MACHINE LEARNING QUICK REFERENCE: BEST PRACTICES

Topic	Common Challenges	Suggested Best Practice
Data Preparation		
Data collection	• Biased data • Incomplete data • The curse of dimensionality • Sparsity	• Take time to understand the business problem and its context • Enrich the data • Dimension-reduction techniques • Change representation of data (e.g. COO)
"Untidy" data	• Value ranges as columns • Multiple variables in the same column • Variables in both rows and columns	Restructure the data to be "tidy" by using the melt and cast process
Outliers	• Out-of-range numeric values and unknown categorical values in score data • Undue influence on squared loss functions (e.g. regression, GBM, and *k*-means)	• Robust methods (e.g. Huber loss function) • Discretization (binning) • Winsorizing
Sparse target variables	• Low primary event occurrence rate • Overwhelming preponderance of zero or missing values in target	• Proportional oversampling • Inverse prior probabilities • Mixture models
Variables of disparate magnitudes	• Misleading variable importance • Distance measure imbalance • Gradient dominance	Standardization
High-cardinality variables	• Overfitting • Unknown categorical values in holdout data	• Discretization (binning) • Weight of evidence • Leave-one-out event rate
Missing data	• Information loss • Bias	• Discretization (binning) • Imputation • Tree-based modeling techniques
Strong multicollinearity	Unstable parameter estimates	• Regularization • Dimension reduction
Training		
Overfitting	High-variance and low-bias models that fail to generalize well	• Regularization • Noise injection • Partitioning or cross validation
Hyperparameter tuning	Combinatorial explosion of hyper-parameters in conventional algorithms (e.g. deep neural networks, Super Learners)	• Local search optimization, including genetic algorithms • Grid search, random search
Ensemble models	• Single models that fail to provide adequate accuracy • High-variance and low-bias models that fail to generalize well	• Established ensemble methods (e.g. bagging, boosting, stacking) • Custom or manual combinations of predictions
Model Interpretation	Large number of parameters, rules, or other complexity obscures model interpretation	• Variable selection by regularization (e.g. L1) • Surrogate models • Partial dependency plots, variable importance measures
Computational resource exploitation	• Single-threaded algorithm implementations • Heavy reliance on interpreted languages	• Train many single-threaded models in parallel • Hardware acceleration (e.g. SSD, GPU) • Low-level, native libraries • Distributed computing, when appropriate
Deployment		
Model deployment	Trained model logic must be transferred from a development environment to an operational computing system to assist in organizational decision making processes	• Portable scoring code or scoring executables • In-database scoring • Web service scoring
Model decay	• Business problem or market conditions have changed since the model was created • New observations fall outside domain of training data	• Monitor models for decreasing accuracy • Update/retrain models regularly • Champion-challenger tests • Online updates

MACHINE LEARNING QUICK REFERENCE: RESOURCES

Publications

Statistical Modeling, The Two Cultures – Leo Breiman

- http://projecteuclid.org/euclid.ss/1009213726

Fifty Years of Data Science – David Donoho

- http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Pattern Recognition and Machine Learning – Christopher Bishop

- https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/bishop-regression.pdf

Machine Learning with SAS Enterprise Miner – SAS White Paper

- http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/machine-learning-with-sas-enterprise-miner-107521.pdf

An Overview of Machine Learning with SAS® Enterprise Miner™ - 2014 SGF Paper (SAS313-2014)

- http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf

Posts

An Introduction to Machine Learning – Patrick Hall on sas.com

- http://blogs.sas.com/content/sascom/2015/08/11/an-introduction-to-machine-learning/

7 Common Mistakes of Machine Learning – Cheng-Tao Chu on KDNuggets

- http://www.kdnuggets.com/2015/03/machine-learning-data-science-common-mistakes.html

How to build a deep neural network in SAS Enterprise Miner – Answer on SAS Data Mining community

- https://communities.sas.com/t5/SAS-Communities-Library/How-to-build-a-deep-learning-model-in-SAS-Enterprise-Miner/ta-p/231190

Repos

A curated list of awesome Machine Learning frameworks, libraries and software

- github.com/josephmisiti/awesome-machine-learning

Benchmark tests/results for open source implementations of the top machine learning algorithms

- github.com/szilard/benchm-ml

Code/materials for integrating SAS with popular open source analytics technologies like Python and R.

- github.com/sassoftware/enlighten-integration

Quick reference tables for machine learning best practices and algorithm usage

- github.com/sassoftware/enlighten-apply/tree/master/ML_tables

Library of SAS Enterprise Miner process flow diagrams to help you learn by example

- github.com/sassoftware/dm-flow