Data Analysis Notes

Jeffrey Uslan

Sunday, July 26, 2015

Statistical Inference

Interval Estimate

1. Compare point estimate to known distribution (normal or t). x+-z(or t)*sd(distribution)

Hypothesis Testing

1. z=(x-mu)/s_x, s_x=sd/sqrt(n)
2. p-value is the probability of that z-score within the distribution

Distribution comparison

1. T-Test

- Test if the mean of two distributions are equal

2. ANOVA

- Test if the mean of two or more distributions are equal

3. Chi-square

- Test if the distribution of categoricals in two distributions are equal

Exploration

General behavior basic

- Summary Stats
- Histograms
- Box plots

Unsupervised Learning

- Hierarchical clustering
- PCA [prcomp]
- K-means clustering

Trend finding

- Auto-Correlation Function
- Cross-correlation Function
- Plot first difference
- Moving Averages
- Kernel smoothing (weight moving averages)
- k-Nearest neighbors
- Locally weighted estimates
- Spline (Window regression)
- Fourier Analysis/ Periodogram
- must not have linear trend or be de-trended

Cleaning

- Denoising
- SVD denoising
- Moving window filter
- Outlier Detection
- Mahalanobis distance
- Single Variable standard deviations

Imputation

- Single point averaging
- k-means

Modeling

Model Choice

- Quantitative Prediction
- Linear regression [lm] - Assumptions:
  - Linear relationship.
  - Residual normality.
  - No or little multicollinearity. Correlated variables.
  - No autocorrelation. Values are not time dependant.
  - Homoscedasticity. Variance of residuals is unrelated to index.
- Polynomial regression
- Qualitative Classification
- Logistic Regression. *1/(1+eˆ(-t))
- K-nearest neighbors classifier
- Naive Bayes [nb]
- Linear Discriminant Analysis [lda]

- Decision Trees [rpart], Random Forests [rf]
- SVM
  - Maximal Margin Classifier
  - exact hyperplane separation
  - support vector classifier
  - inexact hyperplane separation
  - support vector machine
  - non-linear “hyperplane” separation
- Time Series
- Seasonal Decomposition
- Autoregression
- step function fits
- Trig fits

Model Correction

- Subset Selection
  - Stepwise
  - Correlation cutoff
  - fractal dimension impact
- Shrinkage
  - Ridge regularization (L1)
  - Suppresses coefficient
  - Lasso (L2)
  - Allows suppression to zero
- Dimension reduction
  - principal components
- Bootstrapping ##Model Selection and Validation
- Cross Validation
  - Leave one out
  - k-folds
- Quantitative Prediction
  - MSE
- Qualitative Classification
  - ROC Curves