Other 2-Class Prediction Problems

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Log loss probability prediction

There are other (besides 0-1 loss) prediction problems sometimes considered in the 2-class classification model. (These often show up as alternatives to use of classification error rate as test criteria in prediction contests, particularly ones where there is extreme class imbalance.)

One is class probability prediction using the $K = 2$ version of cross-entropy loss (and negative Bernoulli log-likelihood). Where y is in $\{0, 1\}$ but $\hat{y}$ is allowed to be any real number in $[0, 1]$, the so-called "log loss"

$$L(\hat{y}, y) = -y \ln \hat{y} - (1 - y) \ln (1 - \hat{y})$$

is sometimes employed. For this loss, a theoretically optimal predictor is

$$f(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}] = P[y = 1|\mathbf{x}]$$

AUC criterion

Another problem uses as loss (1 minus) an "Area Under the Curve" (AUC). One chooses a function $\mathcal{O}(\mathbf{x})$ taking values in $[-\infty, \infty]$ to *order* values of $\mathbf{x}$ (large $\mathcal{O}(\mathbf{x})$ indicating large likelihood that $y = 1$). For independent $\mathbf{x}$ and $\mathbf{x}^*$ with the (P) distributions of $\mathbf{x}|y = 0$ and $\mathbf{x}|y = 1$ the theoretical "AUC" for $\mathcal{O}$ (to be maximized) is

$$P[\mathcal{O}(\mathbf{x}) < \mathcal{O}(\mathbf{x}^*)]$$

Arguments (using "ROC curves" and Neyman-Pearson theory) show the likelihood ratio $\mathcal{L}(\mathbf{x})$ *or* any monotone increasing transform of it (including $P[y = 1|\mathbf{x}]$) to be an optimal $\mathcal{O}(\mathbf{x})$.

An empirical AUC based on a test set (often used to score predictive analytics contests) is

$$\frac{1}{\#\, y_i = 0} \sum_{i \text{ s.t.} y_i = 0} \left(\frac{1}{\#\, y_i = 1} \sum_{j \text{ s.t.} y_j = 1} I[\mathcal{O}(\mathbf{x}_i) < \mathcal{O}(\mathbf{x}_j)] \right)$$

Losses built on $y\hat{y}$

For reasons that will shortly become clear, it is sometimes convenient to use not 0-1 coding but rather -1-1 coding in 2-class classification models, so that y is in $\{-1, 1\}$. Suppose that $\hat{y}$ is allowed to be any real number, then three other (initially odd-looking) losses built on the product $-y\hat{y}$ are sometimes considered, namely

$$L_1(\hat{y}, y) = \ln(1 + \exp(-y\hat{y})) / \ln(2) \quad ,$$
$$L_2(\hat{y}, y) = \exp(-y\hat{y}) \quad , \text{ and}$$
$$L_3(\hat{y}, y) = (1 - y\hat{y})_+$$

Corresponding optimal predictors

For these losses, theoretically optimal predictors are respectively

$$f_1(\mathbf{x}) = \ln\left(\frac{P[y=1|\mathbf{x}]}{P[y=-1|\mathbf{x}]}\right) = \ln \mathcal{L}(\mathbf{x}) \quad ,$$

$$f_2(\mathbf{x}) = \frac{1}{2}\ln\left(\frac{P[y=1|\mathbf{x}]}{P[y=-1|\mathbf{x}]}\right) = \frac{1}{2}\ln \mathcal{L}(\mathbf{x}) \quad , \text{and}$$

$$f_3(\mathbf{x}) = \operatorname{sign}\left(P[y=1|\mathbf{x}] - P[y=-1|\mathbf{x}]\right)$$