Lecture 12

Penalized Regression:

Penalized regression is a type of regression analysis that involves adding a penalty term to the ordinary least squares (OLS) objective function in order to encourage simpler models that generalize better to new data. The penalty term is a function of the model parameters and is designed to reduce overfitting, which occurs when a model is too complex and fits the training data too closely, leading to poor performance on new data.

There are two main types of penalized regression: Lasso regression and Ridge regression.

Lasso regression adds an L_1 (linear, absolute value) penalty term to the objective function, which encourages sparsity in the model by setting some of the model coefficients to zero. This has the effect of selecting only the most important predictors and can help to prevent overfitting.

Ridge regression adds an L_2 (squared) penalty term to the objective function, which has the effect of shrinking the model coefficients towards zero. This can also help to prevent overfitting by reducing the impact of predictors that are less important.

Penalized regression can be particularly useful when dealing with high-dimensional datasets with many predictors, where traditional regression methods may struggle to produce accurate models. By reducing the complexity of the model, penalized regression can produce more reliable and interpretable results, while still accounting for the relationships between the predictors and the response variable.

Ridge regression is a type of penalized regression that adds an L_2 penalty term to the ordinary least squares (OLS) objective function. The L_2 penalty term is a function of the model coefficients (also called the regression parameters or weights) and is designed to shrink them towards zero. This has the effect of reducing the impact of predictors that are less important and can help to prevent overfitting, which occurs when a model is too complex and fits the training data too closely, leading to poor performance on new data.

The Ridge regression objective function is:

$$RSS \, + \, \alpha \sum \beta_i^2$$

where RSS is the residual sum of squares (the difference between the predicted and actual values of the response variable), β_i is the ith regression coefficient, and α is the regularization parameter that controls the strength of the penalty. The larger the value of α, the greater the penalty and the more the coefficients are shrunk towards zero. (Sometimes λ is used instead of α. In either case, the value must be greater than zero.)

The Ridge regression solution is obtained by minimizing the objective function with respect to the coefficients, subject to the constraint that the sum of their squares is less than or equal to a certain value (called the ridge constraint). This constraint ensures that the coefficients are not too large and the model is not too complex.

One important feature of Ridge regression is that it can handle multicollinearity, which occurs when two or more predictors are highly correlated with each other. In this case, the OLS method may produce unstable or unreliable estimates of the regression coefficients, but Ridge regression can help to stabilize the estimates and improve the predictive performance of the model.

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a type of penalized regression that adds an L_1 penalty term to the ordinary least squares (OLS) objective function. The L1 penalty term is a function of the absolute values of the model coefficients (also called the regression parameters or weights) and is designed to force some of them to exactly zero. This has the effect of selecting only the most important predictors and can help to prevent overfitting, which occurs when a model is too complex and fits the training data too closely, leading to poor performance on new data.

The LASSO regression objective function is:

$$RSS + \alpha \sum |\beta_i|$$

where RSS is the residual sum of squares (the difference between the predicted and actual values of the response variable), β_i is the ith regression coefficient, and α is the regularization parameter that controls the strength of the penalty. The larger the value of α, the greater the penalty and the more the coefficients are shrunk towards zero.

The LASSO regression solution is obtained by minimizing the objective function with respect to the coefficients, subject to the constraint that the sum of their absolute values is less than or equal to a certain value (called the LASSO constraint). This constraint ensures that some of the coefficients are exactly zero, leading to a sparse model.

One important feature of LASSO regression is that it can perform variable selection, which means that it automatically selects a subset of the predictors that are most relevant for predicting the response variable. This can be useful for reducing the dimensionality of the problem and improving the interpretability of the model.

To perform ridge regression or LASSO regression, we can use the glmnet package in R. This package uses alpha to mean something else. We will set alpha to be 0 to get ridge regression, equal to 1 is LASSO regression, and a value between 0 and 1 equivalent to elastic net, which is a type of mixed penalty. This package will also allow us to select the best value of the parameter (λ) to get the best results.

One of the biggest differences between Ridge and LASSO regression is that LASSO regression is a form of model selection in that it can slink coefficients in the model all the way to zero. Ridge regression can make them very small, but they will remain non-zero.

AIC & BIC:

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are statistical measures that can be used to compare the fit of different regression models. They are based on the principle of parsimony, which suggests that simpler models are generally better than more complex models, all else being equal.

AIC and BIC take into account both the goodness of fit of the model and the complexity of the model. The lower the AIC or BIC value, the better the model is considered to be. AIC is defined as:

$$AIC = -2\, log\ likelihood + 2k$$

where k is the number of parameters in the model. The log likelihood measures how well the model fits the data, and the penalty term $2k$ is added to account for the complexity of the model.

BIC is similar to AIC, but the penalty term is more severe:

$$BIC = -2\, log\ likelihood + log(n)\, k$$

where n is the sample size. The penalty term $log(n)$ is larger than $2k$ for large sample sizes, which means that BIC tends to prefer simpler models than AIC.

In general, AIC and BIC can be used to compare different regression models with different numbers of parameters. If two models have similar goodness of fit, the model with the lower AIC or BIC is considered to be the better model. However, AIC and BIC are not absolute measures of model quality, and should always be used in conjunction with other model diagnostics and expert judgment.

Model selection – review and continuation

Model selection methods in multiple regression refer to the process of selecting a subset of predictors (independent variables) from a larger set of potential predictors in order to build the best possible regression model for a given data set. There are several methods for performing model selection in multiple regression, including:

Forward selection: This method starts with an empty model and sequentially adds predictors to the model that provide the greatest improvement in fit until a stopping criterion is met.

Backward elimination: This method starts with a model that includes all predictors and sequentially removes the predictor with the smallest contribution to the model until a stopping criterion is met.

Stepwise selection: This method combines forward selection and backward elimination by alternating between adding and removing predictors until a stopping criterion is met.

All subsets regression: This method involves fitting all possible models that can be formed from a given set of predictors and selecting the best one based on a model selection criterion.

Information criteria: These are statistical criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), that provide a way to compare the fit of different models based on the tradeoff between goodness of fit and model complexity.

Each of these methods has its own advantages and disadvantages. For example, forward selection is computationally efficient but can be prone to overfitting, while all subsets regression is more accurate but can be computationally expensive for large numbers of predictors. Information criteria provide a way to balance model fit and complexity, but can be sensitive to the sample size and the type of data being analyzed.

Resources:

1. http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/
2. https://machinelearningmastery.com/penalized-regression-in-r/
3. https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf
4. https://bookdown.org/andreabellavia/mixtures/penalized-regression-approaches.html
5. https://www.r-bloggers.com/2021/05/lasso-regression-model-with-r-code/
6. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/selection-process-for-multiple-regression/
7. https://statisticsbyjim.com/regression/model-specification-variable-selection/
8. https://www.statology.org/ridge-regression-in-r/
9. https://www.statology.org/lasso-regression-in-r/