Loss Functions in Practice

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

January 13, 2020
Introduction

- seek a model that predicts $Y$ from $X$, denote $f(X)$
- how do we find $f(X)$?
- need mechanism to quantify “goodness” of predictions
- a ‘loss function’ $L(Y, f(X))$
Minimizing loss

- $f(X)$ can be selected by minimizing the average loss, or ‘expected loss’ or ‘expected prediction error’:
  \[ EPE(f) = E[L(Y, f(X))] \]
- the average, or ‘expectation’ is taken over the joint distribution of $X$ and $Y$: $Pr(X, Y)$
The EPE has two purposes:

1. To identify what $f(X)$ should look like

2. To evaluate the predictive quality of a fitted model: The EPE can be approximated using testing data:

$$EPE(f) = E[L(Y, f(X))] \approx \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} L(y_i, f(x_i))$$
Squared-error loss

- loss:

\[ L(Y, f(X)) = (Y - f(X))^2 \]

- \( f(x) \) that minimizes expected loss is mean of \( Y \) given \( X \)

\[ \hat{Y} = f(X) = \mu_{Y|X} = \hat{E}_{Y|X}[Y] \]
Squared-error loss

- For $L_2$ loss, $\hat{Y} = f(X) = \text{mean of } Y \text{ given } X$.
- This is based on decision theory; don’t need data.
- In practice, we need a model (e.g., linear or $k$-NN) of the association between $Y$ and $X$. We need to “fit” the model to training data, and do that by minimizing the training error:

$$\overline{\text{err}}(f) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} L(y_i, f(x_i))$$
Linear model with squared error loss

- given input vector $X$, generate prediction about $Y$ as follows:

$$\hat{Y} = f(X) = \hat{\beta}_1 + \hat{\beta}_2 X$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the values that minimize the training error (average sum of squares):

$$\overline{\text{err}}(\beta_1, \beta_2) = \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} L(y_i, f(x_i))$$

$$= \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} (y_i - \beta_1 + \beta_2 x_i)^2$$

- where there are $n$ training examples $(y_i, x_i)$
- least-squares problem; we can compute $\hat{\beta}_1$ and $\hat{\beta}_2$ easily
- However, it’s not always that easy, for other models or loss functions, we may need to find $\beta$ using numerical optimization methods.