THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Feature / Variable Selection

10/1/2019

Alternative Fitting - Sub-setting Data

- Often the variables in our data set need to be modified
- This includes analyzing a subset of that data
- Or identifying key elements or series of the data to use in our models

Alternative Fitting - Sub-setting Data

- In some cases our n (no. of observations) is not significantly larger than our p (no of coefficients) so ordinary least squares (OLS) regression becomes unreliable
- However if we can reduce the coefficients we can improve our model performance
- This can be done through feature selection methods

Example of featured selection

- Recent headline news: "Security Firm Says Extremely Creepy Mask Cracks iPhone X's Face ID"
  Source: NPR.org, November 13, 2017.
- What features on your face does iPhone look at?
  - Eyes? Eyeglasses? Eyebrows?
  - Nose? Pimples?
  - Lips?
  - Hair? Beard? Sideburns?
  - Cheekbone?
- https://youtu.be/i4YQRLQVixM (1:34)
- Imagine what the "match score" should be a function of.

Variable Reduction

- Examples that we are going to explore
  - Subset Selection (e.g. Forward / Backward Step-Wise Selections)
  - Dimension Reduction and Clustering (e.g. PCA, K-means)
  - Shrinkage/Regularization (e.g. Ridge and Lasso regression)
  - Decision Tree methods

Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

Subset (Manual)

- This process is fairly straight forward, we run our model and then depending on the output, we subset the data.frame using the data option inside the linear model function

- Use the variables shown to be most highly related to our predictor variable

Model Selection (Subset)

- However, we can also use internal R functions to find the smallest combination of variables that minimizes our RSS and maximizes our R^2

Model Selection (Subset)

- Best Fit – Models that considers all the different model possibilities to produce the one with the "best fit". In general, there are 2^p possibilities.

- When you have a significant number of variables this gets to be too computationally expensive

Model Selection (Subset)

- Forward Step-Wise is a solution that only adds the variable that contributes the most to the existing model
  - Example from readings: p = 20
  - No. of Models = $1+p(p+1)/2 = 211$ compared to millions in the best fit approach

- Use cross validation prediction error, AIC, BIC, adjusted R^2 as examples

Model Selection (Subset)

- Backward Step-Wise – Essentially the same as forward except it starts with the full model then eliminates predictors
- *n* must be larger than *p,* which is not true of forward selection
- You can also implement a hybrid system that forward selects but then also removes predictors that don't contribute to the model
- How do we select:
  - Cross-validate using estimates of test error such as: C_p , AIC, BIC, adjusted R^2 as examples

Model Selection (Subset)

- **C_p** – Mallow's C_p statistics estimate the average MSE. It adds an error to our test model development to account for test error results usually being higher than what is occurring. More predictors, higher the adjustment. We want this to be **low**
- **Adjusted R^2** - Unlike R^2 which always increase with model size, this penalize larger numbers of predictors, after variance explanation has plateaued. Want this to be **high**
- **Akaike Information Criterion (AIC)** – Very similar to C_p measure used in often with Logistic regression as it's a measure of maximum likelihood.
- **Bayesian Information Criterion (BIC)** – Similar to C_p and Akaike, includes a log conversion of the error adjustment and as a result places a higher penalty for more predictor variables

R Libraries

- leaps::regsubsets( )
  - ✓ Easy to use
  - ✓ Different parameters and options, such as BIC/R^2/adjR^2/C_p , etc
  - ✓ Nice visual presentation of results
  - ✓ Gives good visual results.
  - ✗ Does not handle logistic regression
- bestglm::bestglm( )
  - ✓ Can handle logistic regression
  - ✓ Additional criteria available such as AIC/CV (cross-validation) etc
  - ✗ Must clean dataframe with target variable renamed as *y*, and all other unused variables removed form dataframe
  - ✗ Lack intuitive visual presentation of results
  - ✗ Limits on max number of features