Bias Variance Tradeoff

Statistical Learning

April 22, 2024

This essay discusses four statistical learning methods whose objective functions involve an adjustment to the model fit using regularization.

Often, there is a case that some or many of the variables used in a multiple regression model are in fact not associated with the response and including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables, that is, by setting the corresponding coefficient estimates to zero, we can obtain a model that is more easily interpreted.

We choose methods such as Subset Selection, Shrinkage,and Dimension Reduction among many others to select the p predictors. The subset selection methods involves fitting a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.Regularizing the coefficient estimates significantly reduces the variance without substantial increase in its bias.Regularization technique is therefore used in statistical learning to prevent over fitting by adding a penalty term to the objective function. This penalty term discourages overly complex models, leading to better generalization performance on unseen data.

Few best-known techniques using regularization the regression coefficients towards zero are ridge regression, lasso regression, Elastic Net Regression, and Support Vector Machines.

1. Ridge Regression:

Ridge regression minimizes the residual sum of squares (RSS) along with a penalty term that is proportional to the sum of the squared coefficients.

Mathematical formulation = minimize ($RSS + \lambda \sum_{j=1}^{p} \beta_j^2$),

where the penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called Shrinkage Penalty added to the RSS, where λ is the regularization parameter. Ridge regression shrinks the coefficients $\beta_1, \beta_2, ..., \beta_p$ towards zero, but they never reach exactly zero.

We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when all(xi=0).

As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.Unlike best subset,forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.

2. Lasso Regression:

Lasso regression minimizes the RSS along with a penalty term that is proportional to the sum of the absolute values of the coefficients.

Mathematical formulation = minimize ($RSS + \lambda \sum_{j=1}^{p} |\beta_j|$)

Lasso regression imposes sparsity on the coefficients by shrinking some of them exactly to zero. This leads to automatic feature selection, making it useful for models with a large number of predictors.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.However, in the case of the lasso, the 1% penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression. We say that the lasso yields sparse models—that is, sparse models that involve only a subset of the variables.

3. Elastic Net Regression:

Elastic Net regression is a hybrid approach that combines the penalties of both Lasso (L1 regularization) and Ridge (L2 regularization) regression. It aims to address the limitations of each method while leveraging their strengths,providing a balance between ridge and lasso regression.

Objective Function = minimize ($RSS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$)

In Elastic Net regression, the objective function consists of two penalty terms: one that penalizes the absolute values of the coefficients (L1 penalty) and another that penalizes the squared values of the coefficients (L2 penalty). The objective function seeks to minimize the sum of the squared residuals along with the combined regularization penalties. Elastic Net addresses some of the limitations of ridge and lasso regression. It can handle highly correlated predictors better than lasso and selects groups of correlated predictors together.

- Elastic Net combines the advantages of Lasso and Ridge regression: it handles multicollinearity better than Lasso and performs feature selection by shrinking some coefficients to exactly zero, while still maintaining the advantages of Ridge regression in stabilizing the estimates.
- It is particularly useful when dealing with datasets with a large number of predictors, many of which may be correlated.
- Elastic Net encourages grouped selection of correlated variables, which can be advantageous when predictors are grouped together.
- The choice of λ_1 and λ_2 parameters is crucial, and techniques such as cross-validation are often used to select optimal values.

4. Support Vector Machine:

Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find the hyperplane that maximizes the margin between different classes in the feature space.

Objective Function: The objective of SVM is to find the optimal hyperplane that separates the data into different classes while maximizing the margin between the classes. Additionally, SVM introduces a regularization parameter (often denoted as C) to control the trade-off between maximizing the margin and minimizing the classification error.

Objective Function= minimize ($1/2||w||^2 + C\sum_{i=1}^{n} \xi_i$) subject to $y_i(w^T x_i + b) >= 1 - \xi_i$, for all i=1,2,3,...,n - SVM is effective in high-dimensional spaces and is particularly useful when the number of features exceeds the number of samples.

- It uses the kernel trick to transform the input space into a higher-dimensional feature space, allowing it to handle non-linear decision boundaries.
- SVM aims to maximize the margin between classes, making it less prone to overfitting.
- The regularization parameter C controls the trade-off between maximizing the margin and minimizing the classification error. A smaller C value leads to a larger margin but may result in more misclassifications on the training data.

Overall, all the above are powerful techniques in statistical learning that incorporate regularization to improve generalization performance and handle complex data. They are widely used in various domains for predictive modeling tasks.
