Statistical Decision Theory

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Statistical Decision Theory: Quantitative Response

- Let Y be a quantitative output and $\boldsymbol{X}$ be the input where $Y \in \Re$ and $\boldsymbol{X} \in \Re^p$.
- Let $P(\boldsymbol{X}, Y)$ be the joint distribution.
- Goal is to find a function, $\hat{Y} = f(\boldsymbol{X})$ for predicting Y.
- Criterion or loss function, $L(Y, f(\boldsymbol{X}))$ for scoring the performance of the learned function or model.
- Bayes risk

$$L^* = \inf_{f(\boldsymbol{X}) \in \mathfrak{D}} \left(E[L(Y, f(\boldsymbol{X}))]\right)$$

- $f^*(\boldsymbol{X})$ is Bayes optimal if $E[L(Y, f^*(\boldsymbol{X}))] = L^*$

Squared Error: Bayes Optimal Predictor

- Square error loss: $[Y - f(\boldsymbol{X})]^2$
- Expected square prediction error or

$$\begin{aligned}
\mathrm{EPE}(f) &= E[Y - f(\boldsymbol{X})]^2 \\
&= \int [Y - f(\boldsymbol{X})]^2 p(x, y) dxdy \\
&= E_{\boldsymbol{X}} \left[E_{Y|\boldsymbol{X}}([Y - f(\boldsymbol{X})]^2 | \boldsymbol{X}) \right]
\end{aligned}$$

- The Bayes optimal predictor is given by the conditional mean, $f(\boldsymbol{X}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$
- The universal machine learning solution

Qualitative Response

- For a categorical response, $G \in \mathcal{G}$, we want a decision function or rule, $d(\boldsymbol{X}) \in \mathcal{G}$, where $\mathcal{G} = \{g_1, \ldots, g_K\}$
- Need a loss function $L(g_i, d(\boldsymbol{X}))$ for classifying an observation as $d(\boldsymbol{X})$ when $G = g_i$.
- A common choice is a zero-one loss function:

$$c_{i,j} = \begin{cases} 1 & g_i \neq g_j \\ 0 & g_i = g_j \end{cases}$$

Optimal Qualitative Classification

- The expected prediction error is

$$\begin{aligned}\text{EPE}(d) &= E_{G,\boldsymbol{X}}[L(G, d(\boldsymbol{X}))] \\ &= E_{\boldsymbol{X}}\left(\sum_{k=1}^{K} L(\mathcal{G}_k, d(\boldsymbol{X}))Pr(\mathcal{G}_k|\boldsymbol{X})\right)\end{aligned}$$

- The minimum of the EPE is the Bayes risk
- For $\boldsymbol{X} = \boldsymbol{x}$, the Bayes optimal decision rule is

$$d(\boldsymbol{x}) = argmin_{c\in\mathcal{G}} \sum_{i=1}^{K} L(g_i, c)Pr(G = g_i|\boldsymbol{X} = \boldsymbol{x})$$

Optimal 0-1 Loss Classification

When the loss function is 0-1, the Bayes optimal rule is

$$d(\boldsymbol{X}) = \text{argmax}_{g_j \in \mathcal{G}} Pr(G = g_j | \boldsymbol{X} = \boldsymbol{x})$$

Break ties arbitrarily.

Optimal Bayes Classifier

- We have methods for constructing classification functions $d(\boldsymbol{X})$ that are Bayes optimal under known distributions.
- We also estimate likelihoods $h(\boldsymbol{X} = \boldsymbol{x} | G = g_i)$ and apply Bayes rule to get $Pr(G = g_i | \boldsymbol{X} = \boldsymbol{x})$.