More on Optimal 2-Class Classifiers

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Notations for 2-class models

We have identified a theoretically optimal (0-1 loss) K-class classifier as

$$f(\mathbf{x}) = \arg\max_{a} P[y = a|\mathbf{x}]$$

By far, the most important version of this is the $K = 2$ case. And for this case, there are some very important additional general insights to be had.

For $K = 2$, for various purposes different ones of the (arbitrary and completely equivalent) codings for the possible values of y

$$\{0, 1\}, \{1, 2\}, \text{ and } \{-1, 1\}$$

prove useful. For the time being, employ the first and abbreviate $P[y = 1]$ as π (so that $P[y = 0] = 1 - \pi$), and write $p(\mathbf{x}|1)$ and $p(\mathbf{x}|0)$ for the two class-conditional densities for $\mathbf{x}$.

Optimal 0-1 loss classification

Since

$$P[y=1|\mathbf{x}] = \frac{\pi p(\mathbf{x}|1)}{\pi p(\mathbf{x}|1) + (1-\pi) p(\mathbf{x}|0)} \quad \text{and}$$
$$P[y=0|\mathbf{x}] = \frac{(1-\pi) p(\mathbf{x}|0)}{\pi p(\mathbf{x}|1) + (1-\pi) p(\mathbf{x}|0)}$$

an optimal classifier is

$$\begin{aligned} f(\mathbf{x}) &= I[P[y=1|\mathbf{x}] > .5] \\ &= I[P[y=1|\mathbf{x}] > P[y=0|\mathbf{x}]] \\ &= I\left[\mathcal{L}(\mathbf{x}) > \frac{(1-\pi)}{\pi}\right] \end{aligned}$$

and $f(\mathbf{x}) = 1$ when $P[y=1|\mathbf{x}]$ is large, or equivalently the likelihood ratio $\mathcal{L}(\mathbf{x})$ is large.

N-P theory, asymmetric loss classification

Notice that this makes connection to classical statistical theory and identifies the optimal classifier as a Neyman-Pearson test of the simple hypotheses $H_0 : y = 0$ versus $H_a : y = 1$ with "cut-point" the ratio $(1 - \pi) / \pi$.

As a slight generalization of this development, note that for constants $L_0 \geq 0$ and $L_1 \geq 0$ and an asymmetric loss

$$L(\widehat{y}, y) = L_y I[\widehat{y} \neq y]$$

an optimal classifier is

$$f(\mathbf{x}) = I\left[\mathcal{L}(\mathbf{x}) > \frac{(1 - \pi) L_0}{\pi L_1}\right]$$

Shifting class probabilities

An important issue in classification models is the effect of changes in π on both $P[y = 1|\mathbf{x}]$ and (optimal classifier) $f(\mathbf{x})$. There are situations, for example, in which π is very extreme (one class is rare and the terminology "extreme class imbalance" is commonly used) and it is then common practice to build a predictor using a training set made with relative frequency of $y = 1$ that is π^*, a value that is much more moderate (nearer to .5) than π. The obvious question is how to translate results for the synthetic value π^* to results for the real value π.

Effects on input-conditional class probabilities

Since

$$P[y=1|\mathbf{x}] = \frac{\mathcal{L}(\mathbf{x})}{\mathcal{L}(\mathbf{x}) + \frac{(1-\pi)}{\pi}}$$

it follows that

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

So subscripting P with π or π^* depending upon which marginal probability of $y=1$ is operating (in models with the same class-conditional densities $p(\mathbf{x}|1)$ and $p(\mathbf{x}|0)$),

$$P_\pi[y=1|\mathbf{x}] = \frac{\frac{(1-\pi^*)}{\pi^*}\left(\frac{P_{\pi^*}[y=1|\mathbf{x}]}{1-P_{\pi^*}[y=1|\mathbf{x}]}\right)}{\frac{(1-\pi^*)}{\pi^*}\left(\frac{P_{\pi^*}[y=1|\mathbf{x}]}{1-P_{\pi^*}[y=1|\mathbf{x}]}\right) + \frac{(1-\pi)}{\pi}}$$

Effect on optimal 0-1 loss classifiers

From above it is obvious how to translate an estimate of $P_{\pi^*}[y = 1|\mathbf{x}]$ made from a synthetically balanced training set to one for the real situation described by π. Further, an optimal classifier is

$$I\left[\left(\frac{P_{\pi^*}[y=1|\mathbf{x}]}{1-P_{\pi^*}[y=1|\mathbf{x}]}\right) > \frac{\pi^*(1-\pi)}{(1-\pi^*)\,\pi}\right]$$

and it is obvious how to translate an estimate of $P_{\pi^*}[y = 1|\mathbf{x}]$ made from a synthetically balanced training set to an approximately optimal classification for the real situation described by π.

For example, considering the k-nearest neighbor set-up using a training set made with relative frequency of $y = 1$ that is π^* when the real probability that $y = 1$ is π, the right use of a neighborhood of $\mathbf{x}$ containing $n_1(\mathbf{x})$ cases $\mathbf{x}_i$ with $y = 1$ and $n_0(\mathbf{x}) = k - n_1(\mathbf{x})$ with $y = 0$, is to classify according to

$$I\left[n_1(\mathbf{x})(1-\pi^*)\,\pi > n_0(\mathbf{x})\,\pi^*(1-\pi)\right]$$