MACHINE LEARNING QUICK REFERENCE: ALGORITHMS - 1

Algorithm Type	Common Usage	Suggested Usage	Suggested Scale	Interpretability	Common Concerns
Penalized Regression	• Supervised regression • Supervised classification	• Modeling linear or linearly separable phenomena • Manually specifying nonlinear and explicit interaction terms • Well suited for ***N*** << ***p***	Small to large data sets	High	• Missing values • Outliers • Standardization • Parameter tuning
Naïve Bayes	Supervised classification	• Modeling linearly separable phenomena in large data sets • Well-suited for extremely large data sets where complex methods are intractable	Small to extremely large data sets	Moderate	• Strong linear independence assumption • Infrequent categorical levels
Decision Trees	• Supervised regression • Supervised classification	• Modeling nonlinear and nonlinearly separable phenomena in large, dirty data • Interactions considered automatically, but implicitly • Missing values and outliers in input variables handled automatically in many implementations • Decision tree ensembles, e.g. random forests and gradient boosting, can increase prediction accuracy and decrease overfitting, but also decrease scalability and interpretability	Medium to large data sets	Moderate	• Instability with small training data sets • Gradient boosting can be unstable with noise or outliers • Overfitting • Parameter tuning
***k*-Nearest Neighbors (kNN)**	• Supervised regression • Supervised classification	• Modeling nonlinearly separable phenomena • Can be used to match the accuracy of more sophisticated techniques, but with fewer tuning parameters	Small to medium data sets	Low	• Missing values • Overfitting • Outliers • Standardization • Curse of dimensionality
Support Vector Machines (SVM)	• Supervised regression • Supervised classification • Anomaly detection	• Modeling linear or linearly separable phenomena by using linear kernels • Modeling nonlinear or nonlinearly separable phenomena by using nonlinear kernels • Anomaly detection with one-class SVM (OSVM)	• Small to large data sets for linear kernels • Small to medium data sets for nonlinear kernels	Low	• Missing values • Overfitting • Outliers • Standardization • Parameter tuning • Accuracy versus deep neural networks depends on choice of nonlinear kernel; Gaussian and polynomial often less accurate
Artificial Neural Networks (ANN)	• Supervised regression • Supervised classification • Unsupervised clustering • Unsupervised feature extraction • Anomaly detection	• Modeling nonlinear and nonlinearly separable phenomena • Deep neural networks (e.g. deep learning) are well suited for state-of-the-art pattern recognition in images, videos, and sound • All interactions considered in fully connected, multilayer topologies • Nonlinear feature extraction with autoencoder and restricted Boltzmann machine (RBM) networks • Anomaly detection with autoencoder networks • Clustering and visualization with self-organizing maps (SOMs)	• Usually small to medium data sets • Stochastic gradient descent (SGD) optimization drastically increases scalability	Low	• Missing values • Overfitting • Outliers • Standardization • Parameter tuning

MACHINE LEARNING QUICK REFERENCE: ALGORITHMS - 2

Algorithm Type	Common Usage	Suggested Usage	Suggested Scale	Interpretability	Common Concerns
Association Rules	• Supervised rule building • Unsupervised rule building	Building sets of complex rules by using the co-occurrence of items or events in transactional data sets	Medium to large transactional data sets	Moderate	• Instability with small training data • Overfitting • Parameter tuning
***k*-Means**	Unsupervised clustering	• Creating a known a priori number of spherical, disjoint, equally sized clusters • *k*-modes method can be used for categorical data • *k*-prototypes method can be used for mixed data	Small to large data sets	Moderate	• Missing values • Outliers • Standardization • Correct number of clusters is often unknown • Highly sensitive to initialization • Curse of dimensionality
Hierarchical Clustering	Unsupervised clustering	Creating a known a priori number of nonspherical, disjoint, or overlapping clusters of different sizes	Small data sets	Moderate	• Missing values • Standardization • Correct number of clusters is often unknown • Curse of dimensionality
Spectral Clustering	Unsupervised clustering	Creating a data-dependent number of arbitrarily-shaped, disjoint, or overlapping clusters of different sizes	Small data sets	Moderate	• Missing values • Standardization • Parameter tuning • Curse of dimensionality
Principal Components Analysis (PCA)	Unsupervised feature extraction	• Extracting a data-dependent number of linear, orthogonal features, where $N >> p$ • Extracted features can be rotated to increase interpretability, but orthogonality is usually lost • Singular value decomposition (SVD) is often used instead of PCA on wide or sparse data • Sparse PCA can be used to create more interpretable features, but orthogonality is lost • Kernel PCA can be used to extract nonlinear features	• Small to large data sets for traditional PCA and SVD • Small to medium data sets for sparse PCA and kernel PCA	Generally low, but higher for sparse PCA or rotated solutions	• Missing values • Outliers
Nonnegative Matrix Factorization (NMF)	Unsupervised feature extraction	Extracting a known a priori number of interpretable, linear, oblique, nonnegative features	Small to large data sets	High	• Missing values • Outliers • Standardization • Correct number of features is often unknown • Presence of negative values
Random Projections	Unsupervised feature extraction	Extracting a data-dependent number of linear, uninterpretable, randomly-oriented features of equal importance	Medium to extremely large data sets	Low	Missing values
Factorization Machines	• Supervised regression and classification • Unsupervised feature extraction	• Extracting a known a priori number of uninterpretable, oblique features from sparse or transactional data sets • Can automatically account for variable interactions • Creating models from a large number of sparse features; can outperform SVM for sparse data	Medium to extremely large sparse or transactional data sets	Moderate	• Missing values • Outliers • Standardization • Correct number of features is often unknown • Less well suited for dense data