CIS 8695

Regularization: Ridge Regression & Lasso

Ling Xue

Computer Information System

Georgia State University

Key Issues

- Main purpose of regularization
- Model specifications
  - Ridge regression
  - Lasso
- Cross-validation
- Prediction using regularization

Main Purpose of Regularization

- **Regularization**. This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

- Using a penalty
  - L2-norm (**Ridge Regression**): the normalization term is the sum of the squared weights
  - L1-norm (**Lasso**): the normalization term is the sum of the absolute values of weights

Ridge Regression

- Linear regression minimizes **Residual Sum of Squares (RSS)**

$$RSS = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

- Ridge regression minimizes:

$$\sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2 + \lambda * \beta_1^2 = RSS + \lambda * \beta_1^2$$

- $\lambda \geq 0$ is a tuning parameter, to be determined separately. The term $\lambda\beta_1^2$ is a shrinkage penalty that decreases when the β parameters withdraw (shrink) towards the zero. Parameter λ controls the relative impact of the two components: RSS and the penalty term. If $\lambda = 0$, the Ridge regression coincides with the least squares method. If $\lambda \rightarrow \infty$, all estimated coefficients tend to zero. The regression Ridge produces different estimates for different values of λ. The optimal choice of λ is crucial and is usually done with cross-validation.

Lasso Regression

- Ridge regression minimizes:

$$\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2 + \lambda * \beta_1^2 = RSS + \lambda * \beta_1^2$$

- Lasso regression minimizes:

$$\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2 + \lambda * |\beta_1| = RSS + \lambda * |\beta_1|$$

- The term λ|β1| is a shrinkage penalty for the Lasso regression. It is basically minimizing the sum of the absolute differences between the target value and the estimated values.

Ridge vs Lasso

- The Ridge regression produces a model with all the variables, of which the part with coefficients is closer to zero. Increasing λ forces more coefficients to be close to zero, but almost never exactly equal to zero, unless $\lambda = \infty$. For forecasting this is not a problem, while interpretation can sometimes be problematic. Lasso regression tries to overcome this aspect.

- The Lasso regression penalty term, using the absolute value (rather than the square, as in the regression Ridge), forces some coefficients to be exactly equal to zero, if λ is large enough. In practice, Lasso automatically performs a real selection of variables.

Cross-Validation: K-Fold

- **Objective:** ensure that learning results are generalizable
- **Steps**
  - The original sample is randomly partitioned into k equal sized subsamples.
  - Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.
  - The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data.
  - A single estimation is produced through aggregation. There are different ways of aggregation, e.g.,
    - Averaging
    - Optimal selection