Kaggle cheatsheet

Toh Wei Zhong

29 December 2015

Exploratory data analysis and data preprocessing

EDA and viz

- Specific to response variable
  - Check distribution of categorical response using table()
  - Check distribution of numerical responses using hist() and boxplot.stats()
  - Use make.names() for valid class labels for caret::train()
- Categorical variables
  - Look at number of unique values per variable
- Numerical variables
  - Look at distributions
  - Look at pairwise correlations
  - Build GLM, look at VIF
- Temporal components
  - Check seasonality (of response variable)
  - Check for breaks in data
- ML-based visualizations
  - Visualise using PCA
  - Visualise by clustering on samples (e.g kmeans)
  - Visualize by clustering on variables (e.g. hiearchical clustering)
- Also check for overlap between training and testing sets (both variables and samples)

Dealing with missing values and outliers

- Omission
  - Omit sample
  - Omit variable
- Simple imputation
  - Impute by measures of central tendencies (mean, median, mode)
  - Impute by an arbitrary number (when variable is physically irrelevant to the sample)
  - Impute by extreme values (when truncation is an issue)
- Imputation by prediction
  - Use DMwR::knnImputation for kNN imputation on numerical variables
  - Use decision tree imputation (used in SAS EM)

Dealing with categorical variables

- Convert to numerial variables
  - Compute frequency (probability) using STANDARDWORKFLOW::Cate2Prob()
  - Compute chi-sq contributions
  - Convert to one-hot encoding

Feature engineering

- Less categories (categorical)
  - Consider set membership, collapse classes
- More categories (categorical)
  - Look for Simpson's Reversals (context mining in Redhyte)
- Mathematical operations on multiple variables (numerical)
  - Simulate statistical interactions using +, -, / (%), *
  - Bin using <, >, == (binary)
- Enhancing signal of a single variable (numerical)
  - Compute x^2, x^3, etc.
- ML-based
  - Include cluster membership from a clustering algorithm as a variable
  - Include PC scores from PCA as a variable

Specific to response variable

- Dealing with imbalance (categorical)
  - Use caret::upSample() or caret::downSample()
  - Use DMwR::SMOTE()
  - Collapse classes within the response variable, then do modeling within class
- Skewed distribution (numerical)
  - Transform using Box-Cox, log(x + 1), if normality is desired
  - Discretize (mean, median, quantile), then do modeling within class

---

Predictive Modeling

Miscellaneous

- For large datasets, use a smaller subset for quick EDA and visualizations

To add:

- which models need scaling
- feature selection (eg. hypothesis testing)
- model strategies (one vs many, group vs group etc)
- meta algorithms
- list of cv methods
- stacking
- large p small n, large n small p (ARM)