

ML ALGORITHM

**In this session**

* Supervised vs Unsupervised
* Algorithm group
* Linear regression
* Logistic regression
* Decision trees
* Neural networks
* Support vector machines (SVMs)
* Bayesian methods
* Considerations when choosing an algorithm
* Cheat Sheet
* Algorithm’s performance comparison

Supervised vs Unsupervised

**Supervised**

* Train with know answer
* Can give answer with any new input, after sufficient training
* Create a function from inputs to give answer
* If the answers are expressed in classes, it is called classification problem
* If the answer space is continuous, it is called regression problem.

**Unsupervised**

* Training with unknown answer
* Can find the structure or relationships between different inputs
* Most important = clustering
* Anomaly detection

Algorithm group

Supervised: Make predictions based on a set of examples

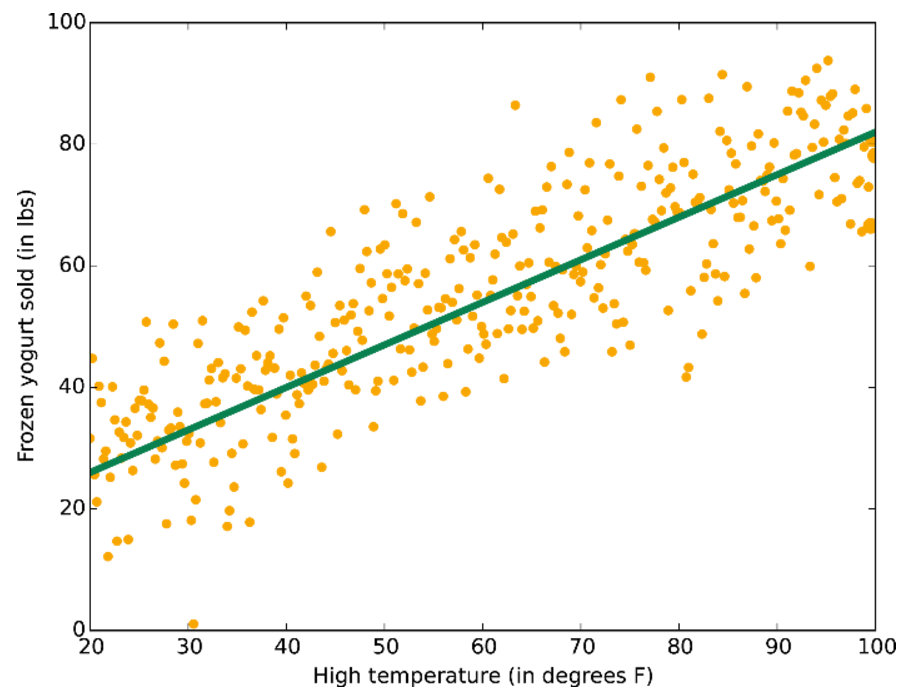
* Classification: predict a category
* Regression: predicted a value
* Anomaly detection: identify data unusual

Unsupervised: data points have no labels associated with them

* Clustering: discovering structure

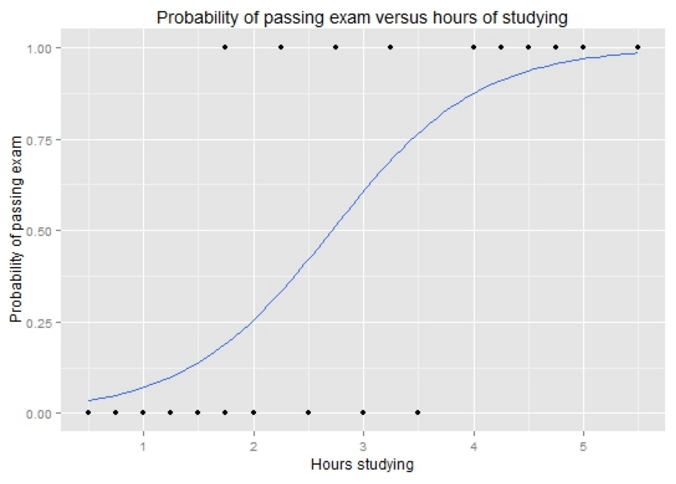
Linear regression

* Use when data fits a line
* It's a workhorse
* Simple and fast
* May be overly simplistic for some problems.

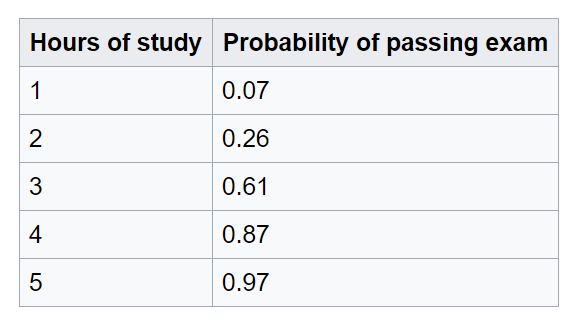


Higher temperature predicts better frozen yogurt sold

Logistic regression

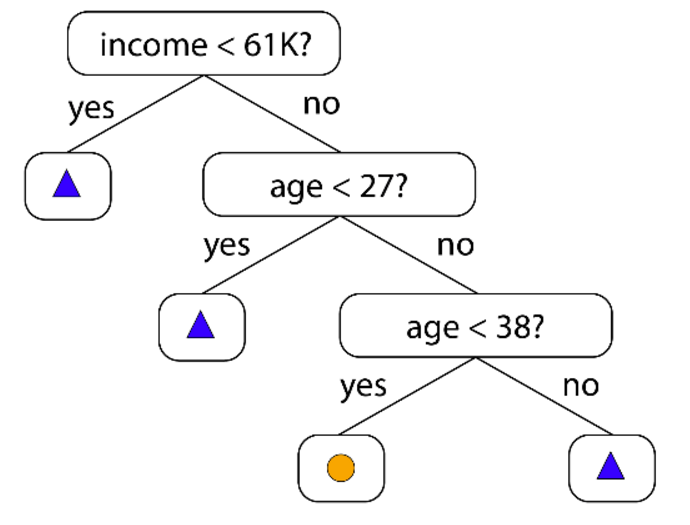
* Tool for two-class and multiclass classification
* Fast and simple
* Uses an 'S'-shaped curve
* Fit for dividing data into groups
* Linear approximation

Graph of a logistic regression curve showing probability of passing an exam versus hours studying



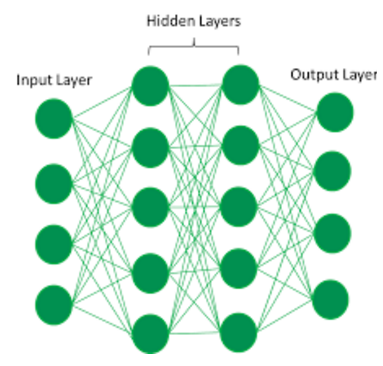
Decision trees

* Subdivide the feature space into regions with mostly the same label
* Decision forests (regression, two-class, and multiclass)
* Decision jungles (two-class and multiclass)
* Boosted decision trees (regression and two-class)
* Foundational machine learning concept



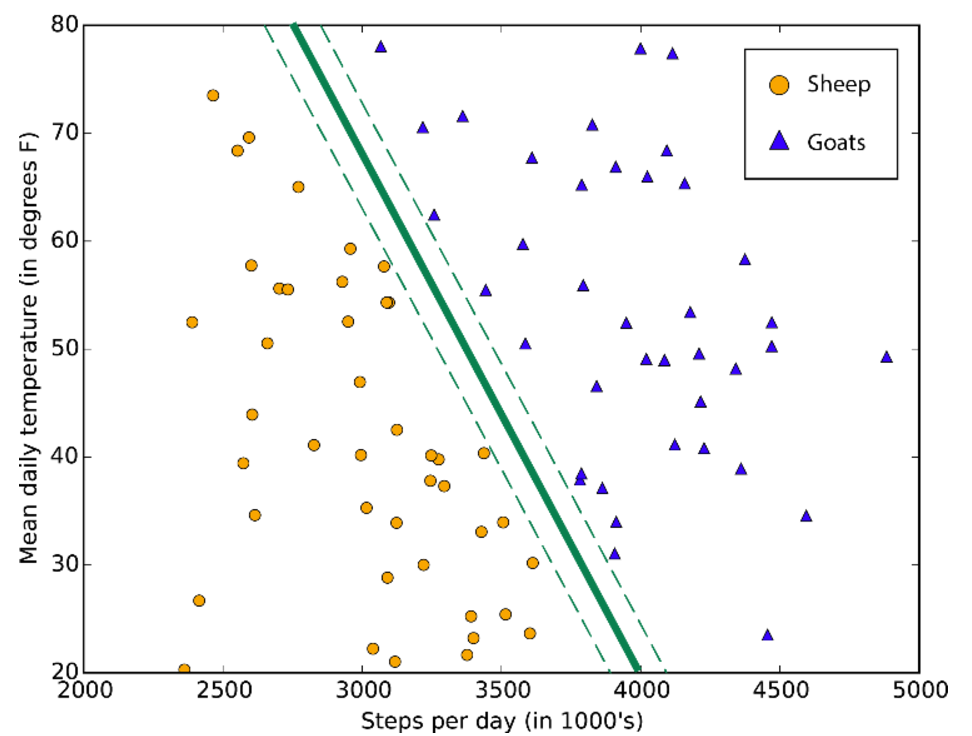
A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences

Neural networks

* Brain-inspired
* Multiclass, two-class, and regression
* Many-layered networks = "deep learning"
* Take a long time to train
* Have more parameters

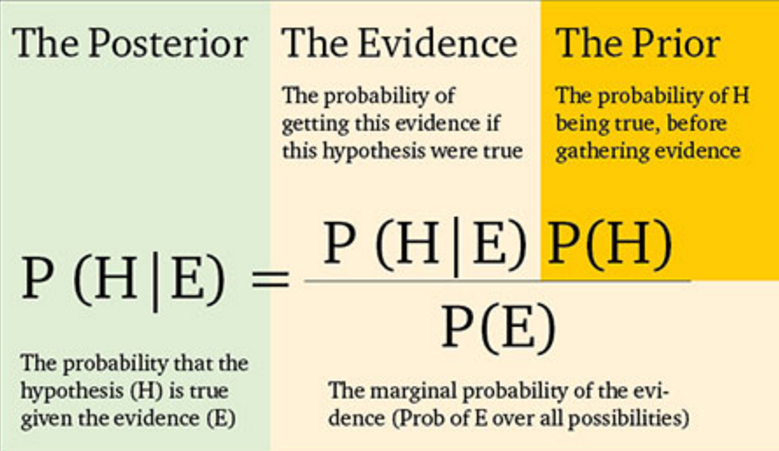
Deep learning use a cascade of many layers of nonlinear processing units for feature extraction and transformation

Support vector machines (SVMs)

* Find the boundary that separates classes
* When the two classes can't be clearly separated
* Uses a linear kernel
* Run fairly quickly
* Feature-intense data (DNA)
* Requiring only a modest amount of memory

A typical support vector machine class boundary maximizes the margin separating two classes

Bayesian methods

* Make the assumption of data points
* One data point is related with others
* Number of minutes until the next subway train arrives
* Two measurements taken a day apart are independent
* Two measurements taken a minute apart are not independent
* The value is highly predictive

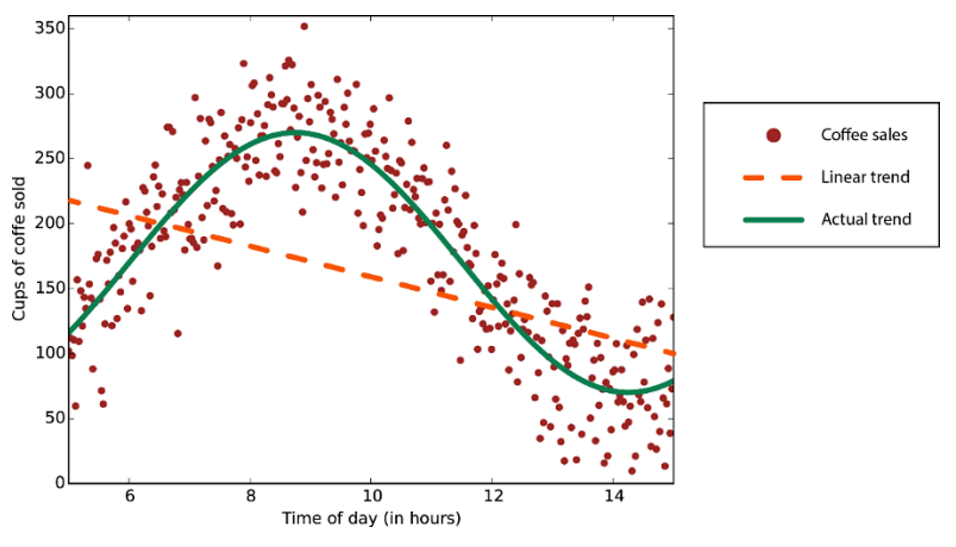
This expression describes how an existing belief (“prior”) held before any evidence is considered, is updated by the evidence to produce a new level of belief (“posterior”).

Considerations when choosing an algorithm

Considerations when choosing an algorithm

* Accuracy: most accurate isn't always necessary
* Training time: more accuracy = longer time
* Linearity: most are liner but not always

Considerations when choosing an algorithm

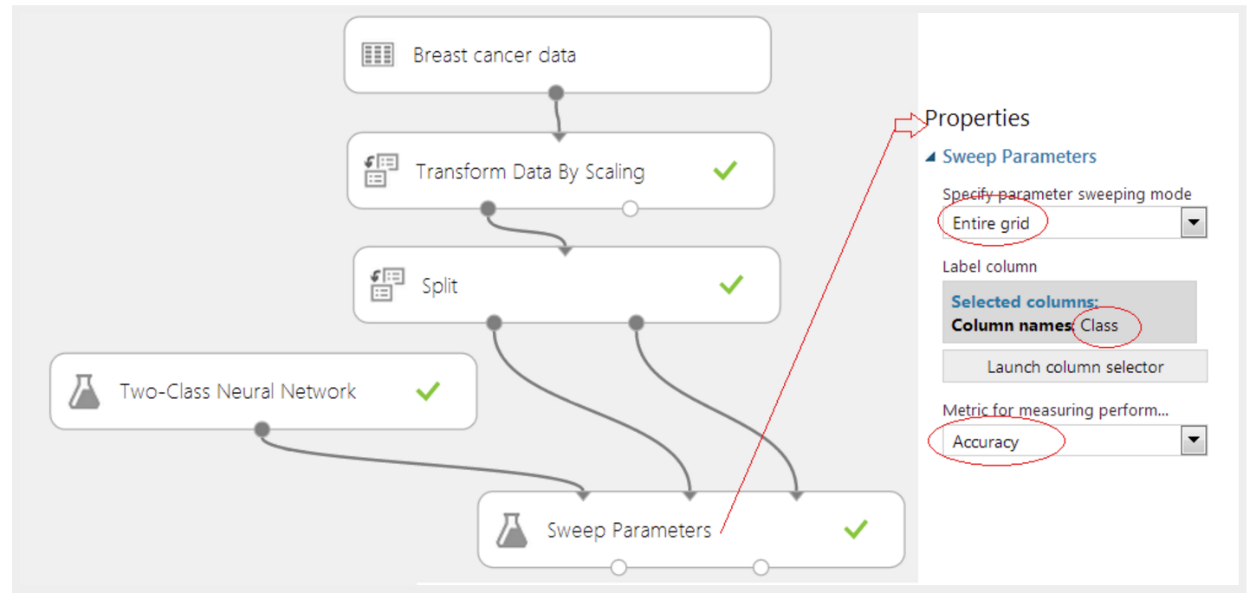


Data with a nonlinear trend - using a linear regression method would generate much larger errors than necessary

Considerations when choosing an algorithm

Algorithm’s parameters

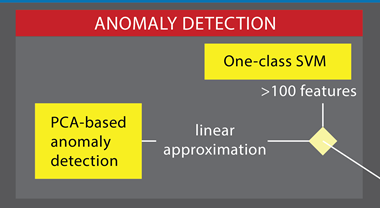
* Are the knobs a data scientist
* Turn when setting up an algorithm
* Affect the algorithm's behavior
* Must understand the in-side out of algorithm
* Use parameter sweeping to automatically tries all parameter



Considerations when choosing an algorithm

Number of features

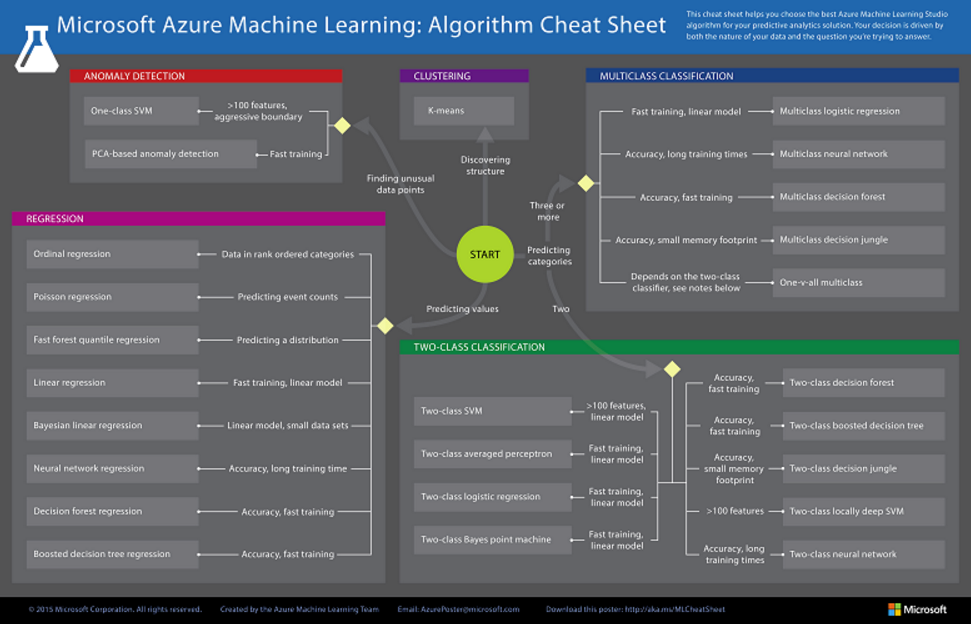
* Can be very large for genetics or textual data
* The large number can bog down some algorithms
* Making training time long
* Go deep
* Support Vector Machines (SVM)



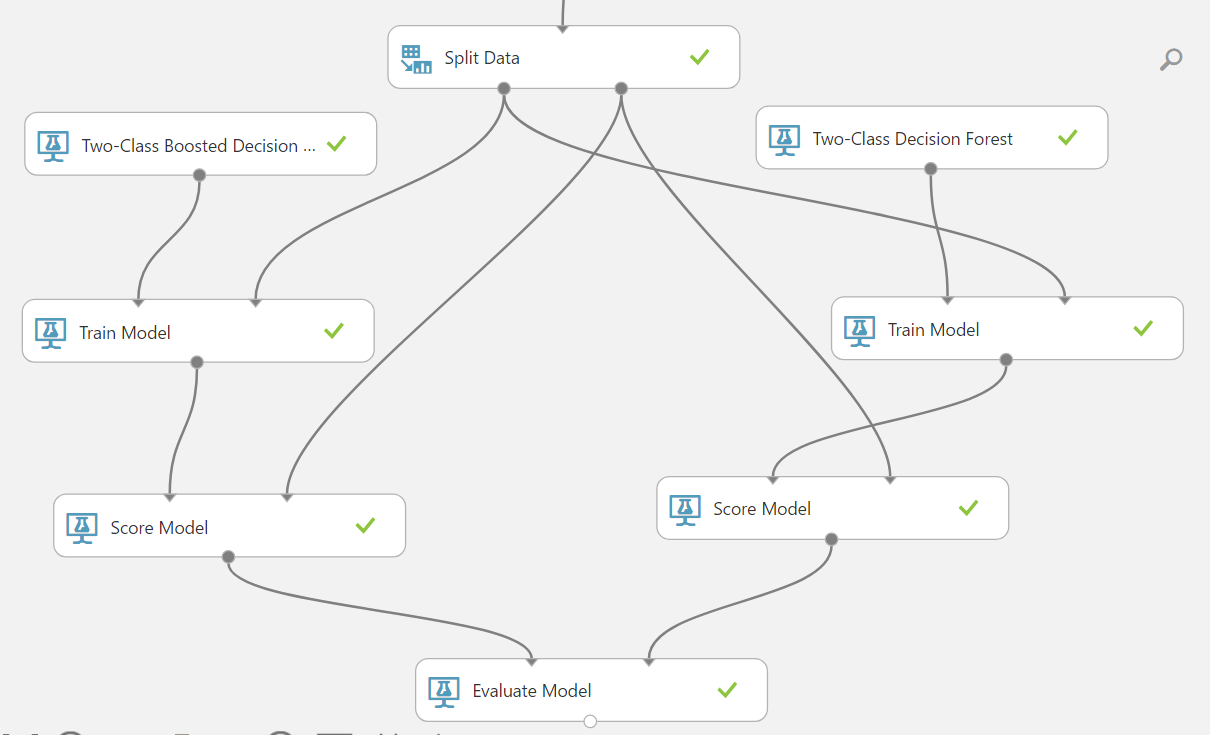
Cheat Sheet

Machine Learning Algorithm Cheat Sheet (11x17 in.)

http://download.microsoft.com/download/A/6/1/A613E11E-8F9C-424A-B99D-65344785C288/microsoft-machine-learning-algorithm-cheat-sheet-v6.pdf



Algorithm’s performance comparison



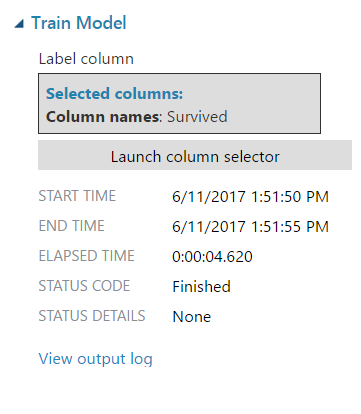
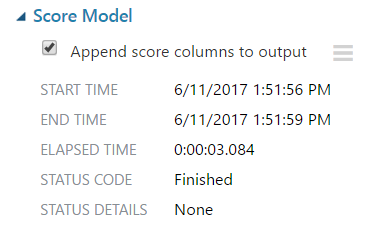
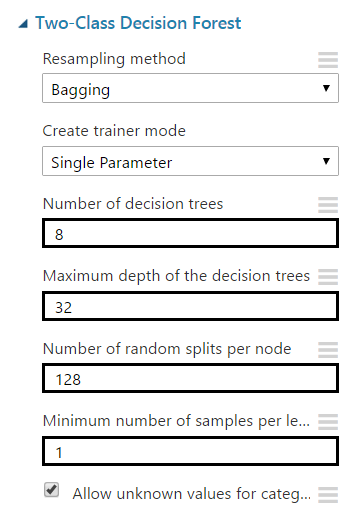
Algorithm’s performance comparison

Algorithm’s performance comparison

1. Open Experiment Titanic
2. Save as Titanic two algorithm
3. Drag & drop modules
   1. Two-Class Decision Forest module
   2. Train Module
   3. Score Module
4. Set module properties
5. Save Experiment
6. Run Experiment
7. View Visualize / ROC Curve and Evaluation metrics

Algorithm’s performance comparison

Modules properties setting



More information

A Tour of Machine Learning Algorithms

http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

