Visual Data Mining & Machine Learning on Viya

Lorne Rothman
Agenda

• Machine Learning
• Viya and VDMML
• Demonstration
Machine Learning
Machine Learning

There is a little Hype

Machine Learning is at the “Peak of Inflated Expectations”

Right next to:
- Cognitive Experts Advisors
- Connected home
- Software Defined Security
- Autonomous vehicles
“ML”—What is it?

Machine Learning Algorithms

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders

Deep Learning

- Random Forest
- Gradient Boosting Machines (GBM)
- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (Blending)
- Gradient Boosted Regression Trees (GBRT)
- Radial Basis Function Network (RBFN)

Neural Networks

- Perceptron
- Back-Propagation
- Hopfield Network
- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least Angle Regression (LARS)
- Cubist
- One Rule (OneR)
- Zero Rule (ZeroR)
- Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
- Linear Regression
- Ordinary Least Squares Regression (OLSR)
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)
- Logistic Regression

Regularization

- Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bayesian Network (BN)

Bayesian

Decision Tree

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5
- C5.0
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- Conditional Decision Trees
- M5

Dimensionality Reduction

- Principal Component Analysis (PCA)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Principal Component Regression (PCR)
- Partial Least Squares Discriminant Analysis
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Flexible Discriminant Analysis (FDA)
- Linear Discriminant Analysis (LDA)

Instance Based

- k-Nearest Neighbour (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Clustering

- k-Means
- k-Medians
- Expectation Maximization
- Hierarchical Clustering
Statistics, Data Mining, Machine Learning

Machine Learning
-- “A field of study that gives computers the ability to learn without being explicitly programmed.” --
Arthur Samuel, 1959

Classical Statistics

*Humans know best*

- Theoretically driven
- Pre-specified model form
- Emphasis on causal
- Insights driven
- Experimentation & hypothesis testing
- P-values & distributional assumptions
- Model interpretability
- “Randomized Control Designs as Gold Standard”
- Few variables, smaller data

Data Mining

- T-tests, ANOVA, GLMs
- MIXED models
- No Variable Selection
- Stepwise methods
- Data partitioning
- Tune models on validation data
- LARS, LASSO
- Decision Trees
- Bagging & Boosting
- Ensemble Models

Machine Learning

*Machines are smarter*

- Limited theory & human intervention
- Data driven modeling
- Machine driven automation
- Emphasis on prediction
- Fewer assumptions
- Limited hypothesis testing and interpretation
- Limited interpretability
- Complex non-linear pattern
- Many variables, bigger data

- Random Forests
- Gradient Boosting
- Neural Networks
- Support Vector Machines
- Hyper-Parameter Auto-Tuning
- Deep Learning Neural Networks
Auto-Tuning Machine Learning Algorithms

Automation

• Auto-search to find optimal combinations of parameters that maximize model performance using validation or cross-validation

• Decision Trees, Random Forests, Gradient Boosting, Neural Networks, Support Vector Machines, Factorized Machine Learning

• Tuning parameters include Tree Depth, Lasso & Ridge Regularization, Number of Trees, Learning Rate, Hidden Units & Hidden Layers in a Neural Net, C penalty & polynomial degree in an SVM, etc.

• AuROC, GINI, KS, Misclassification, Average Squared Error assessments are supported, and more...

“Try them ALL, pick the best one, apply it, generate score code, and send me a report when you’re done.”
SAS VDMML on Viya
SAS Viya Platform
Accelerating Analytic Innovation

Cloud Analytic Services (CAS)

In-Memory Engine

Microservices

ANY DATA

Real time

Cloud

Hadoop

Database

ON PREMISE

SAS Cloud

Public Cloud

Solutions

Visual Experience

Programming Experience

Applications (API)
SAS Viya Interfaces

The Platform

SAS Viya

Controller

Workers

APIs

SAS

python

R
proc print data = hmeq (obs = 10);
run;

df = s.CASTable('hmeq')
df.head(10)

df <- defCasTable(s, 'hmeq')
head(df, 10)

[table.fetch]
  table.name = "hmeq"
  from = 1 to = 10
**VISUAL STATISTICS**

**Model**
- GENSELECT: Generalized Linear Models
- REGSELECT: Linear Regression
- LOGSELECT: Logistic Regression
- NLMOD: Nonlinear Regression
- PLSMOD: Partial Least Squares
- QTRSELECT: Quantile Regression
- TREESPLIT: Decision Trees*

**Explore & Modify**
- KCLUS: K-means and K-modes Clustering
- PCA: Principal Components Analysis
- VARIMPUTE: Impute missing values
- VARREDUCE: Reduce variables with Supervised & Unsupervised methods.
- CARDINALITY: Explore number of levels in variables
- BINNING: Bin interval variables
- PARTITION: Sample (random & stratified), oversample, partition data

**Assess & Score**
- ASSESS: Assess & compare models
- Score data with Data Step code

**VISUAL DATA MINING & MACHINE LEARNING LEARNING**

**Model**
- NNET: Neural Networks*
- SVMACHINE: Support Vector Machines *
- GRADBOOST: Gradient Boosting Machines*
- FOREST: Random Forests*
- FACTMAC: Factorized Machine Learning for predictive models and Recommender systems on sparse data*

**Explore & Modify**
- SVDD: Support Vector Data Description. Useful for outlier detection and rare class (e.g. fraud) problems
- MWPCA: Explore changes in principal components through time
- RPCA: Robust PCA. Useful for image recognition tasks
- NETWORK: Network analyses

**Text Mine**
- TEXTMINE: Language processing, SVD, topic discovery
- BOOLRULE: Extract Simple Boolean Rules from transactional text data.
- Score
  - Score data using Data Step and Analytic Store files
  - TMSCORE: Score text data.

* - Supports Auto-tuning
The V in DMML stands for Visual!
Thank You