CS6301: R For Data Scientists

LECTURE 13: CROSS VALIDATION

We Have A Problem …

We have seen how to build a model that will fit a given data set – but is this what we really want to do?

We really want to fit future data points well

How do we know our model will provide good future predictions?

We need a new approach …

Assessing Model Fit

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Residual Sum of Squares – want this small

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}},$$

Residual Standard Error – point estimator for standard deviation of noise, ϵ

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

R2 – how much of TSS is accounted for by model. Want this close to '1'

where $\mathrm{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

Mean Square Error – approximation to $E(y_i - \hat{y}_i)^2$

Validation Set (Simple Approach)

- Randomly divide dataset into training and test subsets
- Create model on training set
- Calculate test MSE on test set
- Problem: May be highly variable; different selections of training sets can produce very different test MSEs for the same model

LOOCV

- For each data point in the dataset, do the following:
  - Compute the model leaving the point (x_i, y_i) out
  - Compute $MSE_i = (y_i - \hat{y}_i)^2$
- Then, compute an overall average test MSE:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

K-fold Cross Validation

- Divide the dataset in *k* folds, and for each fold:
  - Compute the model leaving the fold out
  - Compute $MSE_i = \frac{1}{|F|}\sum(y_i - \hat{y}_i)^2$, the MSE for the fold
- Then, compute an overall average test MSE:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

K-fold Versus LOOCV

- Test MSE is made up of two parts: bias and variance
- LOOCV tends to overestimate test MSE variance
  - The models are highly correlated
- K-fold tends to overestimate the bias
- Generally, K-fold is more accurate