Machine Learning Algorithm Cheat Sheet

September 09, 2014

Here is a cheat sheet that shows which algorithms perform best at which tasks.

Algorithm	Pros	Cons	Good at
Linear regression	- Very fast (runs in constant time) - Easy to understand the model - Less prone to overfitting	- Unable to model complex relationships -Unable to capture nonlinear relationships without first transforming the inputs	- The first look at a dataset - Numerical data with lots of features
Decision trees	- Fast - Robust to noise and missing values - Accurate	- Complex trees are hard to interpret - Duplication within the same sub-tree is possible	- Star classification - Medical diagnosis - Credit risk analysis
Neural networks	- Extremely powerful - Can model even very complex relationships - No need to understand the underlying data – Almost works by "magic"	- Prone to overfitting - Long training time - Requires significant computing power for large datasets - Model is essentially unreadable	- Images - Video - "Human-intelligence" type tasks like driving or flying - Robotics
Support Vector Machines	- Can model complex, nonlinear relationships - Robust to noise (because they maximize margins)	- Need to select a good kernel function - Model parameters are difficult to interpret - Sometimes numerical stability problems - Requires significant memory and processing power	- Classifying proteins - Text classification - Image classification - Handwriting recognition
K-Nearest Neighbors	- Simple - Powerful - No training involved ("lazy") - Naturally handles multiclass classification and regression	- Expensive and slow to predict new instances - Must define a meaningful distance function - Performs poorly on high-dimensionality datasets	- Low-dimensional datasets - Computer security: intrusion detection - Fault detection in semiconducter manufacturing - Video content retrieval - Gene expression - Protein-protein interaction