are largely additive. I don't know how to envision a global assessment of non-additivity along the lines you described unless you frame it as a contest between SM and ML in one dataset. But by then you've done double work. More feasible may be to get a statistical estimator of the total impact of all two-way interactions. Methods needed.

^ | v • Reply • Share ›

Vladimir Rainish ➚ Frank Harrell • 4 years ago • edited

Prof. Harrell , thanks a lot for your response and clarification. A thought I have after reading and thinking about your last response :

can an ML vs SM "contest" you mentioned be an actual integral part of training

process ?

And even more, why ML and SM should be mutually exclusive ?

As in many cases in science, life, industry etc , some sort of cooperational competition (or competitive cooperation) between them can provide the best possible solution and may be they will be able not just barely "tolerate" each other presence but even create some form of synergy ?

A dynamic, "self tuning/learning" combining of these two worlds (ML and SM) doesn't look, at least immediately, as a complete non-starter.

Of course "framing the contest" (not even the contest , but "competitive cooperation") in mathematical terms might be untrivial, but hopefully not impossible.

It would be very fascinating to see as an arrow/pointer on ML/SM gadget/scale indicating a balance between ML and SM contributions moves/shifts as process of training is advancing.

• Reply • Share ›

Frank Harrell Mod ➚ Vladimir Rainish • 4 years ago

The combination approach may be valuable but I was referring to something else: an indirect way to measure how much non-additivity (interaction) there is in the data. Have a personal contest for just your data. Develop independently a SM and a ML, forcing the SM to be additive but allowing everything to be nonlinear. Put no restriction on ML. Compare the two in some unbiased (not always easy) fashion using some function of the log-likelihood (including pseudo R^2, AIC, likelihood ratio chi-square minus effective number of parameters, leave out one log likelihood, cross-validated log likelihood). Try to learn from that how much non-additivity ML is capitalizing on.

• Reply • Share ›

Alan Hochberg • 4 years ago

I wonder if Avati et al. aren't predicting mortality not through objective assessment of medical information, but rather through changes in the EHR as the physician perceives that the patient is approaching end-of-life. They might order fewer tests, eliminate imaging studies almost entirely, make less-frequent and less-detailed notes, allow increased dosing of opiates and make other medication adjustments, etc. I'm not convinced that restricting the prediction dates is enough to avoid all the subtle and obvious ways that circularity could come in.

• Reply • Share ›

Frank Harrell Mod ➚ Alan Hochberg • 4 years ago

Excellent points. For some purposes, and perhaps consideration of palliative care could be one of them, I can see using physician behavior as important predictors. But more generally I think machine learning applied to EHR should be done quite cautiously and use objective features to the extent possible. At the very least, ML experts owe it to us to carefully describe the features that dominate the predictions.

• Reply • Share ›

Anil Patwardhan • 5 years ago

Figure 3 and an AUC of 0.93 reminds me of this old paper "What price perfection?...." by Diamond (1992). https://www.ncbi.nlm.nih.go.... He elegantly explains the inherant tradeoff between a discrimination metric (AUC) and an assessment of calibration, something that is often lost on the machine-learning community that frames everything as a classification problem.

• Reply • Share ›

Frank Harrell Mod ➚ Anil Patwardhan • 5 years ago

Hi Anil - when that paper first came out I agreed with it more than I do now. Now I think it's mainly machine learning (especially ML based on classification) that decouples calibration and discrimination. With penalized maximum likelihood estimation you can almost ensure calibration without killing discrimination.

• Reply • Share ›

Michael Webb • 5 years ago

If I were to create a model to predict probabilities instead of to classify, would I validate the model the same way (e.g. a test set with a binary outcome variable)? This makes a lot of sense for a classification model but validating predicted probabilities using binary outcome data is less straight-forward.

• Reply • Share ›

Frank Harrell Mod ➚ Michael Webb • 5 years ago

That's fully worked out in my book Regression Modeling Strategies and in its course notes at http://fharrell.com/doc/rms... . The basic idea of calibration is to model how predicted risk is related to actual outcomes. Using a nonparametric smoother or logistic regression fit you can estimate this calibration curve, and the bootstrap can bias-correct if for overfitting if you are evaluating calibration on the same sample used to develop the model.

 • Reply • Share ›

AS • 5 years ago

"The predictive factors are measured at obvious times. One can be certain that the model did not use information it shouldn't such as the use of certain treatments and procedures that may create a kind of circularity with death."

Would you mind elaborating on this part? What is meant by "obvious times" and "circularity with death"? Is the idea that you don't want to include information that is predictive because of policy/treatment practice rather than underlying biology?

 • Reply • Share ›

Frank Harrell Mod → AS • 5 years ago

I didn't describe those very clearly. "Obvious times" include times such as the day of an index admission or day 3 after such an admission. Certain treatments decisions are too subjective and hard to standardize for use as predictors. So yes there are some practice patterns we may not want to use. And some prediction studies use measurements too close to death (that's a tough call) making the prediction too easy.

 • Reply • Share ›

Related

- Damage Caused by Classification Accuracy and Other Discontinuous Improper Accuracy Scoring Rules
- Classification vs. Prediction
- Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements
- How Can Machine Learning be Reliable When the Sample is Adequate for Only One Feature?
- In Machine Learning Predictions for Health Care the Confusion Matrix is a Matrix of Confusion