High Dimensional Data

Sravani Vadlamani

Agenda

- Dimension Reduction Methods
  - Principal Component Regression
  - Partial Least Squares Regression
- High Dimensional Data
  - Issues and Considerations

What we learnt so far

- Least Squares
- Subset selection
- Ridge and Lasso Regression

- In the above, we used the original predictors X_1, X_2, ….X_p to estimate the model coefficients

Dimension Reduction Methods

- Transform the predictors and fit a least squares model using the transformed variables

- Techniques
  - Partial Least Squares Regression
  - Principal Components Regression

Partial Least Squares

- A supervised dimension reduction method
- Identifies a new set of features Z_1, Z_2, ….Z_m that are linear combinations of the original features and uses the m new features to fit a least square model
- PLS uses the response Y to identify new features that not only approximate the old features well but also that are related to the response variable
- PLS attempts to find direction that help explain both the response and the predictors
- Refer to textbook for the math

Principal Components Analysis

- An unsupervised dimension reduction method
- Principal Component Analysis
  - Used to derive a low dimensional set of features from a large set of variables
  - First principal component is the direction of the data along which the observations vary the most
  - Second principal component is a linear combination of the variables that is not correlated with the first principal component and has largest variance subject to this constraint
  - Similarly, you can construct principal components up to the total number of predictors (say p)

Principal Components Regression

- Involves constructing M principal components Z_1, Z_2, ….$Z_{m,}$ and using these components as predictors to fit a least squares linear regression model
- A small number of principal components are sufficient to explain most of the variability in the data as well as the relationship with the response
- PCR offers a significant improvement over least squares if the first few principal components sufficiently capture most of the variation in the predictors as well as the relation with the response
- Standardizing predictors prior to generating the principal components is recommended when performing PCR

High Dimensional Data

- Low dimensional data
  - Number of observations greater than the number of variables ($n > p$)
  - Regression and Classification can be used
- High dimensional data
  - Contains more features or variables than observations
  - Cannot use least squares as we get coefficient estimates that perfectly fit the data regardless of whether there truly is a relation between the response variable and the predictors
  - Cp, AIC, BIC are also not appropriate in high dimension settings as estimating σ is difficult

Regression in High Dimensions

- Forward stepwise selection, ridge, lasso and principal components regression are useful for high dimensional data as they avoid overfitting by using a less flexible fitting approach
- Regularization or shrinkage is key in high dimensional data and choice of appropriate tuning parameter is crucial for model predictive performance
- Test error increases as the number of predictors increase unless the additional features are truly associated with the response variable. This is known as the curse of dimensionality and the additional features increase the risk of overfitting

Regression in High Dimensions

- High risk of multicollinearity in high dimensional setting. Any variable can be written as a linear combination of other variables which makes it difficult to identify which variables are truly related to the outcome variable. We can never identify the best coefficients for use in the regression and hope that large coefficients are assigned to variables that are truly associated with the outcome
- Traditional model fit measures on training data should not be used to assess model performance. MSE or R^2 on an independent test set is a valid measure of model performance