Linear Regression: Recap

Land Belenky

galvanize

What is a Linear Regression Model?

- A set of values (weights, coefficients, or Beta) such that the sum of the products of the weights with features (plus a constant) produces a value, which is a *prediction*
- Note: the “sum of products of weights and features” is a dot product, if weights and features are both vectors

How do we know if a model is good?

- A good model is one which produces *predictions* that are pretty close to *actual*
- Let’s be more mathematical:

$$Predictions = \beta_0 + Features \cdot Weights$$

$$Predictions = \beta_0 + \sum_i x_i \cdot \beta_i$$

$$Actuals = y$$

$$Error = (Actuals - Predictions)^2$$

$$Error = (y - \beta_0 + \sum_i x_i \cdot \beta_i)^2$$

How are the values Beta determined?

Beta values are determined by minimizing the *Error* (or cost function)

$$Error = (y - \beta_0 + \sum_i x_i \cdot \beta_i)^2$$

Two methods for minimizing the cost function:

1. Analytically - set the derivative of the error function to zero and solve for beta
2. Iteratively - pick random values for Beta, follow gradient descent towards lower error

Error Function

- Note that the error function has a dual purpose:

1. It lets us measure (and communicate) how good our model is
2. It guides us to creating the model

Bias-Variance Trade off and Curse of Dimensionality

- More features will always result in lower measured error on training data
- Necessary to use cross-validation to verify model can predict on *new* data, not just training data.
  - Train-Test Split
  - K-Fold Cross Validation: average out effects of outliers
  - Leave-One-Out Cross Validation: extreme version of K-fold
- Additional Features carry “burden”. Model complexity should only be increased if the improvement in predictions is greater than this burden

How to determine best number of features to use:

1. Exhaustively: try all combinations of features and see which one produces lowest error on hold-out data.
2. Forward Selection: Start with single most predictive feature. Add additional features only if the result is improvement to predictions larger than complexity cost.
3. Backward Elimination: Start with all features, eliminate any that are not predictive

Select the features to include

- Information Criteria:
  - Models are credited for better predictions
  - Models are debited for greater complexity

$$C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2)$$

$$AIC = -2logL + 2 \cdot p$$

$$BIC = \frac{1}{n}(RSS + log(n)p\hat{\sigma}^2)$$

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

  - C_p
  - AIC
  - BIC
  - Adjusted R^2

Next Steps:

- Determine and adjust the *Magnitude* of the weights to use
- Determine the *confidence* in the value of these weights