Feature Engineering and Selection

Exploratory Visualizations

- □ Box Plots, Violin Plots and Histograms
  - □ Understand data distribution
  - □ Augment visualizations through faceting, colours, and shapes
- □ Scatter Plots
- □ Heatmaps
  - □ Categorize predictors
  - □ Form grid, and fill by variables
- □ Correlation Matrix Plot
- □ Line Plots
  - □ Mainly for time-series
  - □ Check trends in relation to time and other variables
- □ Condense many dimensions in 2D plots
  - □ Principal Component Analysis (PCA) – Represent variables in a lower dimensional form (Unsupervised)
  - □ Partial Least Squares (PLS) – Uses label to max inter-class variance (Supervised)
  - □ Multi-Dimensional Scaling (MDS) – Different computation of space (Euclidian distance between rows)

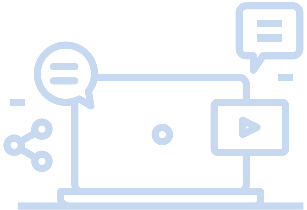
Encoding Categorical Predictors (Qualitative)

- □ Create Dummy Data (Indicator Variables)
  - □ C – 1 Dummy Data (last one can be inferred) – One Hot Encoding (Binary)
  - □ Multiple Predictors – (map keys to hash values) – Feature Hashing (Signed) – (-1, 0, 1)
  - □ Likelihood Encoding – (use statistic such as mean / median to represent factor level)
  - □ For classification – logistic regression p/(1-p) - use Bayesian method for shrinkage
- □ Encoding for Ordered Data
  - □ Treat predictors as unordered factors
  - □ Translate to a single set of numeric scores based on context
  - □ Comparison of mean between more than two different levels of independent var (Polynomial Features)
- □ Features for Text Data
  - □ Removing commonly used “stop-words” such as “is”,” the”, “and” etc.
  - □ Stemming words – such as singular & plural versions are represented with a single entity

Engineering Numeric Predictors

- □ 1 to 1 Transformations
  - □ Box & Cox – (Scaling skewed data)
  - □ Centering predictor (training set average subtracted from predictor individual values)
  - □ Data Smoothing – (Ex. Smoothing Splines – for time / sequence based data)

- □ 1 to Many Transformations
  - □ Basis Expansion – Original column augmented by two features with squared & cubed versions of original
  - □ Polynomial Spline – Connecting knots to cubic functions; Use grid search to determine knots or GCV
  - □ Smoothing Spline Methodology, Multivariate Adaptive Regression Spline (MARS), Hinge Function
  - □ Discretization – Translating quantitative variable to a set of 2 or more categories
  - □ Segmented Regression Models – Separate linear trends in distinct sections
  - □ Rectified Linear Unit



□ Many to Many Transformations

- □ PCA – Find Linear Combinations of original predictors; Use Biplots; Use when Linearly Correlated
- □ kernel PCA – Enables PCA to expand dimension of predictor space; For non-Linear Correlations
- □ ICA – Broader set of trends than PCA, and keeping components independent (non-Gaussianity)
- □ NNMF – Finds best set of coef closer to original data (non-negative); Used for text data
- □ PLS – Supervised version of PCA; Finds linear functions of predictors with optimal covariance
- □ Autoencoders- Make set of nonlinear mapping (original predictors and artificial features) Non Labeled
- □ Global Contrast Normalization – For Image Analysis; Data projected to multidimensional sphere

Detecting Interaction Effects

□ Recognize Types of Interactions

- □ Additive – B3 is not significantly different than 0; interaction x1, x2 not explaining variation
- □ Antagonistic – Coefficient is meaningfully negative; x1, x2 alone also affect response
- □ Positive – Coefficient is meaningfully negative; x1, x2 alone also affect response
- □ Atypical – Coefficient is significantly different than 0; but neither x1, x2 affect response

□ Strategy for Spotting Interactions

- □ Hierarchy Principle – For pairwise interactions; Effect Sparsity; Heredity Principle (Genetic Heredity)
- □ Brute Force – Simple Pairwise screening for small datasets; For false positives: Resampling
- □ Penalized Regression – Case when there are more predictors than samples; minimize lambda
- □ When enumeration is completely impossible: two stage modelling; tree-based method

Handling Missing Data

□ Recognize Types of Interactions

- □ Additive – B3 is not significantly different than 0; interaction x1, x2 not explaining variation
- □ Antagonistic – Coefficient is meaningfully negative; x1, x2 alone also affect response
- □ Positive – Coefficient is meaningfully negative; x1, x2 alone also affect response
- □ Atypical – Coefficient is significantly different than 0; but neither x1, x2 affect response