ch05_notes

August 10, 2019

Table of Contents

---

1 Resampling Methods

- ***Resampling methods*** involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model
- Two of the most commonly used resampling methods are ***cross-validation*** and the bootstrap
- Resampling methods can be useful in ***model assessment***, the process of evaluating a model's performance, or in ***model selection***, the process of selecting the proper level of flexibility.

1.1 Cross-validation

1.1.1 The Validation Set Approach

- Randomly divide the data into a ***training set*** and ***validation set***. The model is fit on the training set and its prediction performance on the test set provides an estimate of overall performance.
- In the case of a quantitative response, the prediction performance is measured by the mean-squared-error. The validation estimates the "true" MSE with the mean-squared error $\mathrm{MSE}_{validation}$ computed on the validation set.

Advantages

- conceptual simplicity
- ease of implementation
- low computational resources

Disadvantages

- the validation estimate is highly variable - it is highly dependent on the train/validation set split
- since the model is trained on a subset of the dataset, it may tend to overestimate the test error rate if it was trained on the entire dataset

1.1.2 Leave-One-Out Cross Validation

Given paired observations $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, for each $1 \leqslant i \leqslant n$: - Divide the data $\mathcal{D}$ into a training set $\mathcal{D}_{(i)} = \mathcal{D}\ \{(x_i, y_i)\}$ and a validation set $\{(x_i, y_i)\}$. - Train a model $\mathcal{M}_i$ on $\mathcal{D}_{(i)}$ and use it to predict $\hat{y}_i$. - The LOOCV estimate for MSE_{test} is

```
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^n \text{MSE}_i$$
```

where $\text{MSE}_i = (y_i - \hat{y}_i)31$

Advantages

- approximately unbiased
- deterministic - doesn't depend on a random train/test split.
- computationally fast in least squares regression

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

where h_i is the Section **??** of point i

Disdvantages

- Computationally expensive32 in general

1.1.3 k-fold Cross-Validation

Given paired observations $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, divide the data $\mathcal{D}$ into K ***folds*** (sets) $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of roughly equal size.33 Then for each $1 \leqslant k \leqslant K$:

- Train a model on $\mathcal{M}_k$ on $\cup_{j \neq k}\mathcal{D}_j$ and validate on $\mathcal{D}_k$.
- The k-fold CV estimate for MSE_{test} is

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_k$$

where MSE_k is the mean-squared-error on the validation set $\mathcal{D}_k$

Advantages

- computationally faster than $LOOCV$ if $k > 1$
- less variance than validation set approach or LOOCV

Disdvantages

- more biased than LOOCV if $k > 1$.

1.1.4 Bias-Variance Tradeoff for k-fold Cross Validation

As $k \to n$, bias $\downarrow$ but variance $\uparrow$

1.1.5 Cross-Validation on Classification Problems

In the classification setting, we define the LOOCV estimate

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Err}_i$$

where $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$. The k-fold CV and validation error rates are defined analogously.

1.2 The Bootstrap

The bootstrap is a method for estimating the standard error of a statistic34 or statistical learning process. In the case of an estimator $\hat{S}$ for a statistic S proceeds as follows:

Given a dataset $\mathcal{D}$ with $|\mathcal{D} = n|$, for $1 \leqslant i \leqslant B$: - Create a bootstrap dataset $\mathcal{D}_i^*$ by sampling uniformly n times from $\mathcal{D}$ - Calculate the statistic S on $\mathcal{D}_i^*$ to get a bootstrap estimate S_i^* of S

Then the bootstrap estimate for the $\mathbf{se}(S)$ the sample standard deviation of the boostrap estimates $S_1^*, \ldots, S_B^*$:

$$\hat{se}(\hat{S}) = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(S_i^* - \overline{S^*}\right)^2}$$

1.3 Footnotes

31. MSE_i is just the mean-squared error of the model $\mathcal{M}_i$ on the validation set $\{(x_i, y_i)\}$. It is an approximately unbiased estimator of MSE_{test} but it has high variance. But as the average of the MSE_i, $CV_{(n)}$ has much lower variance.\

$CV_{(n)}$ is sometimes called the LOOCV error rate – it can be seen as the average error rate over the singleton validation sets $\{(x_i, y_i)\}$

32. Specifically $O(n * \text{model fit time})$

33. LOOCV is then k-fold CV in the case $k = n$. Analogous, CV_k is sometimes called the k-fold CV error rate, the average error over the folds.

34. Recall a statistic S is just a function of a sample $S = S(X_1, \ldots, X_n)$