Explanation of Loss Minimizing Probabilistic Classifier

Matthew Dailey

July 12, 2014

By the minimum risk principle, we want to minimize, for all classification strategies s, the expected loss

$$R(s) = P(1 \to 2 \mid \text{using } s)L(1 \to 2) + P(2 \to 1 \mid \text{using } s)L(2 \to 1).$$

If we are using a probabilistic classifier, we define a feature vector $\boldsymbol{x}$ and build estimators for $P(1 \mid \boldsymbol{x})$ and $P(2 \mid \boldsymbol{x})$, and our strategy is

$$s(\boldsymbol{x}; \Theta) \equiv y(\boldsymbol{x}) = \begin{cases} 1 & \text{if } P(1 \mid \boldsymbol{x})/P(2 \mid \boldsymbol{x}) > \Theta, \\ 2 & \text{otherwise.} \end{cases}$$

In this case our risk function becomes

$$R(s) = P(y(\boldsymbol{x}) = 2 \text{ when } \boldsymbol{x} \text{ is } 1)L(1 \to 2) + P(y(\boldsymbol{x}) = 1 \text{ when } \boldsymbol{x} \text{ is } 2)L(2 \to 1).$$

For a particular $\boldsymbol{x}$, we will always answer with the same class based on our estimates of $P(1 \mid \boldsymbol{x})$ and $P(2 \mid \boldsymbol{x})$. If we choose class 1, we incur a loss of

$$P(2 \mid \boldsymbol{x})L(2 \to 1),$$

and if we choose class 2, we incur a loss of

$$P(1 \mid \boldsymbol{x})L(1 \to 2).$$

We should make the decision that minimizes the loss, i.e., choose 1 if

$$P(2 \mid \boldsymbol{x})L(2 \to 1) < P(1 \mid \boldsymbol{x})L(1 \to 2),$$

or, correspondingly, if

$$P(1 \mid \boldsymbol{x})/P(2 \mid \boldsymbol{x}) > L(2 \to 1)/L(1 \to 2).$$

This means that our threshold should be set as

$$\Theta = L(2 \to 1)/L(1 \to 2).$$

Note that s is optimal if our model $P(1 \mid \boldsymbol{x})$ is exact, but in practice, the best we can obtain is an approximation based on observation of historical data.