Data Splitting for Binary Classification

1 Keywords

- Area Under Curve

$$\text{AUC} \approx \sum_k \frac{\left(\text{TPR}_{\lambda_k} + \text{TPR}_{\lambda_{k+1}}\right)}{2} \left(\text{FPR}_{\lambda_{k+1}} - \text{FPR}_{\lambda_k}\right) \tag{1}$$

- Binary Classification Problem

$$C = 2 \tag{2}$$

- Bin Width

$$\text{BW} = \frac{\max_n X_{n,f} - \min_n X_{n,f}}{K} \tag{3}$$

where K is the number of bins.

- False Negative

$$\text{FN} = \sum_n \mathbf{1}(\widehat{x}_n = 0)\mathbf{1}(y_n = 1) \tag{4}$$

- False Positive

$$\text{FP} = \sum_n \mathbf{1}(\widehat{x}_n = 1)\mathbf{1}(y_n = 0) \tag{5}$$

- False Positive Rate (1-specificity)

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{Precision} = \frac{\sum_n \mathbf{1}(\widehat{y}_n = 1)\mathbf{1}(y_n = 0)}{\sum_n \mathbf{1}(y_n = 0)} \tag{6}$$

- Normalized Histogram $(\tilde{x}_k, h_k)$

$$\begin{cases} \forall k \in \{0, \ldots, K\} & x_k = \min_n X_{n,f} + k\text{BW} \\ \forall k \in \{1, \ldots, K\} & \tilde{x}_k = \frac{x_{k-1}+x_k}{2} \\ \forall k \in \{1, \ldots, K\} & h_k = \frac{1}{N}\sum_n \mathbf{1}(x_{k-1} \leq X_{n,f} < x_k) \end{cases} \tag{7}$$

- Normalized Conditional Histogram $(\tilde{x}_k, h_k^c)$

$$\begin{cases} \forall k \in \{0, \ldots, K\} & x_k = \min_n X_{n,f} + k\text{BW} \\ \forall k \in \{1, \ldots, K\} & \tilde{x}_k = \frac{x_{k-1}+x_k}{2} \\ \forall k \in \{1, \ldots, K\} & h_k^c = \frac{\sum_n \mathbf{1}(x_{k-1} \leq X_{n,f} < x_k)\mathbf{1}(Y_n = c)}{\sum_n \mathbf{1}(Y_n = c)} \end{cases} \tag{8}$$

where $c \in \{0, 1\}$.

- Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\sum_n \mathbf{1}(\widehat{y}_n = 1)\mathbf{1}(y_n = 1)}{\sum_n \mathbf{1}(\widehat{x}_n = 1)} \tag{9}$$

- Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\sum_n \mathbf{1}(\widehat{y}_n = 1)\mathbf{1}(y_n = 1)}{\sum_n \mathbf{1}(y_n = 1)} \tag{10}$$

- ROC curve:

$$\left(\text{FPR}_\lambda, \text{TPR}_\lambda\right)_\lambda \tag{11}$$

- True Negative

$$\text{TN} = \sum_n \mathbf{1}(\widehat{y}_n = 0)\mathbf{1}(y_n = 0) \tag{12}$$

- True Positive Rate (sensitivity)

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall} = \frac{\sum_n \mathbf{1}(\widehat{y}_n = 1)\mathbf{1}(y_n = 1)}{\sum_n \mathbf{1}(\mathbf{1}(y_n = 1)} \tag{13}$$

- True Positive

$$\text{TP} = \sum_n \mathbf{1}(\widehat{y}_n = 1)\mathbf{1}(y_n = 1) \tag{14}$$