Introduction to Statistical Learning

November 29, 2020

Linear Model Selection and Regularization

As seen in previous chapters, linear models are usually fit using the least squares method. Nonetheless, all predictors are not necessarily relevant to predict the response. Other fitting methods are used to improve:

1. Prediction accuracy: when n is not much larger than p, there can be a lot of variabiliy in the least squares fit, resulting in overfitting. When $n < p$, the variance is infinite, so the method cannot be used at all. *Constraining* and *shrinking* the estimated coefficients allow to reduce the variance and thus improve the predictions on data that have not been used to train the model.
2. Model interpretability: *feature selection* and *variable selection* allow to set the coefficients of the irrelevant features to zero. It improves the model interpretability.

Subset Selection

The goal of these methods is to select a subset of relevant predictors out of all the p predictors. These methods result in the selection of a set of models which contains a subset of the p predictors. The best model is then selected among them.

Best Subset Selection

A simple but costly approach We fit the least squares on all possible combinations of p predictors and choose the best model (using AIC, BIC or adjusted R^2). It's computationally very costly as there are 2^p possible models. Infeasible for $p > 40$.

Stepwise Selection

Forward Stepwise Selection Begins with no predictors. Adds one predictor at-a-time, until all predictors are in the model: each time, the predictor that leads to the greatest improvement in prediction accuracy is added. It's a lot more efficient computationally than the best subset selection, but it does not guarantee to select the best out of the 2^p possible models. For instance: best possible one-variable model contains X_1 and best possible two-variable model only contains X_2 and X_3. It is the only subset selection method that can be used in high-dimentional settings where $n < p$.

Backward Stepwise Selection Begins with the full least squares model containing the p predictors. Removes the least useful predictor one-at-a-time.

Hybrid Approaches Combines both forward and backward selection methods: adds the most useful variable sequentially and removes the variables that are no longer useful at each step.

Choosing the Optimal Model

We recall well that training MSE is often an underestimate of the test MSE. RSS and R^2 cannot be good indicators of the goodness of fit, as linear models tend to overfit the training data when degrees of freedom are added. The following technique are ways to estimate the test MSE during the training phase.

C_p C_p adds a penalty to adjust to the fact that the training error tends to underestimate the test error: $C_p = \frac{RSS + 2d\hat{\sigma}^2}{n}$ where d is the number of predictors in the current subset of predictors and $\hat{\sigma}^2$ is a variance estimator. C_p tends to penalize big models, so we tend to choose the model with the lowest C_p.

Akaike Information Criterion (AIC) It is defined by maximum likelihood. In case the Gaussian errors setting, maximum likelihood and least squares are the same thing. In this case, it is given by $AIC = \frac{1}{b\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$.

Bayesian Information Criterion (BIC) It is derived from a Bayesian point of view and is given by $BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$. It tends to select smaller models than C_p as $\forall n > 7, \log(n) > 2$.

Adjusted R^2 It is given by Adjusted $R^2 = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}$. Unlike C_p, AIC and BIC, a model with a larger adjusted R^2 is better. The idea behind adjusted R^2 is to *pay a price* for the inclusion of an additional predictor to the model. Adding a predictor means increasing d and thus increasing $\frac{RSS}{n-d-1}$. In theory, the model with the highest adjusted R^2 will only have relevant variables and no noise variables.

Cross-validation Nowadays, cross-validation is not as computationally costly as it used to be, so cross-validations are often a good way to perform model selection.

Shrinkage Methods

The goal of these methods is to reduce the estimator's variance by shrinking coefficients' estimates towards zero. This set of methods is also called *regularization*.

Ridge Regression

In Ridge Regression, the coefficients are estimated by minimizing the quantity $RSS + \lambda \sum_{j=1}^{n} \beta_j^2$ where $\lambda \geq 0$ is a *tuning parameter*. The second term $\lambda \sum_{j=1}^{n} \beta_j^2$, called the *shrinkage penalty*, is small when the coefficients are close to zero. Hence, the Ridge Regression tends to shrink the estimates towards zero. The tuning parameter λ serves to control the relative incidence of the two terms on the coefficient estimates. The choice of λ is crucial, as $\hat{\beta}_\lambda^R$ is different for each value of λ.

The Lasso

Selecting the Tuning Parameter

Dimension Reduction Methods

Project the predictors in a M-dimensional space with $M < p$ and fit the least squares on these M predictors.