Theoretically Optimal Predictors

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Modeling, loss, and expected loss

- Development of "what would be the best predictor if I knew the case-generating reality" gives a target to shoot for and guide predictor-making

- Modeling

$$(\mathbf{x}, y) \sim P \text{ and } \mathrm{E}(\cdot) \text{ is the corresponding expectation operation}$$

- Loss function for $\hat{y}$ predicted and y observed

$$L(\hat{y}, y) \geq 0$$

- Expected loss ("prediction error") of (theoretical) predictor $f(\mathbf{x})$

$$\mathrm{E}L(f(\mathbf{x}), y) = \mathrm{E}\mathrm{E}\left[L(f(\mathbf{x}), y) \mid \mathbf{x}\right]$$

Theoretically optimal prediction

- The iterated form of the expectation shows what predictor has minimum predictor error … as a function of $\mathbf{x}$ the prediction should minimize conditional expected loss given $\mathbf{x}$

- Optimal theoretical (not-training-set-dependent) predictor is

$$f(x) = \arg\min_a \mathrm{E}\left[L(a, y) \,|x\right]$$

- Its prediction error is as small as possible (setting a limit on what can be achieved) and its form is what one hopes to approximate with a real predictor $\hat{f}$ built using a training set

Squared error loss (SEL)

- Where a target y is quantitative (of "interval type") a common loss is squared error

$$L(\hat{y}, y) = (\hat{y} - y)^2$$

- Here the theoretically optimal predictor is the conditional mean function

$$f(\mathbf{x}) = \mathrm{E}[y \mid \mathbf{x}]$$

(unavailable for use predicting a new output, as P is not known)

- In this context "statistical machine learning" is "regression" and approximation of the conditional mean response given the input based on a training set

K-class 0-1 loss classification

- Where $y \in \{0,1,\ldots,K-1\}$ or $y \in \{1,\ldots,K\}$ a natural loss is

$$L(\hat{y}, y) = I(\hat{y} \neq y)$$

(and the expected loss/prediction error is the overall mis-classification rate)

- Here, with $p(\mathbf{x} \mid y)$ the class-conditional density for $\mathbf{x} \mid y$

$$\begin{aligned} f(x) &= \arg\min_a \sum_{v \neq a} P[y = v|x] \\ &= \arg\max_a P[y = a|x] \\ &= \arg\max_a P[y = a]\, p(x|a) \end{aligned}$$

- One wishes to predict the class with the maximum conditional probability given the input vector

K-class classification with asymmetric loss

- A natural generalization of 0-1 loss in K-class classification is for (possibly different) values $l_y \geq 0$ is

$$L(\hat{y}, y) = l_y I(\hat{y} \neq y)$$

- Here

$$\begin{aligned} f(x) &= \arg\min_a \sum_{v \neq a} l_v P[y = v|x] \\ &= \arg\max_a l_a P[y = a|x] \\ &= \arg\max_a l_a P[y = a] \, p(x|a) \end{aligned}$$

- One wishes to predict the class with the maximum weighted conditional probability given the input vector

Predicting K class probabilities

- In a K-class classification model, one might wish to produce a K-vector $\hat{\mathbf{y}}$

(ultimately representing assessments of how likely it is that $y = k$)

- The "cross-entropy" loss for this problem is

$$L(\hat{\mathbf{y}}, y) = -\sum_k I[y = k] \ln \hat{y}_k$$

(the single observation/sample size 1 multinomial negative log-likelihood)

- Here (use of a Lagrange multiplier argument shows that) an optimal (vector) predictor $\mathbf{f}$ has

$$f_k(\mathbf{x}) = \mathrm{E}\left[I[y = k] \mid \mathbf{x}\right] = P[y = k \mid \mathbf{x}]$$