IE 7374 ST: Machine Learning in Engineering

Spring 2021

LAB 2- Cross Validation and Model Selection, due by March 28, 2021

The data for this lab are in the file lab2.csv. Cross-validation (CV) is a very general tool for:

- Assessing how well a fitted model will predict new data, and/or
- Selecting the best model, based on how well we think the model will predict a new set of "test" data, when we do not have the luxury of setting aside a real set of test data.

CV serves the same purpose as Mallows' C_p statistic or Akaike's AIC_p statistic (both of which covered in statistics course) but is completely general and can be used to assess the quality of virtually any type of fitted model in any scenario. In contrast, theoretically derived criteria like C_p and AIC_p only apply to the specific types of models on which their derivations were based, such as linear least squares regression (for C_p) or maximum likelihood estimation (for AIC_p). In addition, AIC_p and (to a lesser extent) C_p are based on asymptotic approximations that render them strictly valid only for large sample sizes. Whereas there are many assumptions and approximations in the derivation of theoretical criteria like C_p and AIC_p, CV is an entirely empirical measure that involves no assumptions and is, therefore, almost universally applicable.

Suppose we have a data set consisting of n observations of a single response variable y and k predictor variables $x_1, x_2, ..., x_k$ and we want to assess the quality of some model of a particular structure (e.g., a linear regression model containing a subset of the predictors) that we intend to fit to the data. Here, "quality" means that we would like to assess how well the fitted model would do at predicting a new set of data that was not used to fit the model. It is pointless to use the prediction over the training set as a measure of how well the model would predict over a new test set of data. Indeed, a model having more predictors will always fit the training data better than a model with fewer predictors. If we have enough data to set aside a "test" set, while still leaving enough data in the "training" set to fit the model, then we can fit the models to the training data and choose the best model as the one that does the best at predicting the test data. For example, we might set aside 1/3 of the data for test purposes and use the remaining 2/3 for training. The roughly n/3 observations that are put in the test set must be chosen randomly. Conceptually, it looks like figure 1:

If we do this, however, it means that some of our data are not available for fitting purposes, which is generally undesirable. An alternative that does not involve this tradeoff is CV.

We will illustrate the basic idea behind CV with 5-fold CV: First, randomly split the n observations into 5 parts of roughly equal size (as equal as we can get them). Let n_j denote the exact number of observations in Part j ($n_j = n/5$ roughly), so that $n = \sum_{j=1}^{5} n_j$.

The following figure 2 is an illustration:
Then, for $j = 1, 2, ..., 5$, we:

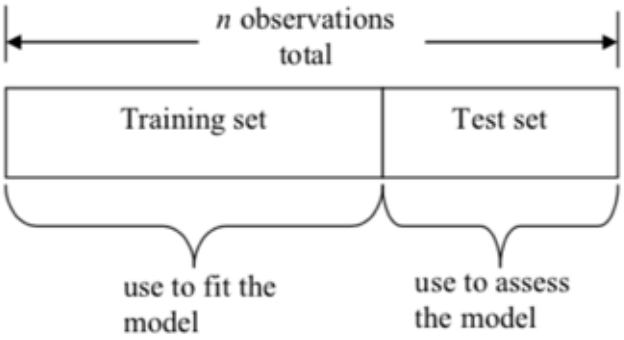

Figure 1: Cross-validation

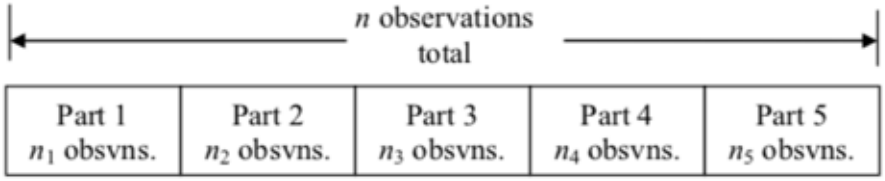

Figure 2: Illustration

1. Set aside Part j as a test set.
2. Fit the model to the remaining 4 parts, taken as a single training set of roughly (4/5)n observations.
3. Use the fitted model from Step 2 to predict the n_j response values in Part j (the test set) and calculate the resulting test error sum of squares for Part j:

$$SSE^j = \sum_{\text{all i in part j}} (y_i - \hat{y}_i^j)^2,$$

where $\hat{y}_i^j$ denotes the prediction of y_i (from Part j) using the model fitted in Step 2 (which excluded Part j). The following figure 3 illustrates for $j = 2$

After repeating Steps 1—3 for $j = 1, 2, ..., 5$ (i.e., for each Part set aside as a test part), we finally calculate the cross-validation SSE:

$$SSE_{CV} = \sum_{j=1}^{5} SSE^j$$

We can use SSE_{CV} as a direct measure of the quality of our model, in terms of how well it would do predicting a new set of data, independent from the training set.

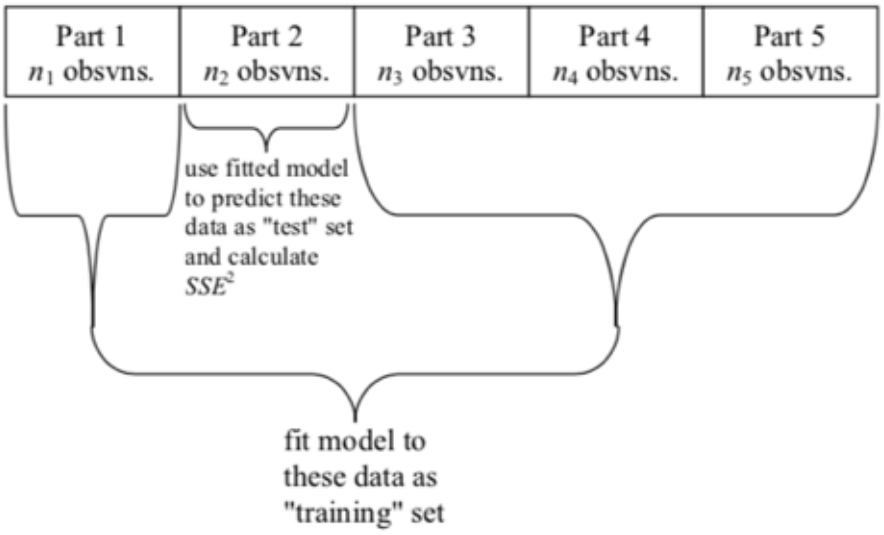

Figure 3: Another Illustration

Comments:

1. Note that we have actually fit five different models (one for each $j = 1, 2, ..., 5$). The final, single model that we would use for any future purposes should be fit to the entire set of n observations. It does not matter that this model will have slightly different parameters than the five models fit in the above procedure; we can still use SSE_{CV} as a measure of how well we expect the final single model to perform.

2. CV is useful for model comparison and selection. Suppose we want to compare a number of different fitted models (e.g., linear regression models with different sets of predictors, ridge regression models with different δ, a neural network model, other nonlinear regression models, etc.) and select the "best" one. For each model we can calculate an SSE_{CV} as described above. Then, to select the "best" model, we simply take the one that has the smallest SSE_{CV} . This is analogous to selecting the model with the smallest C_p in best subsets regression, except CV applies to virtually all types of models.

3. SSE_{CV} measures directly and empirically what C_p and AIC_p (which will be introduced later in this course) are intended to measure analytically. Namely, how well the fitted model can predict a new set of test data. Because CV involves fewer approximations and assumptions than the analytical methods, it is much more broadly applicable and generally more accurate. The only drawback is that CV is more computationally expensive.

4. K-fold CV is the same as the 5-fold CV described above, except that we divide the data into K parts of roughly equal size and repeat Steps 1—3 for $j = 1, 2, ..., K$. The most common choices are $5 \leq K \leq 10$, although n-fold CV is sometimes used. SSE_{CV} for n-fold CV is precisely the $PRESS_p$ statistic that was (or will be) introduced in the context of best subsets regression for comparing linear regression models.

5. The measure of quality you use in CV does not necessarily have to be the SSE. For example, in Step 3 of the CV procedure, you could instead calculate the sum of the absolute error

(SAE):

$$SAE^j = \sum_{\text{all i in part J}} |y_i - \hat{y}_i^j|,$$

6. IMPORTANT REMINDER: The n observations should always be randomly assigned to each Part.

The following is the Lab 2 assignment, which you should turn in as a group.

1. Consider the data in Lab2.csv. The response is y = weight, and there are 8 predictor variables. There are a total of n = 79 observations. Write your own script to perform 10-fold CV to choose between the two models: The eight-predictor model versus the two-predictor model with only x_1 and x_2. Do NOT look for an package or library to do CV automatically. This lab is intended to be an exercise in Python scripting, as well as way to make sure you understand exactly what happens in CV. Your script should do the following:

   - Randomly assigns the 79 observations to the 10 parts, but under the constraint that the sizes of the parts are as balanced as possible (i.e., 8 observations in each of 9 parts, and 7 observations in the last part).
   - Uses the same partition to compare the two models.

   Explain why both of these are important.

2. Use CV to calculate the SSE_{CV} values for the two models and decide which is the better model. What are the respective SSE_{CV} values for the two models?

3. Replicate the CV procedure twenty times, each time using a different random partition of the 79 observations into 10 parts (but, use the same partition for both models). List the SSE_{CV} values for the two models for each of the twenty replicates. Discuss whether or not the results are consistent from replicate-to-replicate. Explain how you might reconcile this inconsistency by combining the results across all twenty replicates.

4. After combining the results across all twenty replicates, what is your conclusion regarding whether the 8-predictor model or the 2-predictor model is preferable? Discuss in some detail.

Turn in your lab report, and also post an electronic version of your Python script and results that are reproducible.