Topic # 07: Data Analytics

Course Review

Instructor: Prof. Arnab Bisi, Ph.D.

Johns Hopkins Carey Business School

Multiple Methods

- Supervised learning: predict output
  - Classification: predict discrete output
  - Regression: predict continuous output
- Unsupervised learning: no output
  - Clustering: study similarity
  - Trees: splitting
  - Evaluate "homogeneity" within each branch/group
  - Fitting multiple trees often works better (forests)

Linear Regression

- Linear regression is a tool for predicting a quantitative response
- Parameters can be found by the *least squares approach*
- Given estimates we make predictions using a formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

- Key elements: p-values, t-tests, R^2, confidence intervals, coefficients

Model Selection

- Subset selection
  - Best subset selection
  - Forward stepwise selection
  - Backward stepwise selection
- Key elements: AIC, BIC, over-fitting

Logistic “Regression”

- Tool for predicting a quantitative (categorical) response

$$q = \Pr(y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}.$$

- Parameters estimated by method of maximum likelihood
- Key elements: Deviance, confusion tables, error rates

Clustering

- K-means clustering: partitioning observations into a pre-specified number of groups
    - Multiple starts due to randomness of initial assignment
- Hierarchical clustering: “bottom up” approach that creates a tree-like representation of a data set
- Key elements: Euclidian distance, sum of squares

Decision Trees

- Can be used for regression (means) or classifications (modes)
  - Iteratively split variables into groups
  - Split where maximally predictive
  - Evaluate “homogeneity” within each branch
  - Fitting multiple trees often works best
- Key elements: splitting rules, how to follow logic of tree

Odds and Ends

- Validation
  - Training vs. Testing data
  - Cross Validation
- KNN – use similarity of “neighbors” to make predictions or assign to groups
- Cases: applications of ideas to real data

Dimension Reduction

- Principal Component Analysis is used for dimension reduction
  - A large set of correlated X variables
  - Principal Component summarizes this set with small number of representative variables
  - Principal Component Regression uses principal components as predictors in the regression model
  - Fitting multiple trees often works best
- Key elements: proportion of variance explained

Learn More About Data Analytics

- Get more info
  - OpenIntro: http://www.openintro.org/
  - Coursera: https://www.coursera.org/
  - edX: https://www.edx.org/
  - Big data university http://bigdatauniversity.com/
- More about R:
  - http://www.r-tutor.com/
  - http://www.revolutionanalytics.com/

Zoom Presentations

- Each group has about 12 minutes for video
- 3-4 present in each group

Questions, Comments?

Let's move to the presentations.