Method

Liang Hao, Andrew Shibata

10/24/2016

3: Method

The methods that we are evaluating for this project include two Shrinkage Methods, *Ridge Regression* and *Lasso Regression*, and two Dimension Reduction Methods, *Principal Components Regression* and *Partial Least Squares Regression.* We also have the *Ordinary Least Squares Regression* as the benchmark to compare the models.

Ordinary Least Squares Regression (OLS)

Ordinary Least Squares Regression is a method for estiamting the unknow parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses in the given dataset and those predicted by a linear function of a set of explanatory variables. In particular, the OLS regression coefficient estimates $\hat{\beta}^R$ are the values that minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 = RSS$$

Shrinkage Methods:

Shrinkage Methods are devoted to constrain or regularize the coefficients estimates, or equivalently, shrink the coefficient estiamtes towards zero when fitting a model containning all predictors. By shrinking the coefficient estimates, the model could significantly reduce the variance in predictions. The two best-known techniques for shrinking the regression coeffcients towards zero are *Ridge Regression* and *Lasso Regression*, both of which we would discuss in the following sections.

Ridge Regression (RR)

Ridge Regression is very similar to OLS Regression, except that the coefficients are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

The above equation trade of two different criteria. As with least squares, *Ridge Regression* seeks coefficint estiamtes that fit the data well, by making the RSS small. However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, called a *Shrinkage Penalty*, is small when $\beta_1, \beta_2, \ldots, \beta_n$ are close to zero, and so it has the effect of *shrinking* the estimates of β_j towards zero. The tuning parameter λ serves to control the relative impact of these two terms on the regression coeffcient estimates. When $\lambda = 0$, teh penalty term has no effect, and *Ridge Regression* will produce the least squares estimates. However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the *Ridge Regression* coefficient estiamtes will approach zero. Therefore, *Ridge Regression* will produce a different set of coefficient estiamtes, $\hat{\beta}_\lambda^R$, for each value of λ. We will select a good value for λ via cross-validation in this project.

Lasso Regression (LR):

The *Ridge Regression* discussed above has one obvious disadvantage, in that it will include all p predictors in teh final model. Its penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ will shrink all of the coeffcients towards zero, but it will not set any of them exactly to zero (unless $\lambda = 0$). Such a property can create a challenge in model interpretation in settings with the large number of variables p.

Thus, *Lasso Regression* comes in as an alternative to *Ridge Regression* that overcomes this disadvantages. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

We see that *Lasso Regression* differs from *Ridge Regression* in the penalty term. Similarly, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the l_1 penalty has the effect of forcing some of the coeffcient estiamtes to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, the lasso performs *variable selection* in this way. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression. Ee also say that the lasso yields *sparse* models, models that involve only a subset of the variables. As in ridge regression, we would select a λ via cross-validation for this project.

Dimension Reduction Methods:

This class of approaches tries to transform th epredictors and then fit a least squares model using the transformed variables. Such techniques are referred to as *dimension reduction methods*. The term comes from the fact that this approach reduces the problem of estimating the $M + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$ to the simpler problem of estimating the $M + 1$ coefficients $\theta_0, \theta_1, \ldots, \theta_M$, where $M < p$. In other words, the dimension of the problem has been reduced from $p + 1$ to $M + 1$.

All dimension reduction methods work in two steps. First, the transformed predictors $Z_1, Z_2, \ldots, Z_M$ are obtained. Second, the model is fit using these M predictors. However, the choice of $Z_1, Z_2, \ldots, Z_M$, or equivalently, the selection of the reduction methods can be achieved in different ways. In this project, we will discuss two approaches: *Principal Components* and *Partial Least Squares*.

Principal Components Regression (PCR)

Principal Components Regression relies on a similar dimension reductjion techniques as *Principal Components Analysis* (PCA). So we may well discuss PCA before PCR.

PCA is a tehcnique for reducing the dimensio of a $n \times p$ data matrix X. The *first principal component* direction of the data is that along which the observations *vary the most*. That is, if we *projected* the 100 observations onto tis line, then the resulting projected observations onto any other line would yield projected observations with lower variance. Projecting a point onto a line simply involves finding the location on the line which is closest to the point. Then the second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_1, and has the largest variace subject to this constraint. And so on, one can construct up to p distinct principal components in general.

The *Principal Components Regression* (PCR) approach involves constructing the first M principal components, $Z_1, Z_2, \ldots, Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares. The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response. In other words, we assume that *the directions in which $X_1, X_2, \ldots, X_p$ show the most variation are the directions that are associated with Y*. While this assumption is not guaranteed to be true, it often turns out to be a reasonable enough approximation to give good results.

If the assumption underlying PCR holds, then fitting a least squares model to $Z_1, Z_2, \ldots, Z_M$ will lead to better results than fitting a least squares model to $X_1, X_2, \ldots, X_p$, since most or all of th einformation in the data that relates to the response is contained in $Z_1, Z_2, \ldots, Z_M$, and by estimating only $M \ll p$ coefficients we can mitigate overfitting.

Similarly, we will choose the number of principal components, M by cross-validation for this project. Also, we perform standardization for each of the predictors, so as to ensure that all variables are on the same scale.

Partial Least Squares Regression (PLSR)

The PCR approach that we just discussed involves identifying linear combinations, or *directions*, that best represent the predictors $X_1, X_2, \ldots, X_p$. These directions are identified in an *unsupervised* way, since the response Y is not used to help determine the principal component directions. That is, the response does not *supervise* the identification of the principal components. Consequently, PCR may suffer from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

As an alternative, we may now discuss *Partial Least Squares* (PLS), a *supervised* alternative to PCR. Similarly, PLS, as a dimension reduction method, first identifies a new set of features $Z_1, Z_2, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. But unlike PCR, PLS identifies these new features in a *supervised* way, making use of the response Y in order to identify new features that not only approximate the old features well, but also that *are related to the response.* Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.