2-Class Classification, Voting Functions, and Losses

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Voting functions and corresponding classifiers

Empirical search for a good 2-class classifier is essentially search for a good approximation to the likelihood ratio function $\mathcal{L}(\mathbf{x})$. This suggests another kind of consideration for 2-class problems, namely focusing on the building of a good "voting function" $g(\cdot)$ to underlie a classifier.

It's now convenient to employ the -1-1 coding of class labels (use $\mathcal{G} = \{-1, 1\}$) and to without much loss of generality consider classifiers defined for an arbitrary voting function $g(\mathbf{x})$ by

$$f(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$$

(except for the possibility that $g(\mathbf{x}) = 0$, that typically has 0 probability for both classes). Then an optimal voting function for 0-1 loss is

$$g^{\text{opt}}(\mathbf{x}) = \frac{p(\mathbf{x}|1)}{p(\mathbf{x}|-1)} - \frac{P[y=-1]}{P[y=1]}$$

Representation of 0-1 loss error rate

With this notation, a classifier $f(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$ produces loss neatly written as

$$L(y, \hat{y}) = I[yg(\mathbf{x}) < 0]$$

(a loss of 1 is incurred when y and $g(\mathbf{x})$ have opposite signs). So the 0-1 loss expected loss/error rate has the useful representation

$$\mathrm{E}I[yg(\mathbf{x}) < 0]$$

We have seen that a function g optimizing the above is $g^{\text{opt}}(\mathbf{x})$ defined in on the last slide. But the indicator function $I[u < 0]$ involved in the error rate is discontinuous (and thus non-differentiable). For some purposes it would be more convenient to work with a continuous (even differentiable) one in making an empirical choice of voting function.

Bounds on 0-1 loss error rate

If $h(u) \geq I[u < 0]$, it is obvious that

$$\mathrm{E}I[yg(\mathbf{x}) < 0] \leq \mathrm{E}h(yg(\mathbf{x}))$$

So $\mathrm{E}h(yg(\mathbf{x}))$ functions as an upper bound for the 0-1 loss error rate and an approximate (data-based) minimizer of it used as a voting function can be expected to control 0-1 loss error rate. Several different continuous choices of "loss" $h(u)$ can be viewed as motivating popular methods of (voting function and) classifier development. These include:

1. $h(u) = \ln(1 + \exp(-u)) / \ln(2)$ associated with use of logistic regression-based estimated conditional class probabilities to make voting functions,
2. $h(u) = \exp(-u)$ associated with the "AdaBoost" algorithm, and
3. $h(u) = (1 - u)_+$ associated with "support vector machines."

Example h functions

For sake of concreteness, below is a plot of $I[u < 0]$ and the three functions $h(u)$ dominating it discussed on the previous slide.

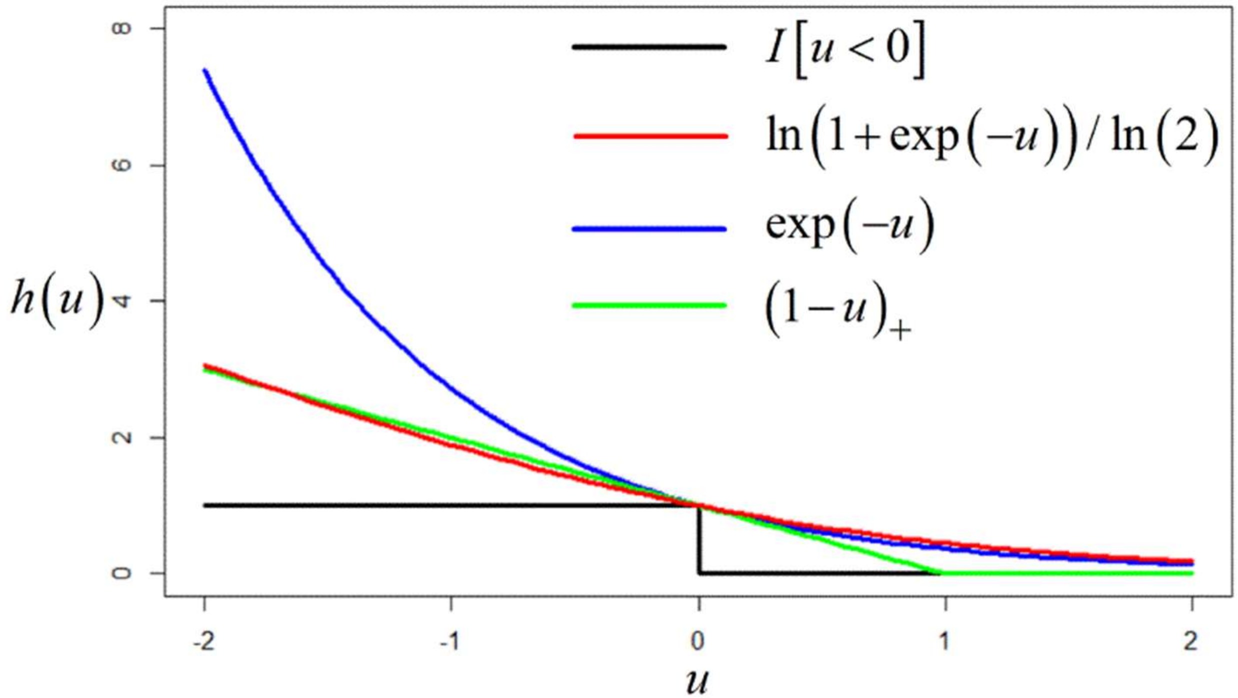

Figure: "Losses" $I[u < 0]$ in black, $h_1(u)$ in red, $h_2(u)$ in blue, and $h_3(u)$ in green.

Optimizers for standard choices of h

Not only does $\mathrm{E}h(yg(\mathbf{x}))$ bound the error rate, but minimizers of $\mathrm{E}h(yg(\mathbf{x}))$ over choice of function $g(\mathbf{x})$ for standard choices of h with $I[u<0] \leq h(u)$ prove to be directly related to the likelihood ratio. Case 1. on the previous slide has optimizing function

$$g^*(\mathbf{x}) = \ln\left(\frac{P[y=1|\mathbf{x}]}{P[y=-1|\mathbf{x}]}\right)$$

and case 2. has an optimizer that is 1/2 of this. Both are monotone transformations of the likelihood ratio and when used as a voting function produce a (0-1 loss) optimal classifier. In case 3. an optimizing function is

$$g^{**}(\mathbf{x}) = \mathrm{sign}\left(P[y=1|\mathbf{x}] - P[y=-1|\mathbf{x}]\right)$$

the optimal classifier itself. **So empirical search for optimizers of (an empirical version of)** $\mathrm{E}h(yg(\mathbf{x}))$ **can produce good classifiers**.