Back to Reality: Predictions With a Single Sample

Contents

5.3 Back to Reality: Predictions With a Single Sample

- Predictive accuracy provides insight into the performance of a predictor function, and can be used to choose between competing ones.
  - The key to this usefulness, however, is that the predictive accuracy can be measured on population $\mathcal{P}$ about which we want to make inference.

- Unfortunately, we typically only have $\mathcal{S}$
  - but not $\mathcal{P}$ nor $\mathcal{T} = \mathcal{P} \setminus \mathcal{S}$.

- So what do we do? All of our $APSE$ calculations have assumed we have $\mathcal{P}$...

- This is the basic problem of inductive inference.
  - Experience says that whenever interest lies in some attribute of the population $a(\mathcal{P})$, we might use $a(\mathcal{S})$ as an estimate of that attribute.

5.3.1 Predictive Accuracy as a Population Attribute

- We cast predictive accuracy as an attribute of population $\mathcal{P}$
  - and then use the corresponding attribute evaluated on $\mathcal{S}$ as its estimate.

- In particular, we care about the attribute

$$a_1(\mathcal{P}) = APSE(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}) = \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \widehat{\mu}_{\mathcal{S}}(\mathbf{x}_u))^2$$

in the **single subset** paradigm

  - this definition relies on the single sample $\mathcal{S}$ and so we will call it the **single subset version** of APSE.

- We also care about the attribute

$$a_2(\mathcal{P}) = APSE(\mathcal{P}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} APSE(\mathcal{P}, \widehat{\mu}_{\mathcal{S}_j})$$

in the **multiple subset** paradigm

  - this definition relies on many (perhaps all possible) samples $\mathcal{S}_1, \ldots, \mathcal{S}_{N_{\mathcal{S}}}$ and so we will call it the **multiple subset version** of APSE.

These are two distinct population attributes, each a slightly different measure of an average prediction squared error.

However, we are usually more concerned with how well each predictor function performs on the population which was **not** used to construct the estimate.

  - Thus the single and multiple subset attributes may be more usefully defined as

$$a_1(\mathcal{P}) = APSE(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}) = \frac{1}{N - n} \sum_{u \in \mathcal{T}} (y_u - \widehat{\mu}_{\mathcal{S}}(\mathbf{x}_u))^2$$

and

$$a_2(\mathcal{P}) = APSE(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} APSE(\mathcal{T}_j, \widehat{\mu}_{\mathcal{S}_j})$$

5.3.1.1 The Single Subset Version

- Suppose we were interested in estimating

$$APSE(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (y_i - \widehat{\mu}_{\mathcal{S}}(\mathbf{x}_i))^2$$

where

  - the predictor function $\widehat{\mu}_{\mathcal{S}}$ is constructed using $\mathcal{S}$
  - the prediction errors are evaluated on $\mathcal{T} = \mathcal{P} \setminus \mathcal{S}$
  - the $|\cdot|$ operator denotes *cardinality*, not an absoluate value

- If all we ever observed was the sample $\mathcal{S}$ from $\mathcal{P}$, we might approximate the single subset version of the APSE by

  - selecting a partition of $\mathcal{S}$ into $\mathcal{S}_0$ and its complement $\mathcal{T}_0$ (i.e., $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{T}_0$, $\mathcal{S}_0 \cap \mathcal{T}_0 = \emptyset$)

- We then use these pieces to *estimate* $\mathcal{P}$, $\mathcal{S}$, and $\mathcal{T}$. In particular:
  - $\widehat{\mathcal{P}} = \mathcal{S} \equiv \mathcal{P}_0$
  - $\widehat{\mathcal{S}} = \mathcal{S}_0$
  - $\widehat{\mathcal{T}} = \mathcal{T}_0$

- The sample estimate of $APSE(\mathcal{T}, \widehat{\mu}_{\mathcal{S}})$ is thus

$$\widehat{APSE}(\mathcal{T}, \widehat{\mu}_{\mathcal{S}}) = APSE(\widehat{\mathcal{T}}, \widehat{\mu}_{\widehat{\mathcal{S}}}) = APSE(\mathcal{T}_0, \widehat{\mu}_{\mathcal{S}_0}) = \frac{1}{|\mathcal{T}_0|} \sum_{u \in \mathcal{T}_0} (y_u - \widehat{\mu}_{\mathcal{S}_0}(\mathbf{x}_u))^2$$

If, alternatively, interest lied in estimating $APSE(\mathcal{P}, \widehat{\mu}_{\mathcal{S}})$ we could do so similarly

$$\widehat{APSE}(\mathcal{P}, \widehat{\mu}_{\mathcal{S}}) = APSE(\widehat{\mathcal{P}}, \widehat{\mu}_{\widehat{\mathcal{S}}}) = APSE(\mathcal{P}_0, \widehat{\mu}_{\mathcal{S}_0}) = \frac{1}{|\mathcal{P}_0|} \sum_{u \in \mathcal{P}_0} (y_u - \widehat{\mu}_{\mathcal{S}_0}(\mathbf{x}_u))^2$$

Training and Test Set

- The set $\mathcal{S}_0$ is sometimes called the **training** set.
  - Because the estimate $\widehat{\mu}_{\mathcal{S}_0}(\mathbf{x})$ is determined only from observations in $\mathcal{S}_0$
  - Because estimation of a prediction function is like **learning** the predictor function from the data (we sometimes say $\mathcal{S}_0$ is used to "train" the predictor function).

- The out of sample set $\mathcal{T}_0$ is often called the **test** set.
  - Because it is used to assess the quality of the "learning".
  - The test set is also more traditionally called a **hold-out sample** to not be used in estimation but to assess the quality of prediction.
  - It has also long been called a **validation** set.

- Performing such a partitioning of your sample $\mathcal{S}$ into a training and a testing set is commonly referred to as **cross validation**

Of course, the million dollar question is how to pick $\mathcal{S}_0$ from $\mathcal{P}_0 = \mathcal{S}$.

5.3.1.2 The Multiple Subset Version

- Suppose we were interested in estimating the average performance over all $N_{\mathcal{S}}$ possible samples

$$APSE(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} APSE(\mathcal{T}_j, \widehat{\mu}_{\mathcal{S}_j})$$

  where
  - $\mathcal{S}_j$ is the j^{th} subset of $\mathcal{P}$, $j = 1, \ldots, N_S$
  - the predictor function $\widehat{\mu}_{\mathcal{S}_j}$ is constructed using $\mathcal{S}_j$ and
  - the prediction errors are evaluated on $\mathcal{T}_j = \mathcal{P} \setminus \mathcal{S}_j$.

- Here we may similarly use the observed sample $\mathcal{S}$ as an estimate of $\mathcal{P}$ (i.e., $\widehat{\mathcal{P}} = \mathcal{P}_0 = \mathcal{S}$)

- Then to mimic taking many samples (and test sets) from $\mathcal{P}$, we do this with $\mathcal{P}_0 = \mathcal{S}$
  - This corresponds to defining many partitions of $\mathcal{S}$: $(\mathcal{S}_{0,j}, \mathcal{T}_{0,j})$, $j = 1, 2, \ldots, N_S$
  - This is precisely what we did in the single subset case, but now we're just repeating it many (N_S) times.

- We then estimate $APSE(\mathcal{T}, \widetilde{\mu})$ by

$$\widehat{APSE}(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_S} \sum_{j=1}^{N_S} APSE(\mathcal{T}_{0,j}, \widehat{\mu}_{\mathcal{S}_{0,j}})$$

- As with a single subset, the question remains as to how to pick the subsets $\mathcal{S}_{0,j}$ from $\mathcal{P}_0$ for $j = 1, \ldots, N_S$.

5.3.2 Choosing the Subsets

5.3.2.1 The Single Subset Version

- It is not always obvious how one should choose $\mathcal{S}_0$ and $\mathcal{T}_0$ in a given situation.

- One guide is that the method of selecting $\mathcal{S}_0$ from $\mathcal{P}_0$ should be as similar as possible to that of selecting the sample $\mathcal{S}$ from the study population $\mathcal{P}$.
  - That is, the same sampling mechanism would be used.

- For example, if $\mathcal{S}$ is a sample chosen at random from $\mathcal{P}$, then so should $\mathcal{S}_0$ be one chosen at random from $\mathcal{P}_0 = \mathcal{S}$.
  - Typically this is what is done.
  - However, in general, there could be different choices.

- A few key questions still need to be addressed when doing this:
  - Should the sampling be done with, or without, replacement?
  - How large should the sample $\mathcal{S}_0$ be?
  - Should $\mathcal{T}_0$ be the full complement of $\mathcal{S}_0$ or just a sample from the complement? And if just a sample from the complement, then how large should $\mathcal{T}_0$ be?

- We address these concerns below.

The Sampling Mechanism

- If predictive accuracy is meant to be an "out-of-sample" assessment, it would seem prudent to restrict ourselves to sampling *without replacement*.
  - There is a clear distinction between the training and test set.
  - Sampling without replacement reduces the possibility of overestimating the predictor's accuracy.

- Sampling with replacement
  - would require redefining APSE to include duplicates in the samples.
  - Unless APSE was calculated using only "out-of-sample" units.

Picking a Training Set Size

- We can gain insight into how large the **training set** should be from the fact that the predicted squared errors are averaged.
  - Recall that
$$\text{SD}\left(\overline{Y}\right) = \frac{\sigma}{\sqrt{n}}$$
  - If the test set $\mathcal{T}_0$ contains $|\mathcal{T}_0|$ units then the standard deviation of the APSE will decrease proportionately to $1/\sqrt{|\mathcal{T}_0|}$.

Thus, the larger $|\mathcal{T}_0|$ is, the better (i.e., less variable) will be our estimate of the APSE.

- Conversely, the larger $|\mathcal{T}_0|$ is, the smaller $\mathcal{S}_0$ will be.
  - The smaller the training set is, the lower the quality of the estimated predictor function $\widehat{\mu}_{\mathcal{S}_0}(\mathbf{x})$
  - (This could easily lead to systematically underestimating the predictor accuracy for the full population)

- Clearly, choosing a sample size requires some trade-off between the variability and the bias of the estimate predictor function.

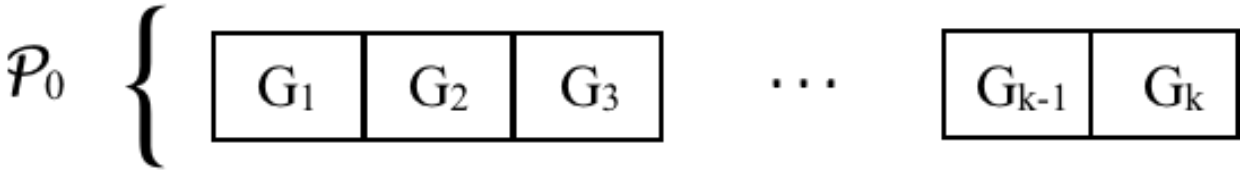
General recommendation: 80-20 split (training / test)

5.3.2.2 The Multiple Subset Version

- Every one of the concerns discussed above is also pertinent here when considering how to choose $\mathcal{S}_{0,j}$ from $\mathcal{S}$, but now there is an additional consideration:
  - How many samples $\mathcal{S}_{0j}$ should we take? One? Many? How many?

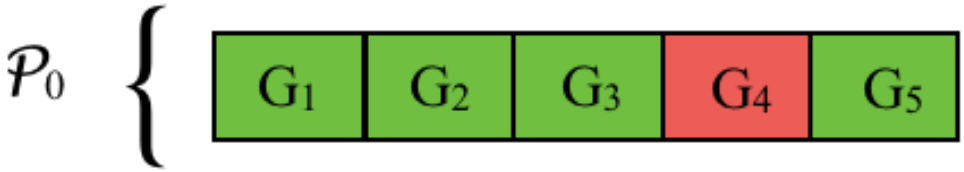
- A simple way to create a collection of samples $\mathcal{S}_{0j}$ is to
  - partition $\mathcal{P}_0$ into pieces, or groups
  - then select some groups to form $\mathcal{S}_{0,j}$ and the remainder to form $\mathcal{T}_{0,j}$.

- Typically, $\mathcal{P}_0$ is partitioned into k groups $G_1, G_2, \ldots, G_k$ of equal size (approximately equal in practice). We call this a ***k*-fold partition** of $\mathcal{P}_0$:



- Selecting any set of groups from the partition will define a sample $\mathcal{S}_{0,j}$ and the remaining groups will define its complement $\mathcal{T}_{0,j}$.

- The most common method of selecting the groups would be to select $k-1$ groups to form $\mathcal{S}_{0,j}$ and the remaining group forms $\mathcal{T}_{0,j}$.
  - For example, when $k = 5$ we have the following partition of $\mathcal{P}_0$ with the green groups forming the sample and the red group forming the test.
  - In this case, $\mathcal{S}_{0,j} = G_1 \cup G_2 \cup G_3 \cup G_5$ and $\mathcal{T}_{0,j} = G_4$.



- Note that for a k-fold partition there can only be k different pairs of sample $\mathcal{S}_j$ and test set $\mathcal{T}_j$. That is $N_{\mathcal{S}} = k$.

Calculating

$$\widehat{APSE}(\mathcal{T}, \widetilde{\mu}) = \frac{1}{N_S} \sum_{j=1}^{N_S} APSE(\mathcal{T}_{0,j}, \widehat{\mu}_{\mathcal{S}_{0,j}})$$

using sampling that selects all $k-1$ groups from a k-fold partition is known as k-**fold cross-validation** in the literature.

Remaining Questions

1. How should the partition be constructed?
   - Simple random sampling is the obvious choice, but there may be contexts where other sampling protocols might also be considered.

2. What value should k take?
   - Clearly a large value of k will produce a large sample $\mathcal{S}_{0,j}$ but a smaller test set $\mathcal{T}_{0,j}$
   - A predictor based on a larger $\mathcal{S}_{0,j}$ should be closer to that based on all of $\mathcal{S}$, but would lead to smaller $\mathcal{T}_{0,j}$ and thus a less precise estimate of the prediction error
   - A predictor based on a smaller $\mathcal{S}_{0,j}$ should perform more poorly (being based on fewer observations) and so tend to systematically overestimate the prediction error
   - This suggests that there exists is a bias-variance trade-off that must be considered when selecting k
   - Choosing an optimal value of k is difficult, but **experience and related literature** suggest that k=5 or k=10 often work well to balance the bias-variance trade-off