Introduction to Machine Learning Applications

Spring 2023

Model overview

Minor Gordon

gordom6@rpi.edu

Rensselaer

“A model is a simplified representation of reality created to serve a purpose.” - Provost & Fawcett

Machine Learning

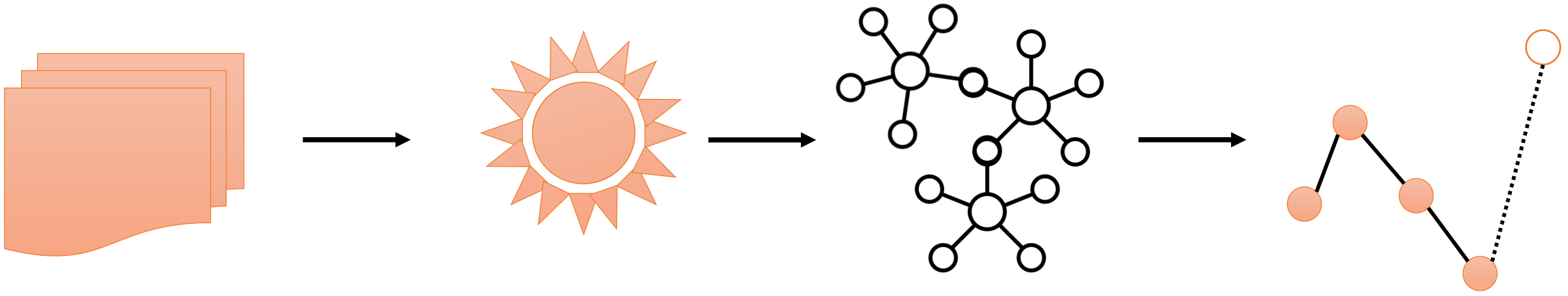
According to Tom Mitchell (1998):

Machine Learning is the study of algorithms that

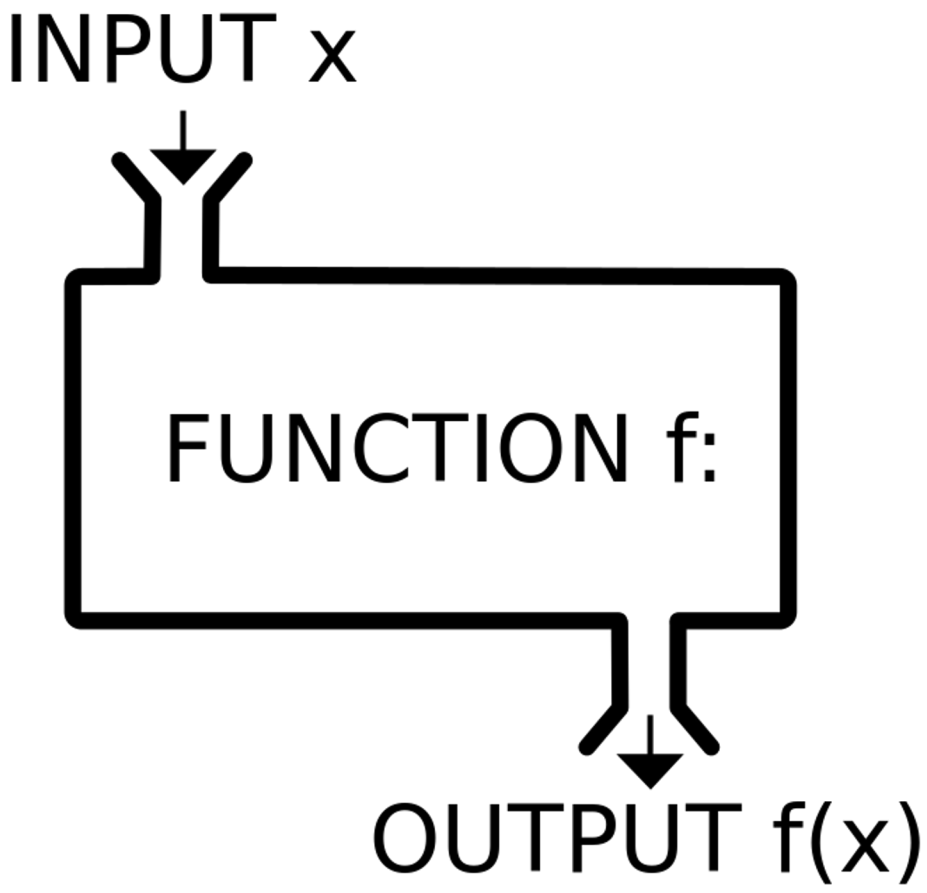
- improve their performance P
- at some task T
- with experience E

Well-defined learning task: <P, T, E>

A standard learning pipeline



Independent or Explanatory Variables

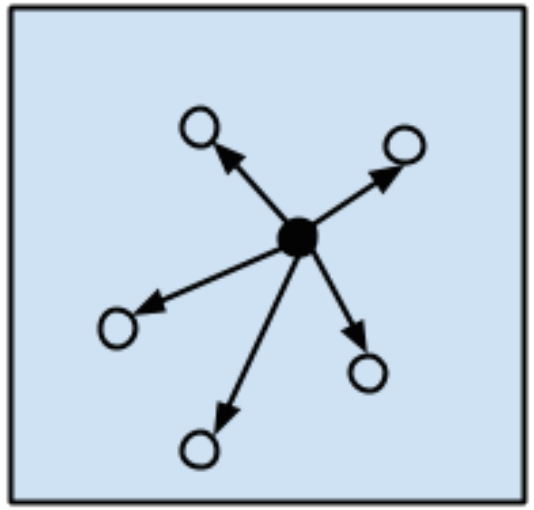


Target or Dependent Variable

Nearly impossible to find an “exhaustive” list of models

Instance-based algorithms

- This model is a decision problem with instances of training data that are deemed important or required to the model.
- Focus is put on the representation of the stored instances and similarity measures used between instances.
- Popular instance-based algorithms:
  - K-Nearest Neighbor (KNN)
  - Support Vector Machines (SVM)
  - Learning Vector Quantization
  - Self-Organizing Maps



Instance-based Algorithms

Model-based learning

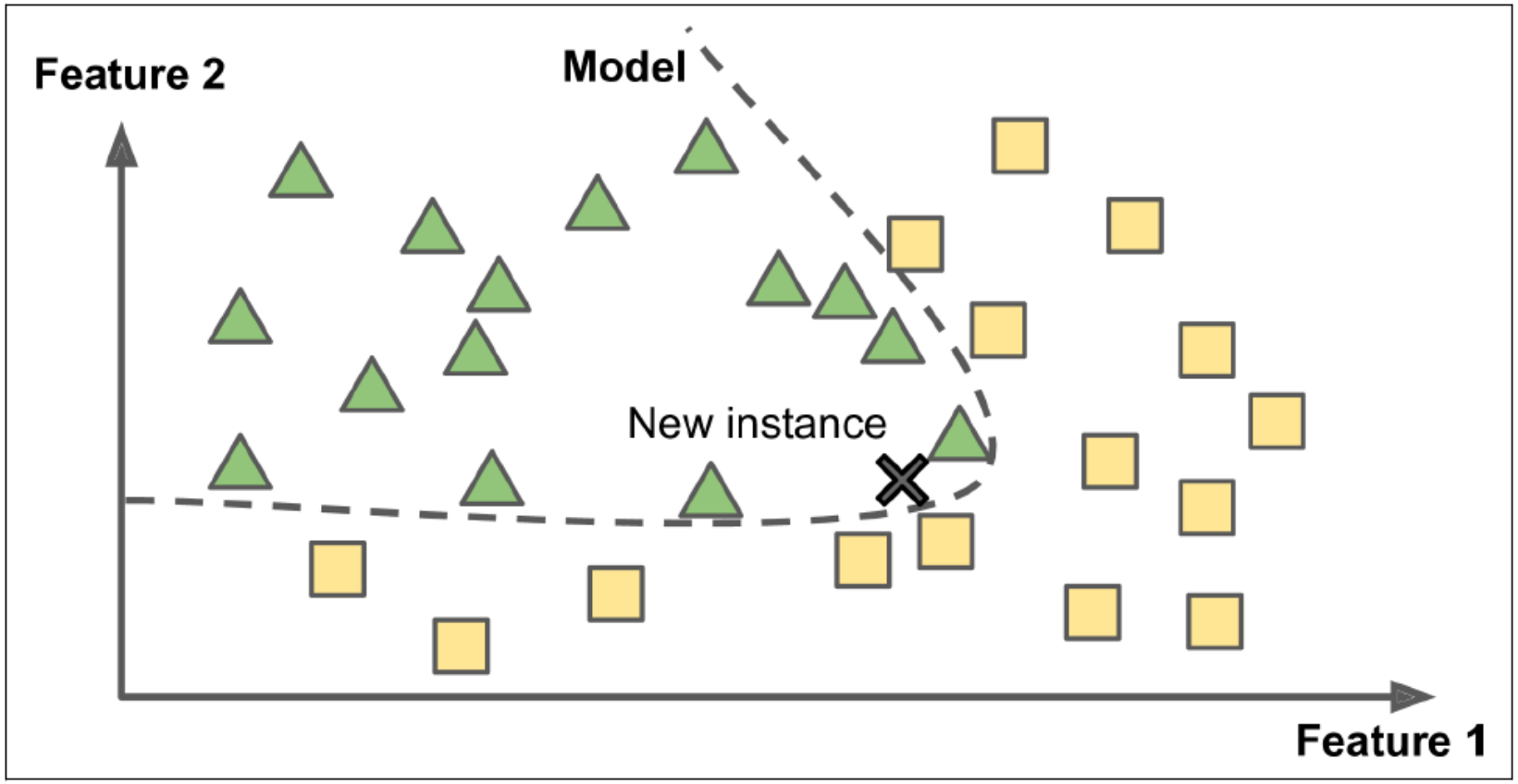


Figure 1-16. Model-based learning

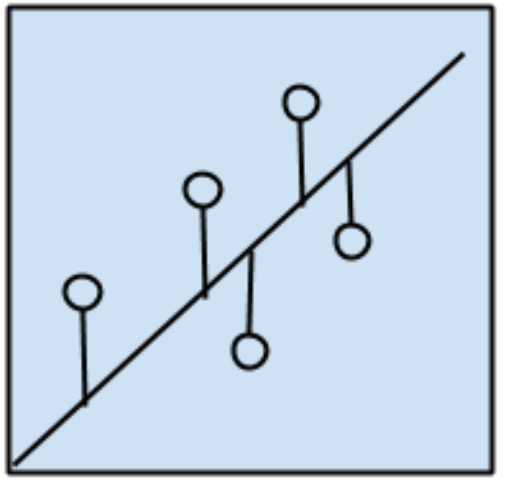
Classification techniques

- Base classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines

- Ensemble classifiers
  - Boosting, Bagging, Random Forests

Regression algorithms

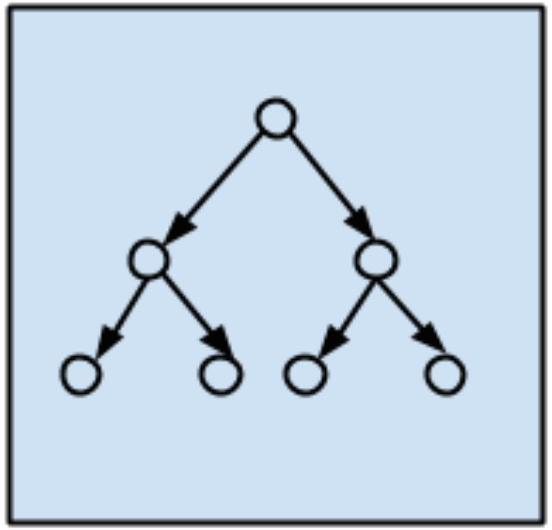
- Modeling the relationship between variables that are iteratively refined using a measure of error.
- Popular regression algorithms:
  - Ordinary least squares regression
  - Linear regression
  - Logistic regression
  - Multivariate adaptive regression splines



Regression Algorithms

Decision Tree-based algorithms

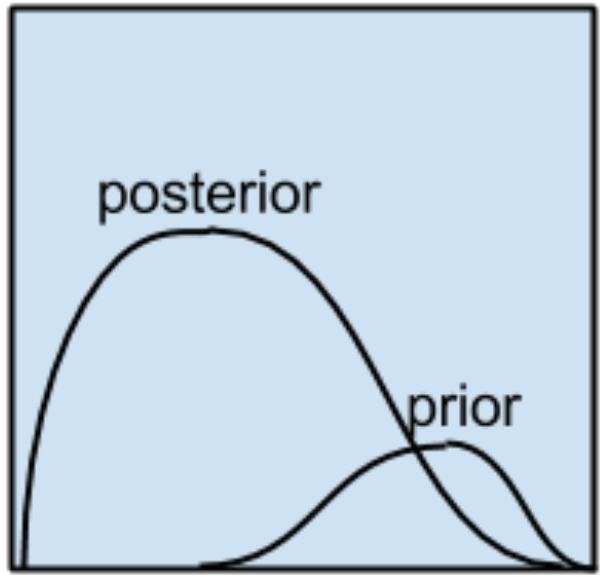
- These methods construct a model of decisions based on the actual values of attributes in the data.
- These decisions built are in the form of a tree.
- Most popular algorithms are:
  - Classification and Regression Tree
  - Conditional Decision Trees
  - ID3
  - C4.5 and C5.0



Decision Tree Algorithms

Bayesian algorithms

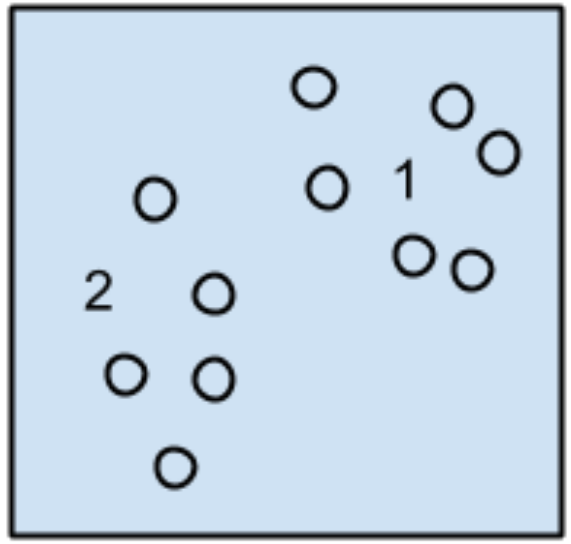
- Bayesian methods explicitly apply Bayes Theorem for problems such as classification and regression.
- Most popular algorithms are:
  - Naïve Bayes
  - Gaussian Naïve bayes
  - Bayesian network
  - Bayesian belief network



Bayesian Algorithms

Clustering algorithms

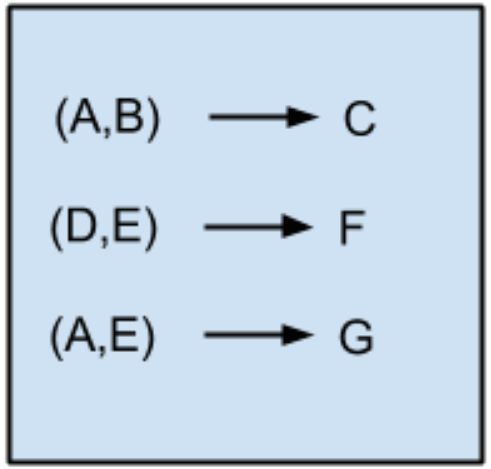
- These algorithms utilize the inherent structures in the data to organize them into various groups.
- Main goal is to find clusters that have high intra similarity and high inter similarity distances.
- Most popular clustering algorithms are:
  - K-Means
  - K-Medoids
  - Expectation Maximization
  - Hierarchical Clustering



Clustering Algorithms

Association Rule learning algorithms

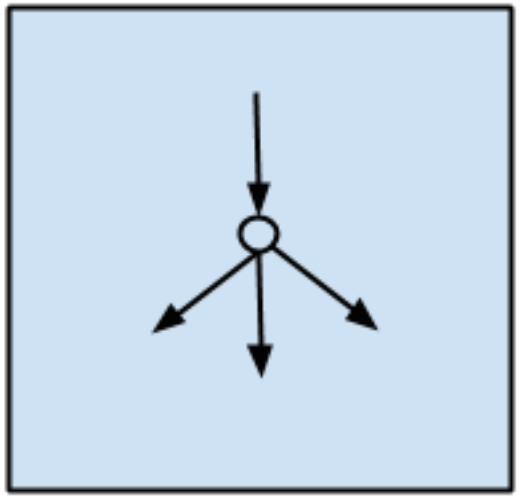
- These methods extract rules that best explain the observed relationships between variables in the data
- Most popular algorithms are:
  - Apriori
  - Eclat
  - FP-growth
  - …



Association Rule
Learning Algorithms

Artificial Neural Network algorithms

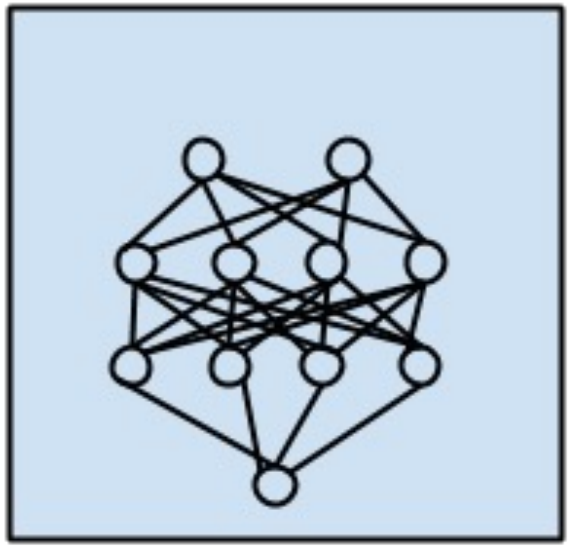
- Models that are inspired by the structure and function of biological neural networks.
- Popular algorithms:
  - Perceptron
  - Multilayer perceptron
  - Backpropagation



Artificial Neural Network Algorithms

Deep Learning algorithms

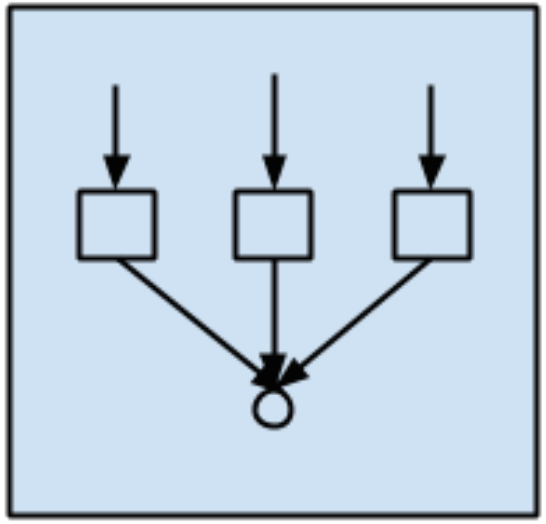
- Update to Artificial Neural Networks
- Goal is to build a much larger and more complex neural network
- Popular algorithms:
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Networks (RNNs)
  - Long Short-Term Memory Networks (LSTMs)
  - Deep Belief Networks (DBN)



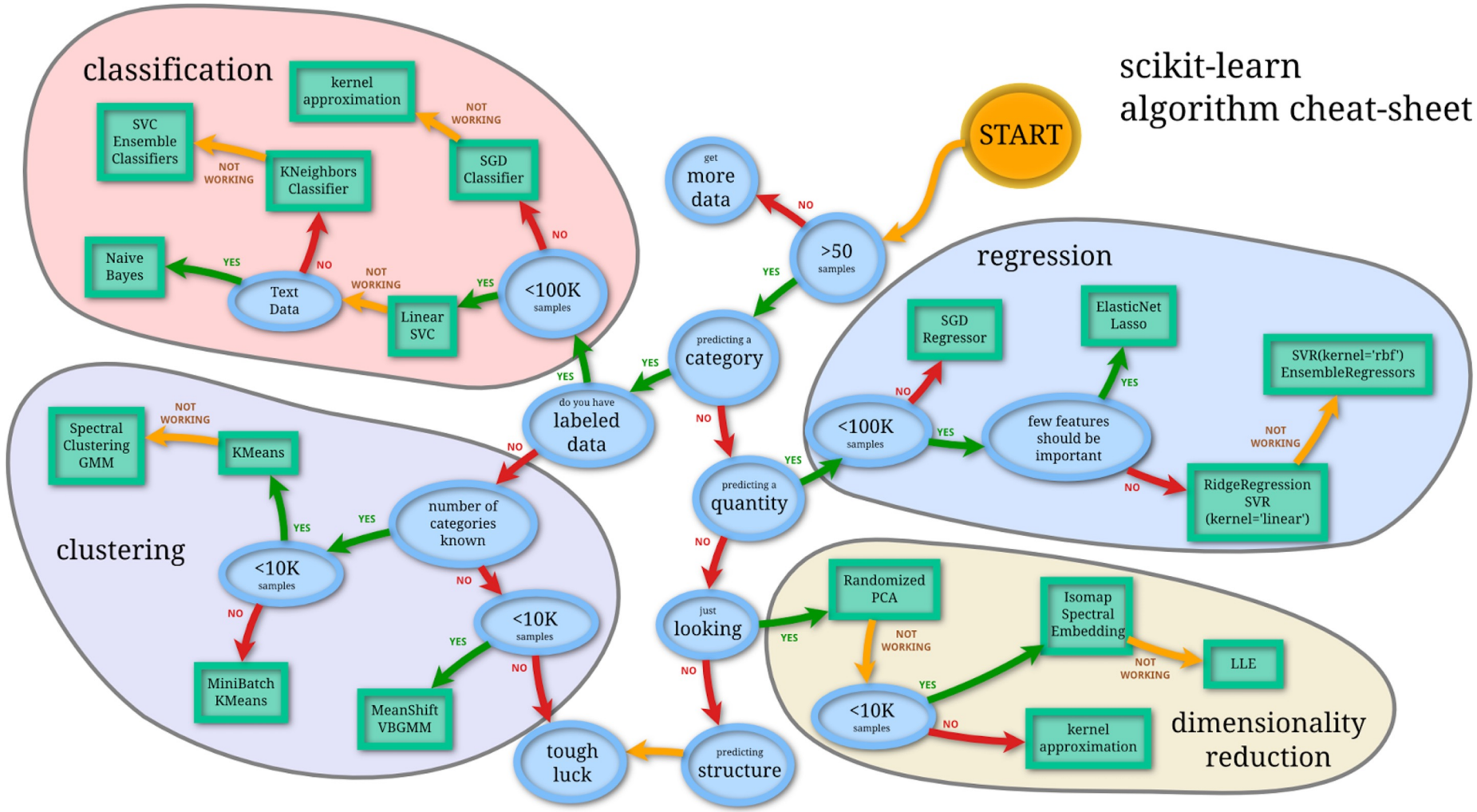
Deep Learning
Algorithms

Ensemble algorithms

- These are the models composed of multiple weaker models that are independently trained and the predictions are combined to make the overall prediction.
- Popular algorithms:
  - Boosting
  - Boostrapped Aggregation
  - AdaBoost
  - Gradient Boosting Machines
  - Random Forest



Ensemble Algorithms


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
NO
predicting a quantity
YES
NO
just looking
YES
NO
predicting structure
tough luck
do you have labeled data
YES
NO
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
YES
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift VBGMM
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
dimensionality reduction
Randomized PCA
YES
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation

What can we learn from this?

- <50 observations…get more data
  - Why? Inadequate power to effectively detect patterns or relationships. Visualization can still be very useful.
  - The power of a statistical test is the probability that it correctly rejects the null hypothesis when the null hypothesis is false.

Which drives the choice of method, the dependent variable or the independent variable?