Mike Beyer • 4 years ago • edited

Thanks for the great piece Dr. Harrell! Completely agree with you.

I've also found that trying to make machine learning (ML) loss functions do "double duty" as decision models is really not a good idea most of the time. You lose a lot of information by not fitting a calibrated probability model to the data.

For example, in the case of customer churn, it may be economically worthwhile to target retention efforts at customers that have only a 10% chance of churning -- it all depends on the economic/decision-making context! In this case, using the a 50% "most likely" classification rule would result in sub-optimal revenue retention.

I have a background in operations research (OR). I often find that ML and OR are complementary tools, where OR's decision-oriented models utilize predictions and estimates from ML and statistical models. However, I have an issue with ML models using loss functions (implicit or explicit) as proxies for a well-formulated decision model that takes into account the utility function and constraints. A mathematical programming formulation seems to be much better suited for this latter task.

Of course, most OR people see that many ML models are also instances of optimization problems -- again, it's a useful exercise to understand the form of this optimization problem to see if you agree with what it is trying to optimize.

Thanks again for the great post. Very helpful!

39 • Reply • Share ›

Frank Harrell Mod ➚ Mike Beyer • 4 years ago

I wish I could have said it this well. Thanks Mike!

• Reply • Share ›

cbeleites • 2 years ago

Dear Frank,
Nice post.

I'd prefer to call what is "prediction" to you "probabilistic prediction" or "probabilistic outcome", since I like many others already use prediction to denote models whose output is in a pre-specified domain (regression, classification, ...). "Probabilistic outcome" is not that much longer and it is IMHO much less ambiguous.

From this point of view, a classifier may have nominal, score, or probabilistic output (in order of increasing information content). And yes, I also prefer probabilistic :-)

OTOH, if a classifier with nominal output is set up with an appropriate figure of merit that is optimized, I'm fine with that as well. Appropriate would for me typically include varying costs of various ways of misclassification, being a proper scoring rule and preferably also appropriate treatment of the underlying prior probabilities for the different classes. I.e. there is IMHO much more to choosing your classifier than the thowing a wildly varying set of algorithms to the wall to see whether something sticks.

(Curiosity) Why do you consider chemical composition as suitable for classification? As analytical chemist and chemometrician, chemical composition to me is first and foremost *continuontus* in |R^n - possibly on a restricted subspace for mixtures (unless we go into single molecule analysis, where counting a discrete number of molecules becomes important). So the underlying nature directly points to regression. We may consider whether the concentrations are above/below some threshold in a next step, and that may indeed be better modeled as predicting probabilities rather than nominal (the relevant norms in analytical chemistry derive above/uncertain/below from probabilities, btw).

Also, I like to further distinguish classifiers into discriminative vs. one-class/single-class vs. what I call "regression in disguise", e.g. above mentioned presence/absence of an analyte. *All of them* may be set up to yield probabilistic output or not. That is IMHO just another aspect into the choice of an appropriate algorithm/model.

--cbeleites

1 • Reply • Share ›

David Rosen • 2 years ago

Just in the interest of mutual communication, I'd like to point out that in machine learning, a "classification problem" usually means one in which the target (dependent) variable is categorical, even when the goal is to estimate class probabilities, in which case it may be further qualified as

"probabilistic classification" or "soft classification" as opposed to "hard classification" where the goal is to jump straight to outright choice of the predicted class for each observation. This is the sense in which "logistic regression is a (probabilistic/soft) classification method", although in statistical parlance you might say it is a regression method. In machine learning, a regression problem is usually defined as one in which the target (dependent) variable itself is continuous.

1 • Reply • Share ›

Demetrius K. Green David Rosen • 2 years ago

I am glad I continued reading the comments in hopes that someone would highlight the differences in terminology and nomenclature! It is indeed where much of the bifurcation between more traditional statisticians and ML practitioners comes into play. Since ML and computer science have more of an intimate relationship, but it has been, at least in more recent history, more or less "guided" by statisticians such as yourself, Professor Harrell!

1 • Reply • Share ›

Frank Harrell Mod David Rosen • 2 years ago

Thanks for writing. We need to work to fix the nomenclature error that has apparently pervaded the ML literature. "Classification" as an active process means only to "classify" so the term is definitely being misused. One can speak of "previously classified observations" (categorical data) or actively classifying an observation into a discrete number of classes. Imprecise language causes imprecise thinking (which you're definitely not guilty of).

• Reply • Share ›

David Rosen Frank Harrell • 2 years ago • edited

Good points. It's hard to revise entrenched terminology, but I think the term "decision" may be less ambiguous than "classification". We might estimate the probability of rain, while the choice to bring an umbrella is perhaps better described as a decision than as a classification of the day as a "rain day". We don't require a discrete choice/classification until we are forced to make a decision based on available information including the (estimated or believed) probabilities of different outcomes.

1 • Reply • Share ›

Frank Harrell Mod David Rosen • 2 years ago

I'm saying this just to be picky :-) - when you decide to take an umbrella you are not concluding that it will rain. You are playing the odds and deciding whether to act as IF it will rain. Probabilistic thinking works.

• Reply • Share ›

Jason H. Moore, Ph.D. • 6 years ago

Probability machines!

https://www.ncbi.nlm.nih.go...

https://www.ncbi.nlm.nih.go...

1 • Reply • Share ›

I_love_han_hye_jin • a year ago

Dear professor, If i use probabilistic forcasting (output probability) in imbalanced case (say ratio between majority and minority class is 100 : 1), I saw that the output probability of data points from majority class is very High (say 99% or so), and much higher than output probability of data points from minority class. The problem is: In case of abnormality detection in banking or in many cases in medical study, we just want to detect the minority class. So i want to increase the probability of minority class. What can we do in this case. I searched many sources on the internet and papers, but did not see any solutions to this problem. Maybe because people in machine learning just care about accuracy, then they just apply under/over sampling.

• Reply • Share ›

Frank Harrell Mod I_love_han_hye_jin • a year ago

Don't think "imbalance". Think "how do we sample customers so that they are representative of future customers" and estimate probabilities from that sample. Don't think "detect minority class"; think "how do we most accurately estimate risk and how do we turn that into a decision rule". There are two types of decision rule, first not really being a formal decision rule: (1) a lift curve where you decide how many customers N you can afford to deal with and deal with the N highest risk customers, or (2) a formal decision rule where you compute the best decision based on expected gains/losses/utility. The best decision is best on the individualized risk estimate, combined with the utility/loss function. Classification plays no rule.

• Reply • Share ›

cbeleites • 2 years ago

Dear Frank,

Nice post

Nice post.

I'd prefer to call what is "prediction" to you "probabilistic prediction" or "probabilistic outcome", since I like many others already use prediction to denote models whose output is in a pre-specified domain (regression, classification, ...). "Probabilistic outcome" is not that much longer and it is IMHO much less ambiguous.

From this point of view, a classifier may have nominal, score, or probabilistic output (in order of increasing information content). And yes, I also prefer probabilistic :-)

OTOH, if a classifier with nominal output is set up with an appropriate figure of merit that is optimized, I'm fine with that as well. Appropriate would for me typically include varying costs of various ways of misclassification, being a proper scoring rule and preferably also appropriate treatment of the underlying prior probabilities for the different classes. I.e. there is IMHO much more to choosing your classifier than the thowing a wildly varying set of algorithms to the wall to see whether something sticks.

(Curiosity) Why do you consider chemical composition as suitable for classification? As analytical chemist and chemometrician, chemical composition to me is first and foremost *continuontus* in |R^n - possibly on a restricted subspace for mixtures (unless we go into single molecule analysis, where counting a discrete number of molecules becomes important). So the underlying nature directly points to regression. We may consider whether the concentrations are above/below some threshold in a next step, and that may indeed be better modeled as predicting probabilities rather than nominal (the relevant norms in analytical chemistry derive above/uncertain/below from probabilities, btw).

Also, I like to further distinguish classifiers into discriminative vs. one-class/single-class vs. what I call "regression in disguise", e.g. above mentioned presence/absence of an analyte. *All of them* may be set up to yield probabilistic output or not. That is IMHO just another aspect into the choice of an appropriate algorithm/model.

--cbeleites

• Reply • Share ›

Frank Harrell Mod cbeleites • 2 years ago

This is not appropriate terminology. "Classifier" comes from the active verb "classify" which connotes putting things into classes. Classification represents a forced choice. "Prediction" implies an estimate that is not tied up with decision making. In the regression world, prediction involves an almost always linear combination of predictors regardless of whether Y is discrete or not. Probabilistic prediction or probabilistic outcome are not bad terms but if you want to be that explicit I would say *probability estimation* or that you are using a *direct probability model*. I'm not sure where I got the chemical composition notion.

• Reply • Share ›

Jake Oaknin • 3 years ago

Dear Prof. Harrell

In "Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization", Tong Zhang proved that many of the usual classification methods (AdaBoost, for instance) can be actually regarded as estimators of the conditional in-class probability. So, even though they are set to minimize the 0-1 loss (that implies equal cost for wrong decisions for both classes), they actually achieve a more general goal: an estimate of the conditional in-class probability.

He also proved that these classifiers can be seen as generalizations of the classical, maximum-likelihood Logistic Regression that use a different bregman distance to measure dissimilarity between their estimates and the true conditional in-class probability, and hence that Logistic Regression doesn't enjoy a special status among them.

I would be very grateful if you could help me understand, if my understanding is erroneous

• Reply • Share ›

Frank Harrell Mod Jake Oaknin • 3 years ago

Thanks for the note Jake. I think we are mixing two ideas: (1) machine learning classifiers that never involve probabilities and for which the output is a forced choice that cannot be converted to a probability (which is the majority of uses) and (2) machine learning classifiers that involve something (e.g. ensemble voting) that can be captured and turned into a probability. I think you are referring to (2) whereas I'm referring to (1). For (2) any resulting forced choice binary output is just ignored.

• Reply • Share ›

Jake Oaknin Frank Harrell • 3 years ago

Dear Prof. Harrell

I actually meant classifiers of kind (1) like AdaBoost that originated in the ML community and that, in principle, were meant to provide a hard decision 0/1 that minimizes the 0-1 loss. It's my understanding that it was later proved that these

classifiers also estimate the conditional in-class probability with the same asymptotic guarantees as Logistic Regression. If so, the original decision they provide can be ignored in favor of any other decision based on the conditional in-class probability.

Reply • Share ›

Frank Harrell Mod ➚ Jake Oaknin • 3 years ago

I can't follow that. A method that outputs a binary quantity does not estimate a continuous quantify so can't get you a probability. Even if you can trick them into giving you a probability I doubt that it would achieve perfect calibration-in-the-small.

1 Reply • Share ›

Jake Oaknin ➚ Frank Harrell • 3 years ago

My understanding is as follows:
the goal of these large-margin classifiers is to estimate a decision function f(x), such that sign(f(x)) minimizes the 0-1 loss, i.e they err whenever Y f(X) <= 0 (with the class label, Y = -1,1). In setting the problem this way they are assuming equal cost for errors on both classes. To achieve their stated goal, however, these methods minimize a convex upper bound to the 0-1 loss, and for some choices of that upper bound, the estimate f(X) can be related to a consistent estimate of the conditional in-class probability. Both Logistic Regression and AdaBoost fall within this category. SVM doesn't. Here are the two papers from which I got this understanding. Please, let me know if I misunderstood them
- "Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization", by Tong Zhang
- "Coherence Functions with Applications in Large-Margin Classification Methods", by Zhihua Zhang, Guang Dai and Michael I. Jordan

Reply • Share ›

Frank Harrell Mod ➚ Jake Oaknin • 3 years ago

I appreciate the additional information. Logistic regression directly provides the needed probabilities so we can leave that out of the discussion. I gather from AdaBoost that it relies on a "confidence in prediction" level (|f(X)|) to turn things into probabilities. This requires the machine learning algorithm to provide continuous f(X), and I can see your point for the subset of algorithms that do that. This subset excludes those methods that provide only the forced choice binary output. As a statistician, these methods seem very indirect when compared to direct probability estimators.

Reply • Share ›

Steve Pitts • 5 years ago

Fascinating and completely new way of thinking about this topic for me. Spiegelhalter article next on my agenda. But there are some huge domain gaps, some of it just vocabulary, between practicing general docs (not specialists) and academic statisticians. I can only imagine what a "machine learning advocate" is because I've never met one or even read of one. The motivation for thinking about numbers comes naturally to clinicians during their training, from statements like "50% of prostate nodules are due to cancer". You can function perfectly well without questioning this of course, but most docs I think would adapt this number to their own practice, i.e. revise this probability. Seeing self-referred unselected patients is like diving into the ocean. Without some immediate classification you drown in uncertainty. I would like to know "does this patient have a gallstone" when I am too busy to get the ultrasound machine, not "what is this gallstone patient's prognosis". In my understanding, prognosis and treatment require prediction and a time-to-event column, but diagnosis does not, if prediction refers to the future (cf Yogi Berra). As far as I know nobody has done a machine learning study with all possible interactions for gallstones. But I bet there have been unfunded attempts to use multiple logistic regression as a classification tool, with very few "cases" and maybe a dozen independent variables, so over-fitting and imprecision will be threats. And then published the result as a "decision rule". This rule ought to subject to modification by setting, e.g. Indian reservation has much higher prevalence, independent of the covariates. Huge files from RCT registries are not available for these sort of questions, which are far more important to practicing general docs than whether to use a particular pharmaceutical or device.

Reply • Share ›

Frank Harrell Mod ➚ Steve Pitts • 5 years ago

Thanks very much for this discussion Steve. The "50% of prostate nodules are due to cancer" has a subtle problem. Don't you want to know instead the probability of cancer given the characteristics and number of the nodules? You raised several other good issues, with the only response for now that I can think of is that using classification instead of risk estimation won't take into account that utilities change, even from day to day, making for different decisions. For example, in making the decision to hospitalize a patient you would probably act differently if the hospital is overflowing vs. has immediate bed availability. I've always felt that if

someone without a college education could master probabilities to win at poker, physicians should also be able to get pretty good at probabilities. But as you mentioned there's not enough time in the day to do everything optimally.

• Reply • Share ›

Steve Pitts ➚ Frank Harrell • 5 years ago

Well yes, I was unclear. What I meant was P(D|Finding), "predictive value" or "post-test probability". I know that you have spoken ill of the "inverse", P(F|D). My point is that this prior probability/prevalence (got to be careful with the vocabulary) is offered to medical students by urologists as an unalterable fact like an anatomic structure, maybe to make them aware that if you feel a nodule you should be alarmed; but even if true in urologist's office, it will not be true for docs who don't limit their practice. This is a huge sore point. Real ER docs mock FPs for referring them pts who don't have appendicitis, and real surgeons mock ER docs for the same thing. For each of these docs the patient has a different "prior", which is just the prevalence of appendicitis among all referrals. And for each the penalty for getting a "non-case" is different. So I agree with the idea that utility functions too are important and overwhelming.

But Diagnosis is still a form of "classification" (the present, i.e. "divination") rather than Prediction (the future) if I understand what you mean by "prediction". But maybe this is not the distinction you had in mind? The language is sloppy, eg "prediction rules" and "predictive value" can be about either diagnosis OR prediction. I think maybe what I don't understand is your defining classification as implying an action. Once I know the patient is classified as having a gallstone, I might take any number of different actions depending on the situation. But now I know it's not a ruptured abdominal aortic aneurysm, I can be much more confident in making decisions, can see my way out of the ocean.

• Reply • Share ›

Frank Harrell Mod ➚ Steve Pitts • 5 years ago

Great points. I have been sloppy in how I use the term 'prediction'. I always mean it to include estimating the probability of a current (but hidden) state (e.g., diagnosis), in addition to forecasting a future event or measurement.

With Bayesian modeling, the prevalence is a prior distribution, i.e., we don't have to nail down a single value. You are using prevalence in the traditional way it's used in medical diagnosis, where we use a single number. I don't find this extremely helpful because everyone has a different idea of prevalence. For example does the prevalence of pregnancy in the US include males? Females younger than 15? Older than 70? Conditional probabilities are much better defined, and everything is conditional on something. This is one of many reasons that I highly question the way this topic is taught in medical school. I'd rather see us start with the idea of a cohort that allows us to estimate P(disease | patient characteristics and test results). It is fruitful to think about how a test moves the probability when compared to a reference point of a completely "normal" test result rather than some kind of mean test result that is envisioned when updating prevalence to a post-test probability. When you use logistic models in this setting, they are more straightforward and flexible than talking about updating a "prevalence".

Steve Pitts ➚ Frank Harrell • 5 years ago

I love that you're thinking of poaching in the medical school curriculum, I've been poaching in epi/stats for decades, with just a few arrows in my back (have practiced emergency medicine mostly full-time, retired from practice now). But you're up against a huge edifice built on Sn/Sp, including Sackett/Guyatt EBM and its stations of the cross, and the Pauker/Kassirer test/treatment threshold models. Even the most thoughtful primary docs live and die based on algorithms that were developed using these approaches. I don't think Duke and Vanderbilt have been in the forefront of these movements though: I wish you a crashing success. BTW, I have a book called "Classification and Prediction" by Pepe, published 2003. It says NOTHING about machine learning. It's time to update the "clinical math" curricula to the era of big data since many docs (esp generalists) may have to fight against "decision aids" made possible by the EHR. I see from previous comments that there is lots of interest in your technical subject now. I very much appreciate your openness to people like me!

• Reply • Share ›

Frank Harrell Mod ➚ Steve Pitts • 5 years ago

Nice points Steve. It is frightening to me how true it is what one professor said "I can teach my students anything I want. Getting them to unlearn what they already know is another matter." I believe that Sackett/Guyatt EBM with respect to diagnosis has done some damage. This way of thinking, to me, has

wasted a lot of time and created unnecessary complexity. I don't think machine learning will be a game changer hear. And I have some reservations about the Pepe approach.

Reply • Share ›

Frank Harrell • 5 years ago

Terrific and sickening example. Makes me wonder how many 'data scientists' understand data science.

Reply • Share ›

Z Bicyclist • 5 years ago

A good example of the overclassification thinking occurs in Larose and Larose's textbook, Data Mining and Predictive Analytics, 2nd edition, on page 422. Since all four combinations of two binary predictors made the same classification prediction ("won't churn", although as different probabilities), they recommend undersampling the data so that some of the probabilities are now greater than .50 and "churn" can be the classification outcome.

As you note, few people in marketing would do this. The probability of losing a customer (churning) would not have to be more than 50% to trigger much concern.

Reply • Share ›

Frank Harrell • 6 years ago

Thanks for the nice comments. I could be proven wrong but I think that the majority of people do not want to be told to bring an umbrella. Classification assumes in effect that everyone has the same utility function, which I know is not the case. My experience with additivity is that I had a grant with Phil Goodman (PI) to study neural networks vs. logistic regression in large medical outcome databases. We found no important interactions in any of the variables in any of the databases.

Reply • Share ›

Matthias Pierce • 6 years ago

Thanks very much for this thought provoking piece! In general, I agree with your point that, in many contexts, predicted probabilities, and their error, will have greater utility than classifications, but acknowledge that at the end there will always be a binary decision. With regards to the umbrella example, I think most people want a recommendation of whether they should bring an umbrella out or not, only some will want the probability of rain. Your preference will likely be dictated by your understanding of risk and uncertainty, and whether you are at the extremes of being worried about getting your hair wet or have a particularly cumbersome umbrella! In the biomedical sciences, I am not clear on what the preference of clinicians would be, but I would hazard a guess that it would context specific.

Also, I am interested in your assertion that the additivity assumption is approximately true ‘much of the time’. Is there a mathematical proof for this? I have a hunch that this is correct, given Taylor-type expansions of most data-generating functions, but given that real-world data comes from unknowable constructs, I am not clear how this can be justified.

Reply • Share ›

Frank Harrell • 6 years ago

I did a very quick read of the second paper for which you are a co-author. It doesn't seem to make the same mistake of underfitting logistic regression as the first paper made (if I'm reading it correctly). It possibly makes the opposite error because I didn't see appropriate penalization used in the logistic regression description. Logistic regression is often superior to machine learning for dealing with 2-way interactions, but you need to apply a penalty function. In my book Regression Modeling Strategies I show how to apply proper hierarchical penalties, e.g., least penalty on linear main effects, more penalty on nonlinear main effects, then on linear interactions, and most penalty on nonlinear interactions. In your case it would just involve putting a fairly heavy penalty (using effect AIC, etc.) on all the linear interaction terms. On a separate issue, the gold standards for comparing various models are the out-of-sample log likelihood (logarithmic probability scoring rule), the mean squared error of predicted logit, and mean absolute error of predicted logit and predicted probability. Precision, as you studied, is also important.

Reply • Share ›

Frank Harrell • 6 years ago

P.S. I was referring to the first paper you listed. Haven't looked at the second yet.

Reply • Share ›

Frank Harrell • 6 years ago

The methods for logistic regression are not well described in the paper, but I strongly suspect that they used logistic regression in a way that ignores every advance since logistic regression was invented by DR Cox in 1959. Some of the advances include regression splines, tensor splines, and penalized maximum likelihood estimation. The calibration curve they published for logistic regression

is flat, making me suspect that they used a vanilla logistic regression with only linear terms when the data were generated to be highly nonlinear. That could have been fixed trivially. So perhaps they gave machine learning every advantage and logistic regression no advantage. If this is indeed the case, that paper is worse than useless.

• Reply • Share ›

Frank Harrell • 6 years ago

There is an approximate way to correct the intercept based on relative odds of disease in the training and the target population, but I've forgotten the reference. The most rigorous way to do this is to have real-world data and to fit a model with just an intercept and with an offset term: the log odds from your model from the oversampled-disease dataset. The new intercept estimate will be the best available frequentist estimate of the correction you need for the intercept to apply your original model to the real world. You can always give up on the idea of estimating absolute risk and just provide relative odds, once you select a reference point (e.g. subject with covariates all equal to the median or mean).

• Reply • Share ›

Mehdi Rostami → Frank Harrell • 2 years ago

Is the reference this: https://gking.harvard.edu/f...
King, Gary, and Langche Zeng. "Logistic regression in rare events data." Political analysis 9.2 (2001): 137-163.

• Reply • Share ›

Frank Harrell Mod → Mehdi Rostami • 2 years ago

Thanks for that reference.

• Reply • Share ›

abbas Al-Shimary • 6 years ago

Many thanks for the interesting article. I am currently working on a data set obtained from a clinical trial in which the prevalence of disease (~ 50%) is by design is significantly higher than that observed in the real world (~ 15%), I am using logistic regression. Your article made me think whether some calibration is in order to apply this model to real world data? If yes then I would be grateful if you could make few suggestions.

• Reply • Share ›

Frank Harrell • 6 years ago

Excellent points all. The "no human in the loop" type of ML classification in my view works best when the signal:noise ratio is high, and only works when one does not desire to use utilities or the utilities are unknowable but we have some vague belief that the classification is implicitly using a reasonable utility function. It also should be noted that many comparisons of performance by ML with probability estimators such as logistic regression have been hurt by the use of an improper accuracy scoring rule.

• Reply • Share ›

Frank Harrell • 6 years ago

It may be just semantics but I don't see a lift curve as supporting classification. True you can solve for a cutpoint in predicted probabilities that yields the first n from a lift curve, but the lift curve can be based on miscalibrated probabilities, relative risks, relative odds, etc., and still work fine. But if you have a probability you have so much more. For example a marketer could change the form of advertisement when the probability of purchasing is lower but the customer is still worth pursuing. Classification just gets in the way of that.

• Reply • Share ›

Unknown • 6 years ago

The roots of machine learning are in settings where one wishes to write a program to make automated decisions (such as character recognition, speech recognition, or computer vision). Attempts to write such programs by hand failed. Machine learning applied to large data sets has succeeded very well. In these settings, there is no human in the loop to look at probabilities or confidences, and there is no desire to make statistical inferences or test scientific hypotheses. In such settings, methods that are trained 'end-to-end' to perform the task have generally given better results than methods based on probability models. This is "Vapnik's Principle" that one should not solve a harder problem (i.e., probability estimation) as an intermediate step to solving an easier problem (i.e., classification). There is also an interesting analysis by Shie Mannor and his students showing that the linear Support Vector Machine is a robust classifier, which is a property that few probabilistic methods share.

But of course as "machine learning experts" started looking at more subtle decision problems, they have reached the same conclusion: in many tasks it is important to estimate conditional probabilities. So today's deep neural networks are essentially multinomial logistic regressions (with very rich internal structure). And machine learning experts have been studying proper scoring rules to understand which loss functions give desired results. The ML Experts at Google and Microsoft are building causal models using propensity scores to make advertising decisions. Many of us employ

Markov Decision Process models to understand optimal sequential decision making.

In short, your depiction of "machine learning experts" is a straw man that may be useful for your argument but is not representative of the good work in ML. Of course, anyone can call themselves an ML researcher (or a statistician) and apply tools naively. Given the hype around ML/Data Science, thousands of people are doing exactly this, unfortunately. --Tom Dietterich

Reply • Share ›

Noah Motion • 6 years ago

I agree that the lift curve alone is not a classifier, but it supports classification. Or, put another way, it functions as part of a classifier, wherein each marketer's classification rule (at any given time) is determined by their budget (at that time), which in turn determines the number of potential customers they can target.

Reply • Share ›

Mark • 6 years ago

It's obviously different if the marketer adjusts how much they spend on advertising to each person based on the probability. The example you provided suggested a fixed cost of marketing to a person and so attempting to maximize revenue by targeting the top N most likely.

Reply • Share ›

Mark • 6 years ago

How is it functionally different from using a breakpoint other than 0.5 to convert probabilistic predictions to classifications? The marketer is saying, I want a classification model that classifies N people as 'market to' and all others as 'do not market to'. The break point for converting the probabilities to labels slides till they get what they want. This isn't terribly different (functionally) from adjusting a breakpoint to improve measures like Sensitivity/Specificity/F1.

Reply • Share ›

Frank Harrell • 6 years ago

The lift curve does not use classification in any way. It uses the predicted probability of purchasing, or anything monotonically related to that probability. And the point where one stops advertising to customers will vary with the advertising budget.

Reply • Share ›

Noah Motion • 6 years ago

This is a thought provoking post. Thanks for writing it (and, more generally, thanks for creating this blog).

It seems to me that you're defining "classification" too narrowly here, though. For example, you write:

To get the "biggest bang for the buck", the marketer who can afford to advertise to n persons picks the n highest-probability customers as targets. This is rational, and classification is not needed here.

This seems like classification with marketer-specific rules to me. The lift curve describes the range of values that could, in principle, be used to classify customers as targets or non-targets, and each marketer is free to implement a rule as desired.

My own training is primarily in statistics and mathematical psychology (focusing mostly on signal detection theory and various related models of perception and [statistical] decision making), and I've only fairly recently started to dig into the machine learning literature. So maybe I have an overly broad definition of what counts as classification.

In any case, I'd be curious to hear more of your thoughts on this.

Reply • Share ›

Ray • 6 years ago

Thank you Professor Harrell
for this GREAT article.
(I had never thought about it
from this clear angle...).

btw:
reached your Blog article
via your (new) Twitter acct! :-)

@SF99
San Francisco

Reply • Share ›

Frank Harrell • 6 years ago

Nice comments Keith - thanks. I didn't make this very clear, but probability has many roles including probability models for data and understanding individual calculated probabilities related to decision making and more. I was discussing more the latter.

• Reply • Share ›

phaneron0 • 6 years ago

> Probabilistic thinking and understanding uncertainty and variation are hallmarks of statistics.

I certainly think it should be and I do think there is a subset of the statistics discipline that understands statistics as primarily about conjecturing, assessing, and adopting idealized representations of reality, predominantly using probability generating models for both parameters and data.

Not sure if its the majority - there is another prospective on statistics, as primarily being about discerning procedures with good properties that are uniform over a wide range of possible underlying realities and restricting use, especially in science, to just those procedures. Here the probability model is de-emphasized and its role can fade into background technicalities.

Also, starting with probability models and explicating their role in representing reality well enough so that we can act in ways that are not frustrated by reality, does seem hard for people. Perhaps more so with those going into machine learning and data science.

Hope you enjoy blogging.

Keith O'Rourke

• Reply • Share ›

Related

- Is Medicine Mesmerized by Machine Learning?
- Damage Caused by Classification Accuracy and Other Discontinuous Improper Accuracy Scoring Rules