CheatSheet

Computational Statistics

Wenjie Tu

Spring 2021

- LASSO is a less flexible approach than linear regression since it is more restrictive in estimating the coefficients and sets a number of them to exactly zero.
- LASSO is more interpretable than linear regression.
- As the flexibility of the statistical learning method increases, we observe a montone decrease in the training MSE and a *U-shape* in the test MSE. This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of the statistical method being used.
- Cross-validation is an important method for estimating test MSE using the training data.
- The expected test MSE can be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ (irreducible error).
- In the simple linear regression setting, the squared correlation and the R^2 statistic are identical.
- The square of each t-statistic is the corresponding F-statistic.
- We can see small individual p-values for each variable even in the absence of any true association between the predictors and the response. If we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship. However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors.
- Backward selection cannot be used if $p > n$, while forward selection can always be used.
- Forward selection is a greedy approach, and might include variables early that later become redundant. Mixed selection can remedy this.
- In multiple linear regression, it turns out that it equals $R^2 = \text{Cor}(Y, \hat{Y})^2$.
- When an additional predictor is added to the model, RSS must decrease. It is possible that models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p. $\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$.
- Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for the true underlying function (the reducible error) and the irreducible error.
- The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
- If there is correlation among the error terms, the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be. In addition, p-values associated with the model will be lower than they should be.
- If the *heteroscedasticity* is violated, the standard errors, confidence intervals, and hypothesis tests associated with the linear model will become invalid.

- In case of *heteroscedasticity*, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.
- Even if an outlier does not have much effect on the least squares fit, it can cause other problems (RSE, confidence intervals, p-values, R^2).
- Outliers are observations for which the response y_i is unusual given the predictor x_i. High-leverage points are observations with unusual value for x_i.
- Removing the high leverage observation has a much more substantial impact on the least squares line than removing the outlier. In fact, high leverage observations tend to have a sizable impact on the estimated regression line.
- The leverage statistic is always between $\frac{1}{n}$ and 1, and the average leverage for all the observations is always equal to $\frac{p+1}{n}$. So if a given observation has a leverage statistic that greatly exceeds $\frac{p+1}{n}$, then we may suspect that the corresponding point has high leverage.
- Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for β_j to grow. Recall that the t-statistic for each predictor is calculated by dividing β_j by its standard error. Consequently, collinearity results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that the power of the hypothesis test—the probability of correctly power detecting a non-zero coefficient—is reduced by collinearity.
- As a rule of thumb, a VIF (variance inflation factor) value that exceeds 5 or 10 indicates a problematic amount of collinearity.
- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximatelt normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes.
- Cross-validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility. The process of evaluating a model's performance is known as *model assessment*, whereas the process of selecting the proper level of flexibility for a model is known as *model selection*.
- Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$; however, in this case, it is possible to construct submodels $\mathcal{M}_0, \cdots, \mathcal{M}_{n-1}$ only, since each submodel is fit using least squares, which will not yield a unique solution if $p > n$.
- Though forward/backward stepwise selection considers $\frac{p(p+1)}{2} + 1$ models, it performs a guided search over model space, and so the effective model space considered contains substantially more than $\frac{p(p+1)}{2} + 1$ models.
- The BIC statistic generally places a heavier penality on models with many variables, and hence results in the selection of smaller models than C_p.
- The rationale for *one-standard-error rule* is that if a set of models appear to be more or less equally good, then we might choose the simplest model - that is, the model with the smallest number of predictors.
- The shrinkage penalty is not applied to the intercept in a linear model.
- The standard least squares coefficient estimates are *scale equivariant* whereas the ridge regression coefficient estimates are not. $X_j\hat{\beta}^R_{j,\lambda}$ will depend not only on the value of λ, but also on the scaling of jth predictor. In fact, the value of $X_j\hat{\beta}^R_{j,\lambda}$ may even depend on the scaling of the other predictor. Therefore, it is best to apply ridge regression after standardizing the predictors so that they are all on the same scale.

- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
- If $p > n$, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance.
- In general, a cubic spline with K knots uses a total of $4 + K$ degrees of freedom.
- The general definition of a degree-d spline is that it is a piecewise degree-d polynomial, with continuity in derivatives up to degree $d - 1$ at each knot.
- Regression splines often give superior results to polynomial regression. This is because unlike polynomials, which must use a high degree to produce flexible fits, splines introduce flexibility by increasing the number of knots but keeping the degree fixed.
- Classification error is not sufficiently sensitive for tree-growing, and in practice two other measures are preferable: *Gini index*, *entropy*.
- When building a classification tree, either the Gini index or entropy are typically used to evaluate the quality of a particular split, since these two approaches are more sensitive to node purity than is classification error rate. Any of these three approaches might be used when pruning the tree, but the classification error rate is preferable if prediction accuracy of the final pruned tree is the goal.
- Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen in this book.
- trees can be very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree.
- *Bootstrap aggregation*, or *bagging*, is a general-purpose procedure for reducing the variance of a statistical learning method.
- Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all.
- In nearly all cases, a simultaneous confidence band will be wider than a pointwise confidence band with the same coverage probability.
- The OLS estimate is still a reasonable estimator in the face of non-normal errors. The Gauss-Markov Theorem states that the OLS estimate is the best linear unbiased estimator of the regression coefficients as long as the errors (1. have mean zero; 2. are uncorrelated;, 3. have constant variance). The normality condition comes into play when you are trying to get confidence intervals and/or p-values.
- Situations where the bootstrap can fail: distributions that do not have finite moments, small sample sizes, estimating extreme values from the distribution, estimating variance in survey sampling.
- Some statistical methods, such as K-nearest neighbors and boosting, can be used in the case of either quantitative or qualitative responses.
- Types of confidence intervals in bootstrap. The **percentile CI** takes the relevant percentile to calculate the confidence intervals. The **normal CI** is a modification of the Wald CI. The **studentized CI** corrects for the spread difference and the skew of the original distribution. The **basic CI** corrects for the weird tails of distribution where the percentile CI gives unreliable results. The **BCA** can be unstable when percentiles are outliers.