Linear Model Performance

UIS CS570:Essentials of Data Science
Mike Bernico

RMSE

For one observation:

$$\hat{y} - y_i$$

For all observations in the test set:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{y} - y$$

The problem with this is that some values will be positive, some negative.
Lets adjust for that...

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y)^2}$$

Scikit doesn't have an RMSE function, but it does have MSE – Can just apply math.sqrt()

Coefficient of Determination

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$SS_{total} = \sum (y_i - \bar{y})^2$$ Says something about the total variance in the data set.

$$SS_{residual} = \sum (y_i - \hat{y})^2$$ Says something about variance from actual to predicted

$$R^2 = 1 - \frac{UnexplainedVariance}{ExplainedVariance}$$

If all the the variance in the model was explained (error was 0) then R^2 = 1

R^2 will be between 0 and 1, larger is better

Is My Model Good Enough?

- That Depends
  - On Application Primarily

  - Metrics are best though of as heuristics

  - Best used to compare two models on the same data