STAT 3690 Lecture 32

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 18, 2022

Misclassification/error rate

- Population: $\Pr(Y \neq h(\mathbf{X}))$
  - $h(\cdot)$: the classifier to be evaluated
- Apparent estimation:
  - Implementation
    1. Have a random split of dataset
    2. Apply the fitted classifier to each observation in the training data
    3. Estimate the error rate by the misclassification proportion
  - Training and testing with identical data points
  - Severe underestimation likely
- Parametric estimation
  - Implementation
    1. Express $\Pr(Y \neq h(\mathbf{X}))$ in terms of unknown parameters
    2. Plug in estimates of unknown parameters
  - Underestimation likely
- Estimation via M-fold cross validation (CV)
  - Implementation
    1. The dataset is randomly partitioned into M chunks.
    2. Train one classifier upon each combination of $M-1$ chunks.
    3. Apply each classifier to the corresponding remaining chunk and compute the empirical error rate.
    4. Estimate the population error rate by averaging these M empirical error rates.
  - Leave-one-out CV $\Leftrightarrow$ n-fold CV
- Estimation via $M \times L$-fold CV
  - Implementation
    1. Repeat the four steps of M-fold CV L times.
    2. Average all the ML resulting empirical error rates.
  - $M \times 1$-fold CV $\Leftrightarrow$ M-fold CV