Statistical Thinking

Classification vs. Prediction

Code

PREDICTION DECISION-MAKING MACHINE-LEARNING ACCURACY-SCORE CLASSIFICATION DATA-SCIENCE 2017

Classification involves a forced-choice premature decision, and is often misused in machine learning applications. Probability modeling involves the quantification of *tendencies* and usually addresses the real project goals.

AUTHOR
Frank Harrell

AFFILIATION
Vanderbilt University
School of Medicine
Department of Biostatistics

PUBLISHED
January 15, 2017

> *It is important to distinguish prediction and classification. In many decisionmaking contexts, classification represents a premature decision, because classification combines prediction and decision making and usurps the decision maker in specifying costs of wrong decisions. The classification rule must be reformulated if costs/utilities or sampling criteria change. Predictions are separate from decisions and can be used by any decision maker.*
>
> *Classification is best used with non-stochastic/deterministic outcomes that occur in say 0.3 - 0.7 of the observations, and not when the simplest classifer (always outputting "positive" or always outputting "negative") is highly accurate or when two individuals with identical inputs can easily have different outcomes. For these situations, modeling tendencies (i.e., probabilities) is key.*
>
> *Classification should be used when outcomes are distinct and predictors are strong enough to provide, for all subjects, a probability near 1.0 for one of the outcomes.*

The field of machine learning arose somewhat independently of the field of statistics. As a result, machine learning experts tend not to emphasize probabilistic thinking. Probabilistic thinking and understanding uncertainty and variation are hallmarks of statistics. By the way, one of the best books about probabilistic thinking is Nate Silver's *The Signal and The Noise: Why So Many Predictions Fail But Some Don't*. In the medical field, a classic paper is David Spiegelhalter's Probabilistic Prediction in Patient Management and Clinical Trials.

Comments

By not thinking probabilistically, machine learning advocates frequently utilize classifiers instead of using risk prediction models. The situation has gotten acute: many machine learning experts actually label logistic regression as a classification method (it is not). It is important to think about what classification really implies. Classification is in effect a decision. Optimum decisions require making full use of available data, developing predictions, and applying a loss/utility/cost function to make a decision that, for example, minimizes expected loss or maximizes expected utility. Different end users have different utility functions. In risk assessment this leads to their having different risk thresholds for action. Classification assumes that every user has the same utility function and that the utility function implied by the classification system is *that* utility function.

Classification is a *forced choice*. In marketing where the advertising budget is fixed, analysts generally know better than to try to classify a potential customer as someone to ignore or someone to spend resources on. Instead, they model probabilities and create a *lift curve*, whereby potential customers are sorted in decreasing order of estimated probability of purchasing a product. 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.

A frequent argument from data users, e.g., physicians, is that ultimately they need to make a binary decision, so binary classification is needed. This is simply not true. First of all, it is often the case that

the best decision is "no decision; get more data" when the probability of disease is in the middle. In many other cases, the decision is revocable, e.g., the physician starts the patient on a drug at a lower dose and decides later whether to change the dose or the medication. In surgical therapy the decision to operate is irrevocable, but the choice of *when* to operate is up to the surgeon and the patient and depends on severity of disease and symptoms. At any rate, if binary classification is needed, it must be done at the point of care when all utilities are known, not in a data analysis.

When are forced choices appropriate? I think that one needs to consider whether the problem is mechanistic or stochastic/probabilistic. Machine learning advocates often want to apply methods made for the former to problems where biologic variation, sampling variability, and measurement errors exist. It may be best to apply classification techniques instead just to high signal:noise ratio situations such as those in which there there is a known gold standard and one can replicate the experiment and get almost the same result each time. An example is pattern recognition - visual, sound, chemical composition, etc. If one creates an optical character recognition algorithm, the algorithm can be trained by exposing it to any number of replicates of attempts to classify an image as the letters A, B, ... The user of such a classifier may not have time to consider whether any of the classifications were "close calls." And the signal:noise ratio is extremely high. In addition, there is a single "right" answer for each character. This situation is primarily mechanistic or non-stochastic. Contrast that with forecasting death or disease where two patients with identical known characteristics can easily have different outcomes.

When close calls are possible, or when there is inherent randomness to the outcomes, probability estimates are called for. One beauty of probabilities is that they are their own error measures. If the probability of disease is 0.1 and the current decision is not to treat the patient, the probability of this being an error is by definition 0.1. A probability of 0.4 may lead the physician to run another lab test or do a biopsy. When the signal:noise ratio is small, classification is usually not a good goal; there one must model *tendencies*, i.e., probabilities.

The U.S. Weather Service has always phrased rain forecasts as probabilities. I do not want a classification of "it will rain today." There is a slight loss/disutility of carrying an umbrella, and I want to be the one to make the tradeoff.

Whether engaging in credit risk scoring, weather forecasting, climate forecasting, marketing, diagnosis a patient's disease, or estimating a patient's prognosis, I do not want to use a classification method. I want risk estimates with credible intervals or confidence intervals. My opinion is that machine learning classifiers are best used in mechanistic high signal:noise ratio situations, and that probability models should be used in most other situations.

This is related to a subtle point that has been lost on many analysts. Complex machine learning algorithms, which allow for complexities such as high-order interactions, require an enormous amount of data unless the signal:noise ratio is high, another reason for reserving some machine learning techniques for such situations. Regression models which capitalize on additivity assumptions (when they are true, and this is approximately true is much of the time) can yield accurate probability models without having massive datasets. And when the outcome variable being predicted has more than two levels, a single regression model fit can be used to obtain all kinds of interesting quantities, e.g., predicted mean, quantiles, exceedance probabilities, and instantaneous hazard rates.

A special problem with classifiers illustrates an important issue. Users of machine classifiers know that a highly imbalanced sample with regard to a binary outcome variable Y results in a strange classifier. For example, if the sample has 1000 diseased patients and 1,000,000 non-diseased patients, the best classifier may classify everyone as non-diseased; you will be correct 0.999 of the time. For this reason the odd practice of subsampling the controls is used in an attempt to balance the frequencies and get some variation that will lead to sensible looking classifiers (users of regression models would never

exclude good data to get an answer). Then they have to, in some ill-defined way, construct the classifier to make up for biasing the sample. It is simply the case that a classifier trained to a 1/2 prevalence situation will not be applicable to a population with a 1/1000 prevalence. The classifier would have to be re-trained on the new sample, and the patterns detected may change greatly. Logistic regression on the other hand elegantly handles this situation by either (1) having as predictors the variables that made the prevalence so low, or (2) recalibrating the intercept (only) for another dataset with much higher prevalence. Classifiers' extreme dependence on prevalence may be enough to make some researchers always use probability estimators like logistic regression instead. One could go so far as to say that classifiers should not be used at all when there is little variation in the outcome variable (prevalence is near 0 or 1), and that only tendencies (probabilities) should be modeled.

One of the key elements in choosing a method is having a sensitive accuracy scoring rule with the correct statistical properties. Experts in machine classification seldom have the background to understand this enormously important issue, and choosing an improper accuracy score such as proportion classified correctly will result in a bogus model. This is discussed in detail here.

Resources

- The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression by van den Goorbergh, van Smeden, Timmerman, Van Calster

Acknowledgement

Thanks to Andrew Howe for suggesting some wording clarifications.

Discussion Archive (2017–2021)