Linear Data Analysis
Cross Validating Linear Regression

Cain Susko

Queen's University
School of Computing

February 9, 2022

a Validation of Linear Regression

Validation is a way of assessing a linear regression. Cross validating linear regression means to validate a regression with a subset of it's data.

A common form of cross validation is k-fold cross validation, where we cross reference with a subset of size k. This topic is covered in Monday, Week 5 Tutorial.

Recall

Each observation is $\vec{x}_i$. A data matrix is a matrix of m observations:

$$A = \begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_m^\top \end{bmatrix}.$$

$$\begin{aligned} &\text{Linear Regression:} && A\vec{u} \approx \vec{c} \\ &\text{Standardize Data:} && A \to X \wedge \vec{c} \to \vec{y} \\ &\text{Standard Problem:} && X\vec{w} \approx \vec{y}. \end{aligned}$$

Validation is to confirm that output of a model is acceptable. Linear Regression is done over the independent data in A and the dependent data $\vec{y}$. The usual technique for Validation is the Root Mean Squares method. Given the proposed solution $\vec{w}$ is:

$$RMS\left(X, \vec{y}; \vec{w}\right) = \sqrt{\frac{[X\vec{w} - \vec{y}]^\top [X\vec{w} - \vec{y}]}{m}}.$$

This measures the *fit* of the model $\vec{w}$ to **all** data.

b Training Sets and Testing Sets

Training is defined as: using data to find $\vec{w}$.
Testing is to evaluate the model using $\vec{w}$.

Previously, we used all data $X, \vec{y}$ to find $\vec{w}$ and to evaluate the regression. Instead, we can leave one or n observations of $\vec{x} \in X$ and $y \in \vec{y}$ out of our training. Then, we can test using the left out observation.

Thus, we should 'Hold Back' $\vec{x}_i, y_i$ and train on the remaining X $(\vec{x}_{i+1} : end)$, then test on $\vec{x}_i, y_i$. We can then repeat this for all observations where we test the data on the respective data left out.

This procedure detects 'singleton' statistical outliers.

c k-fold Cross Validation of Linear Regression

we shall explore what's called k-fold validation.

Consider: the partition of data:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{y} = \begin{bmatrix} \vec{y}_1 \\ \vec{y}_2 \end{bmatrix}.$$

Where we train two models:

$$X_1\vec{w}_1 \approx \vec{y}_1 \quad X_2\vec{w}_2 \approx \vec{y}_2.$$

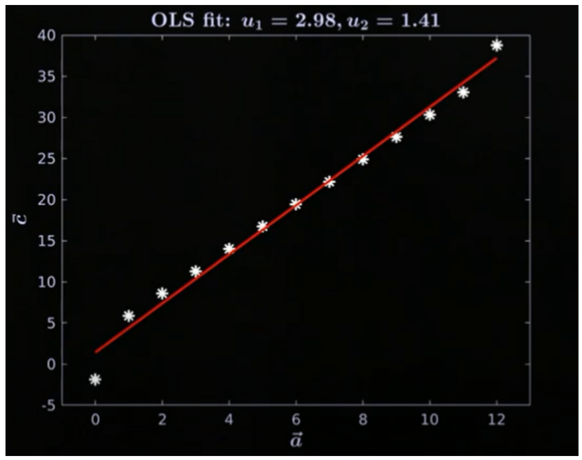
To cross validate these models, we will use the data from each to test the other using RMS.

$$RMS(X_2, \vec{y}_2; \vec{w}_1) \quad RMS(X_1, \vec{y}_1; \vec{w}_2).$$

these partitions 1 and 2 are called 'folds', and thus this is 2 fold cross validation. The rule is, for i in range k: train $\vec{w}_i$ on $data_{i \rightarrow [i+(k-1)] \bmod k}$ and test w_i on $data_{i-1}$. Typical values of k are 5 or 10.

d Example of Five Fold Validation

Given a dataset derived from a 2D line with 2 outliers. With this data $\vec{a}$ we will first augment it to $\vec{a} \to A$ We will then solve the approximation $A\vec{u} \approx \vec{c}$ and set the result to be $\hat{u}$.



Finally, we will compute the Root Mean Square Error $RMS(A, \vec{c}; \hat{u})$
Which results in a Root Mean Square Error of 1.3376.
But how valid is this RMSE? If we do a five fold cross validation on the data where we train on $len(A) - 5$ data and then test on 5 data, And then compare the mean of the RMSE of the training runs with that of the test runs.

OLS: 5-Fold, 10 RMSE

	TRAIN	TEST
	0.7372	1.1496
	1.1504	1.8840
	1.2280	1.7711
	1.3208	1.8735
	1.2622	2.2049
	1.2340	1.8694
	1.2712	1.7658
	0.0000	0.0000
	0.7353	1.1707
	1.3415	1.6849
Mean:	1.0280	1.5374
Std:	0.4246	0.6294

By the result we can see that the mean error of the training is less than the test error. In fact, the error is about 50% higher in testing than training. This implies that the current model is not a good fit for the data.

Learning Summary

Students should now be able to:

- Validate linear regression with RMS Error
- Implement k-fold Cross Validation
- Assess the training errors and testing errors of a linear regression using cross validation