Methods

Algorithms For Prediction

Least Squares Regression:

One way to predict sales based on multiple predictors is to fit separate linear regression model for each predictor, and the other way is to extend the linear model so that it can directly accommodate multiple predictors. For a one predictor case, we can just use a simple linear regression with Balance = Beta-0 + Beta-1*Income where Beta-0 is the intercept and Beta-1 is the slope of the regreession line, and replace Income with the other predictors to get different results. For a multiple linear regression, we need more coefficients to represent the data properly. We can formulate a linear regression with all predictors via the least square criterion which is the same criterion in the simple linear regression.

Ridge Regression:

Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. Ridge regression seeks coefficients that fit the data well by making RSS small which is the same intuition as least square regression. The difference between Ridge Regression and Least Regression is that Ridge Regression adds a l2 penalty to the RSS which is small when Beta is close to zero and shrinks the estimate of Beta to zero. Ridge regression's advantage over least squares is rooted in the bias-variance trade-off.

Lasso Regression:

Lasso Regression shares the same intuition as Ridge Regression which is adding a penalty to RSS to shrink coefficients and make better prediction. The difference between Lasso and Ridge is that Lasso uses a l1 penalty rather than a l2 penalty. In this case, Lasso is able to shrink coefficients to exactly zero and would return only a subset of predictors.

Principal Components Regression:

Principal components analysis (PCA) is a common approach for deriving a low-dimensional set of features from a large set of variables. Principal Components Regression is a dimension reduction technique for regression. The principal components regression (PCR) approach involves constructing the first M principal components, Z1,. . .,ZM, and then using these components as the predictors in a linear regression model that is fit using least squares. we assume that the directions in which X1,. . . , Xp show the most variation are the directions that are associated with Y.

Partial Least Squares Regression:

PCR has a major disadvantage that there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response. Similarly, PLS is a dimension reduction method which first identifies a new set of features Z1,. . .,ZM that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. The difference between PLS and PCR is that PLS learns the new features in a supervised way which means that PLS find directions that help to explain both response and predictors.

Parameter Choosing for Algorithms

Cross Validation:

We use Cross Validation to choose best parameters that give us best prediction accuracy. Specifically we use 10-fold Cross Validation. In this case, The data set is divided into 10 subsets, and the holdout method is repeated 10 times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets form a training set. Then the average error across all 10 trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set 9 times.