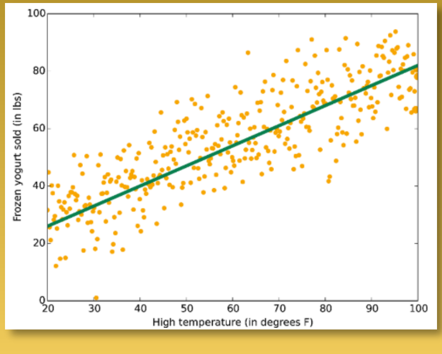
ML Algorithm

ML ALGORITHM



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

ML Algorithm

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

ML Algorithm

Algorithm group

Supervised: Make predictions based on a set of examples

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

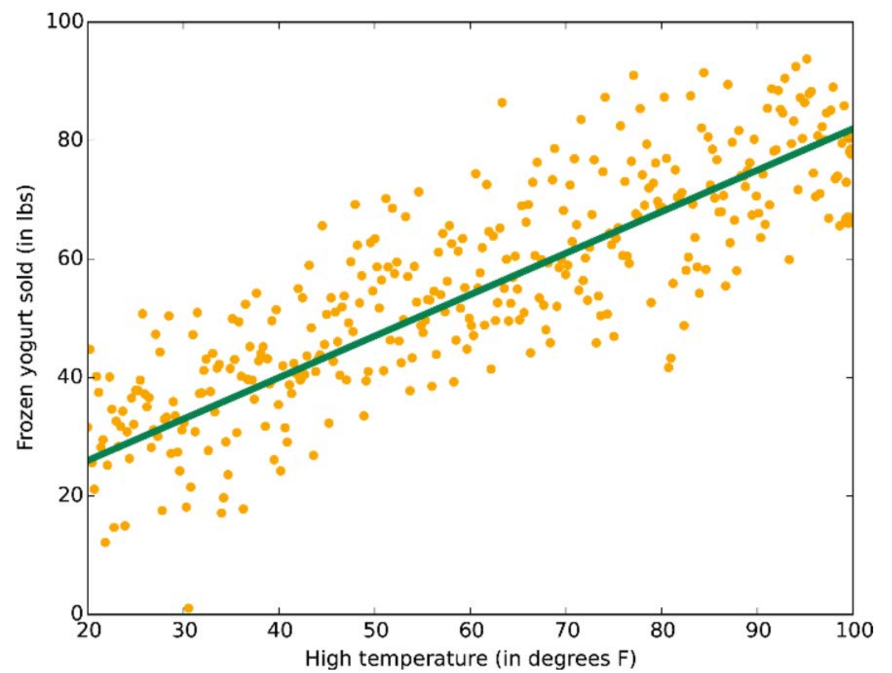
Unsupervised: data points have no labels associated with them

- Clustering: discovering structure

ML Algorithm

Linear regression

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



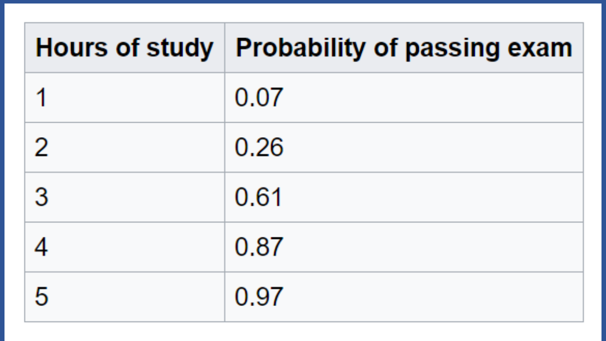
Higher temperature predicts better frozen yogurt sold

ML Algorithm

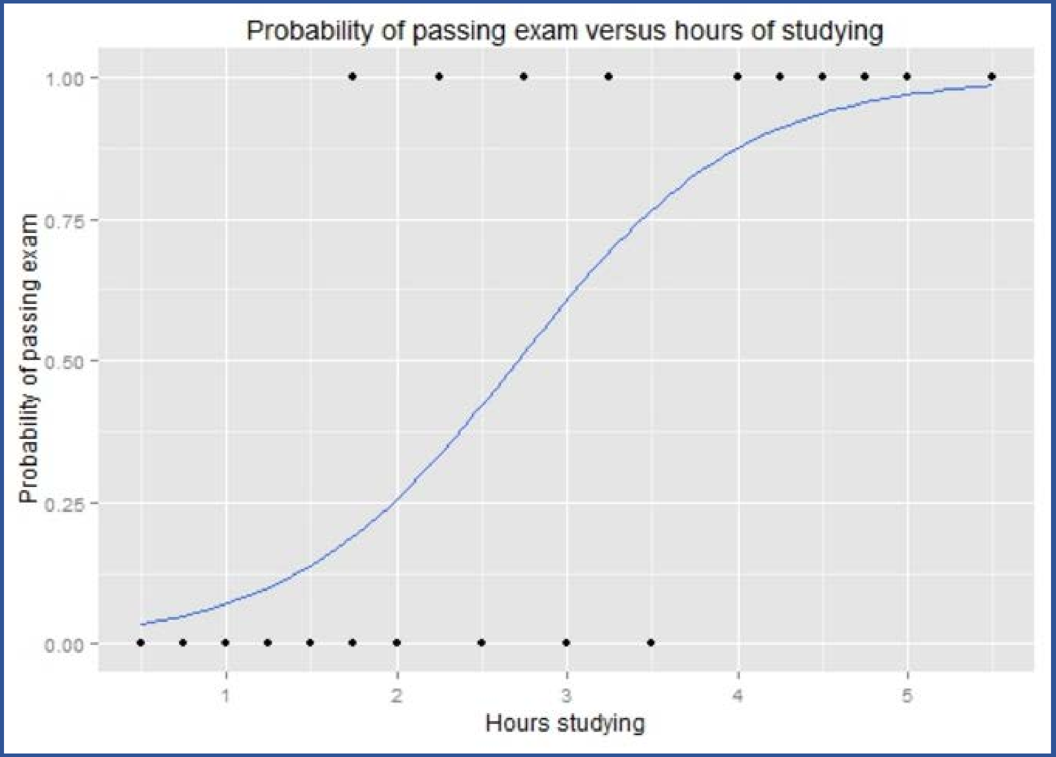
Logistic regression

- 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



Hours of study	Probability of passing exam
1	0.07
2	0.26
3	0.61
4	0.87
5	0.97

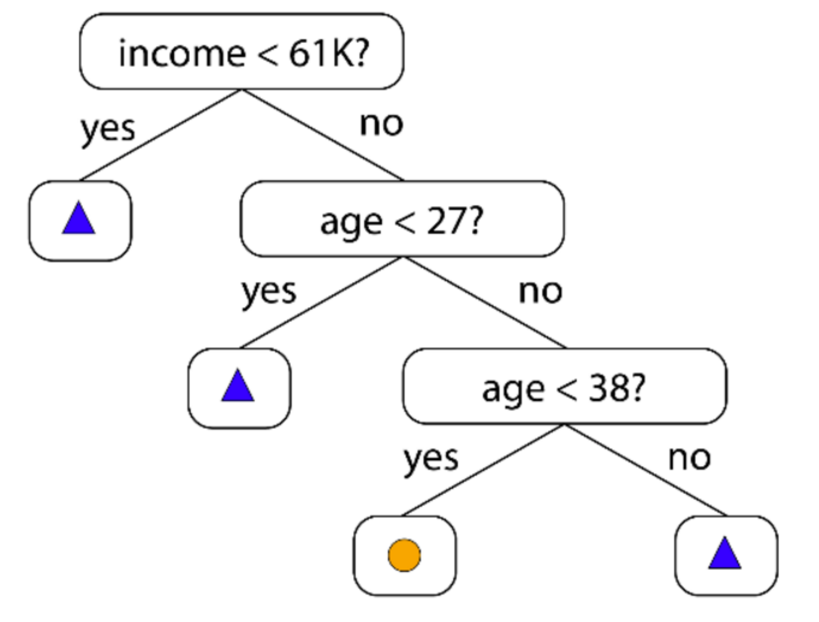


ML Algorithm

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

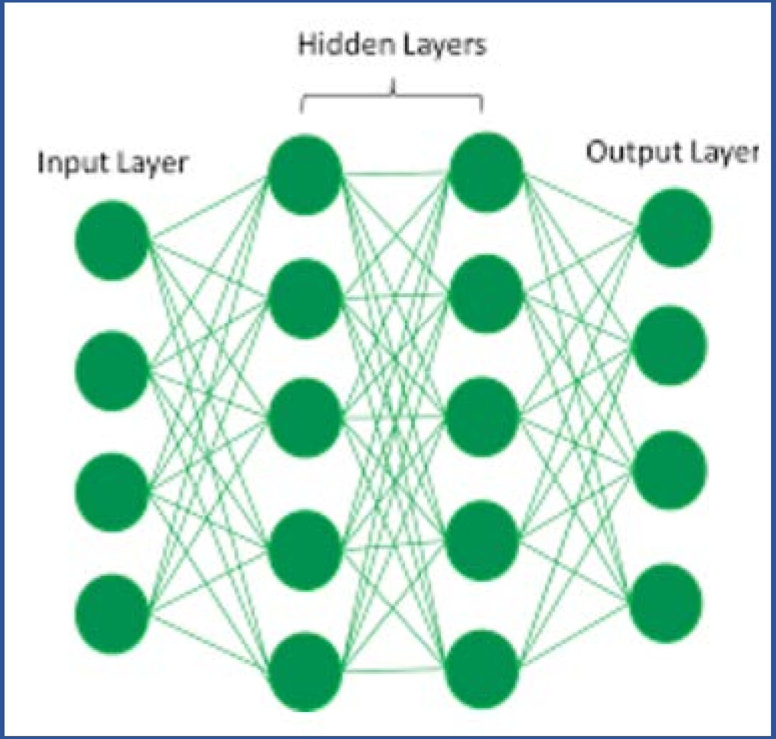
A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences



ML Algorithm

Neural networks

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

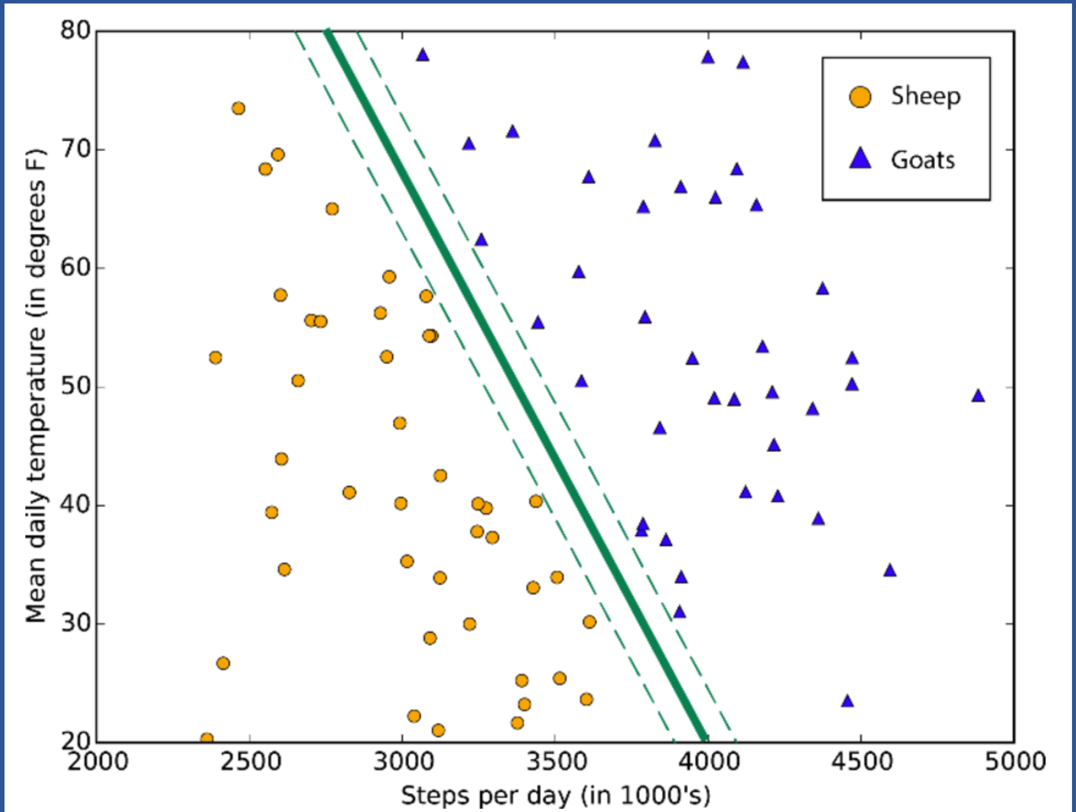


Deep learning use a cascade of many layers of nonlinear processing units for feature extraction and transformation

ML Algorithm

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



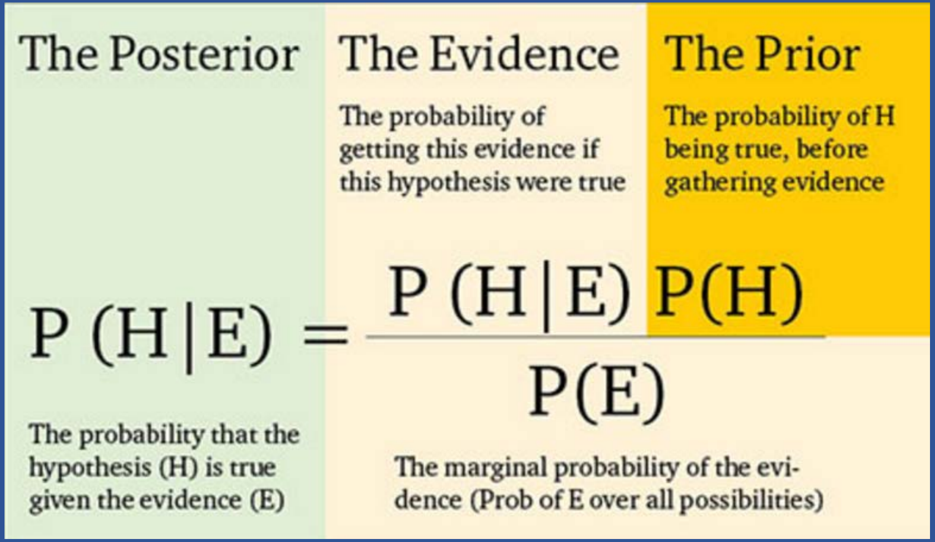
A typical support vector machine class boundary maximizes the margin separating two classes

ML Algorithm

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”).



The Posterior

The Evidence

The probability of getting this evidence if this hypothesis were true

The Prior

The probability of H being true, before gathering evidence

$$P\,(H|E) = \frac{P\,(H|E)\;P(H)}{P(E)}$$

The probability that the hypothesis (H) is true given the evidence (E)

The marginal probability of the evidence (Prob of E over all possibilities)

ML Algorithm

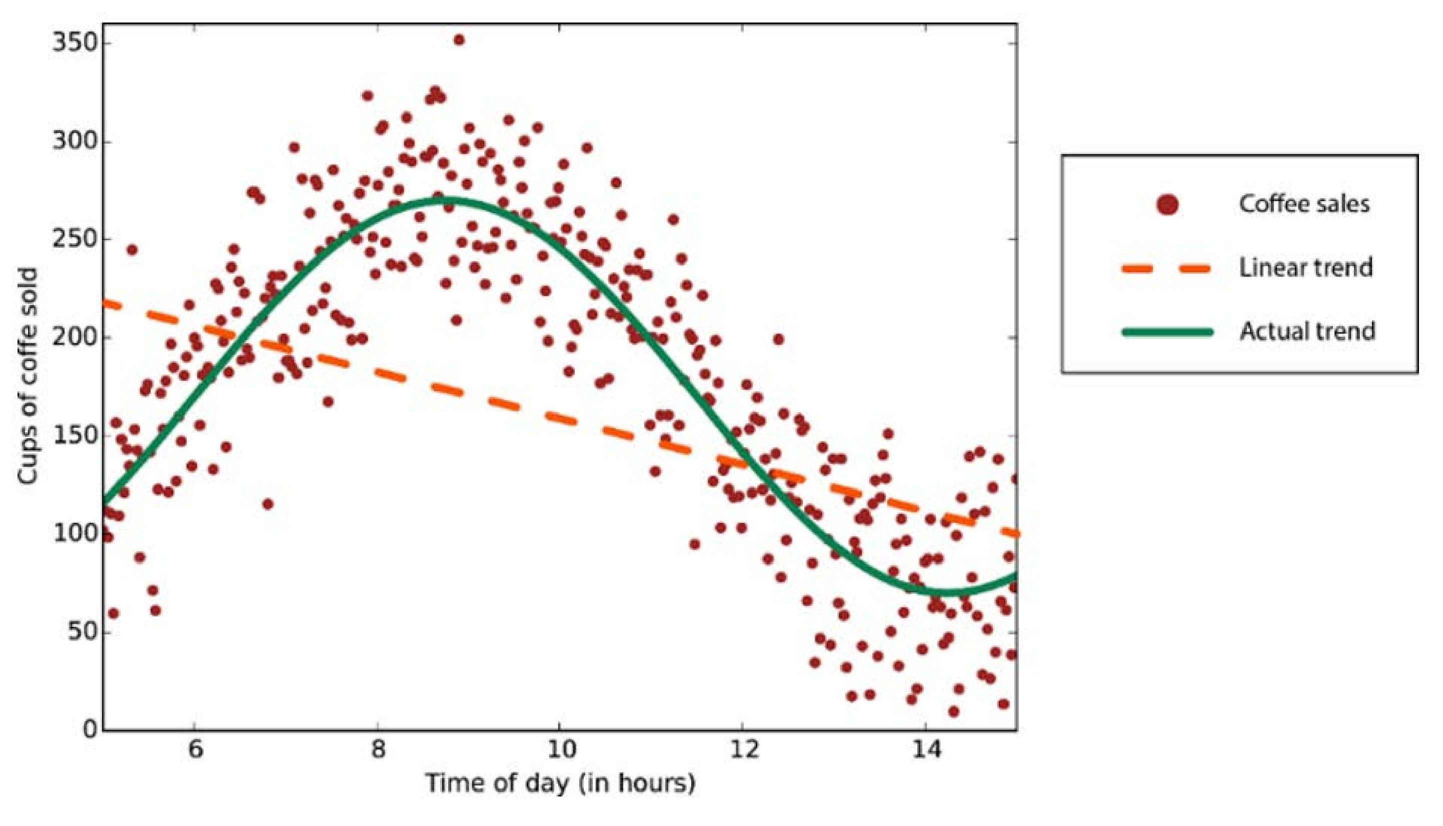
Considerations when choosing an algorithm

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

ML Algorithm

Considerations when choosing an algorithm



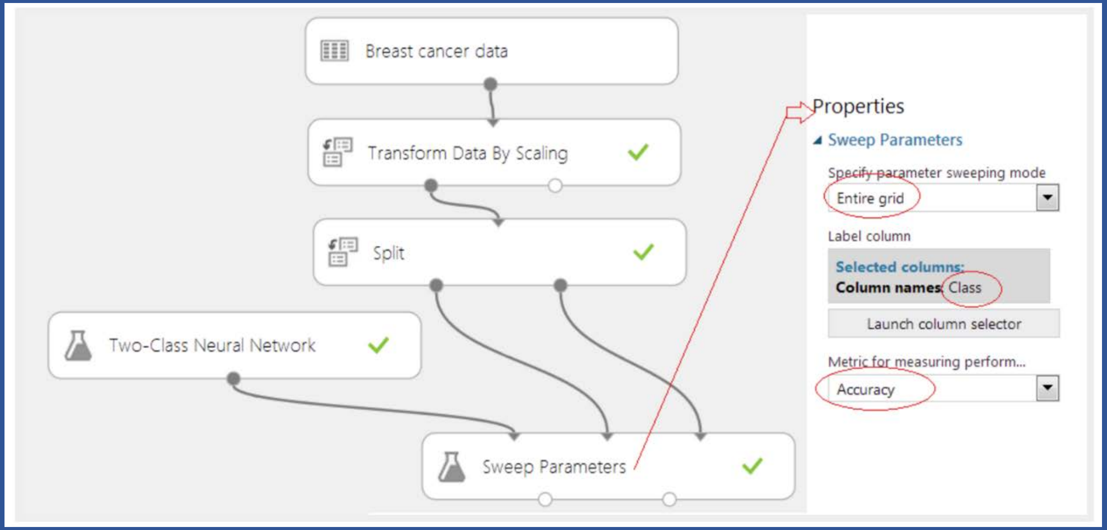
Data with a nonlinear trend - using a linear regression method would generate much larger errors than necessary

ML Algorithm

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



Properties

Sweep Parameters

Specify parameter sweeping mode: Entire grid

Label column

Selected columns:
Column names: Class

Launch column selector

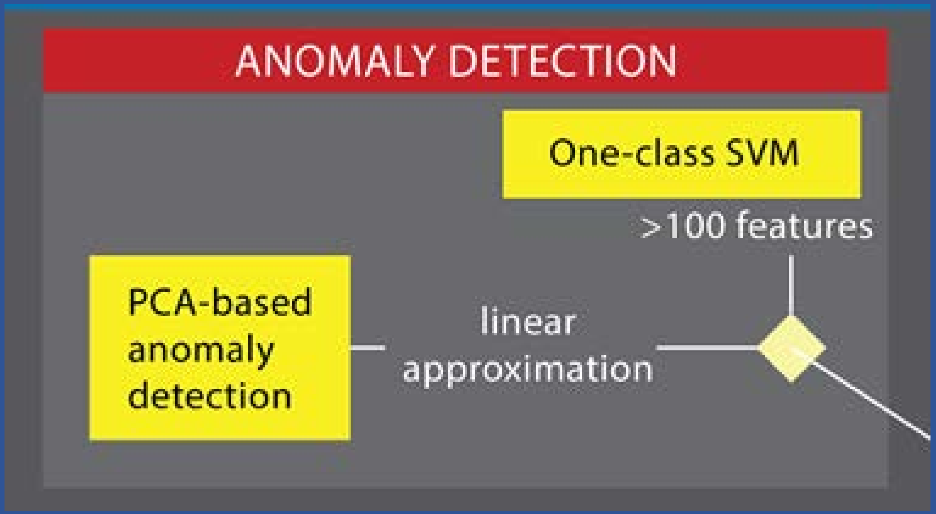
Metric for measuring perform...: Accuracy

ML Algorithm

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)

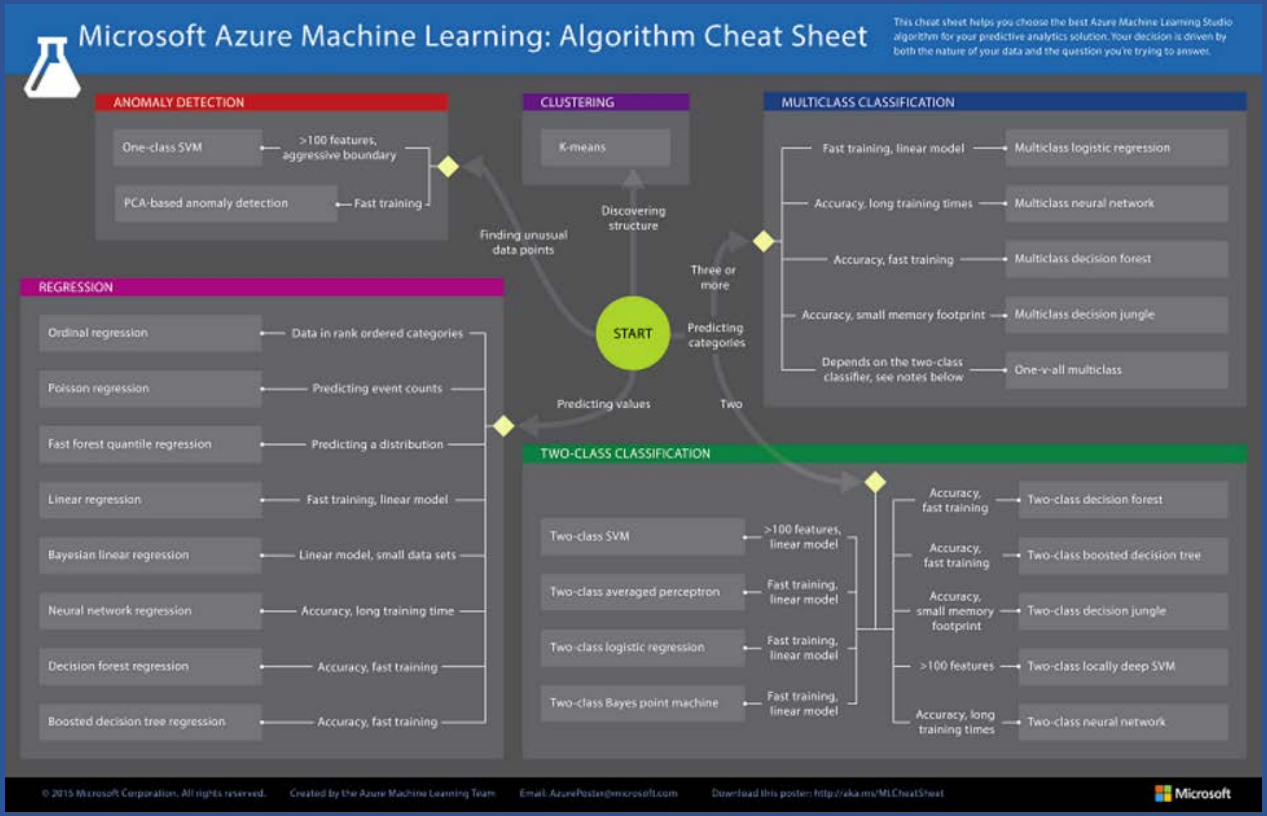


ML Algorithm

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



Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points → ANOMALY DETECTION

- One-class SVM — >100 features, aggressive boundary
- PCA-based anomaly detection — Fast training

Discovering structure → CLUSTERING

- K-means

Predicting categories — Three or more → MULTICLASS CLASSIFICATION

- Fast training, linear model — Multiclass logistic regression
- Accuracy, long training times — Multiclass neural network
- Accuracy, fast training — Multiclass decision forest
- Accuracy, small memory footprint — Multiclass decision jungle
- Depends on the two-class classifier, see notes below — One-v-all multiclass

Predicting categories — Two → TWO-CLASS CLASSIFICATION

- Two-class SVM — >100 features, linear model
- Two-class averaged perceptron — Fast training, linear model
- Two-class logistic regression — Fast training, linear model
- Two-class Bayes point machine — Fast training, linear model
- Accuracy, fast training — Two-class decision forest
- Accuracy, fast training — Two-class boosted decision tree
- Accuracy, small memory footprint — Two-class decision jungle
- >100 features — Two-class locally deep SVM
- Accuracy, long training times — Two-class neural network

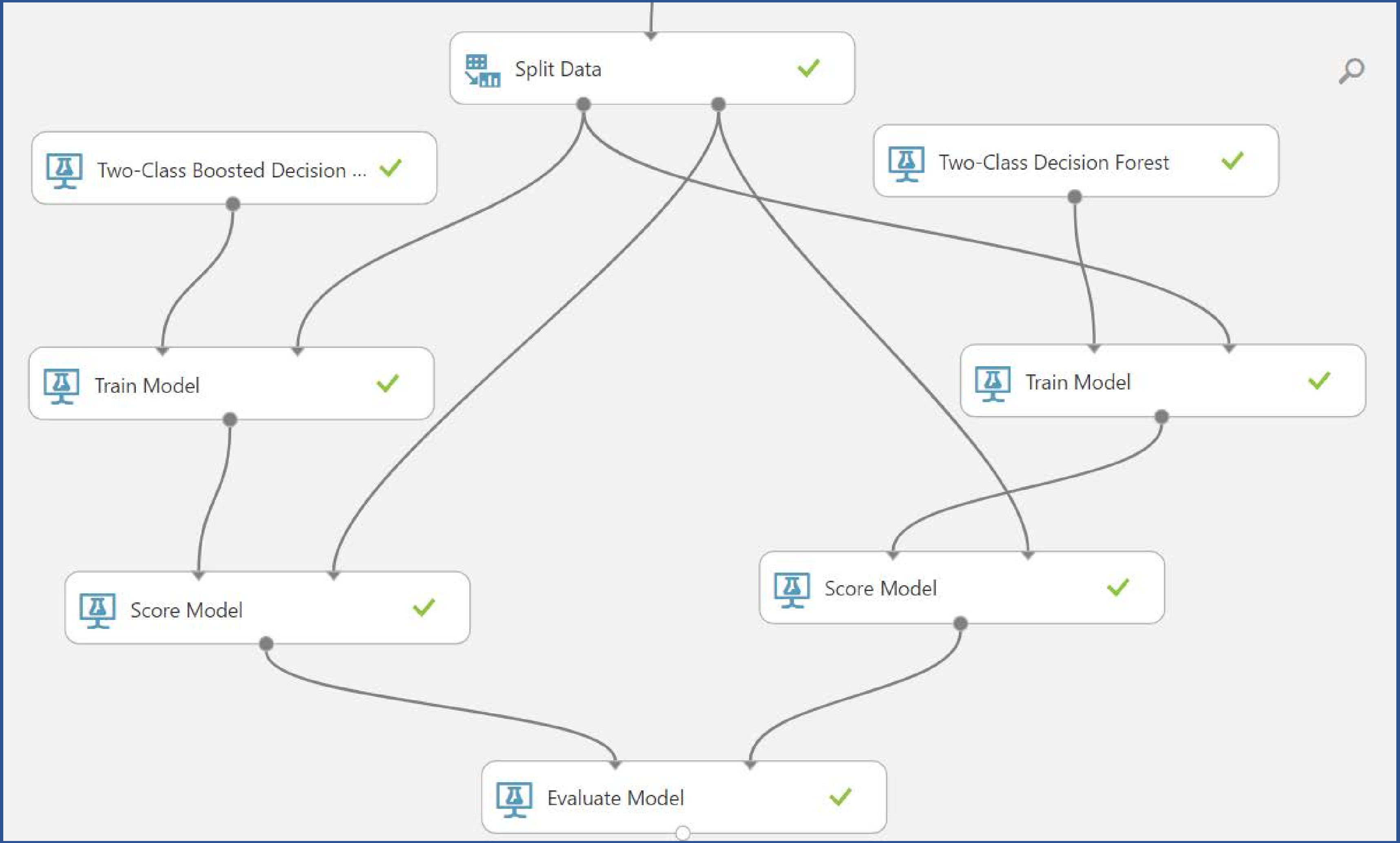
Predicting values → REGRESSION

- Ordinal regression — Data in rank ordered categories
- Poisson regression — Predicting event counts
- Fast forest quantile regression — Predicting a distribution
- Linear regression — Fast training, linear model
- Bayesian linear regression — Linear model, small data sets
- Neural network regression — Accuracy, long training time
- Decision forest regression — Accuracy, fast training
- Boosted decision tree regression — Accuracy, fast training

© 2015 Microsoft Corporation. All rights reserved. Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet

ML Algorithm

Algorithm’s performance comparison



ML Algorithm

Algorithm’s performance comparison

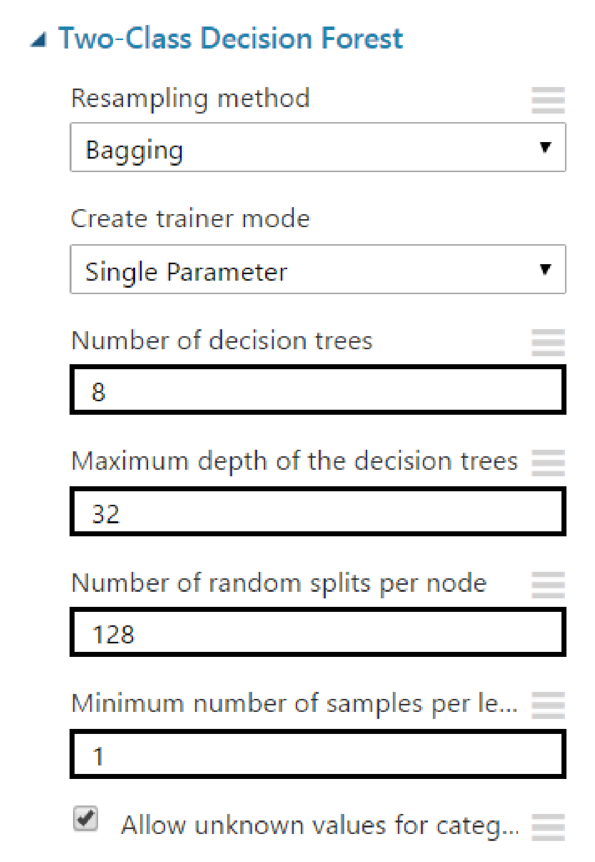
Algorithm’s performance comparison

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

ML Algorithm

Algorithm's performance comparison

Modules properties setting



◢ Two-Class Decision Forest

Resampling method

Bagging ▾

Create trainer mode

Single Parameter ▾

Number of decision trees

8

Maximum depth of the decision trees

32

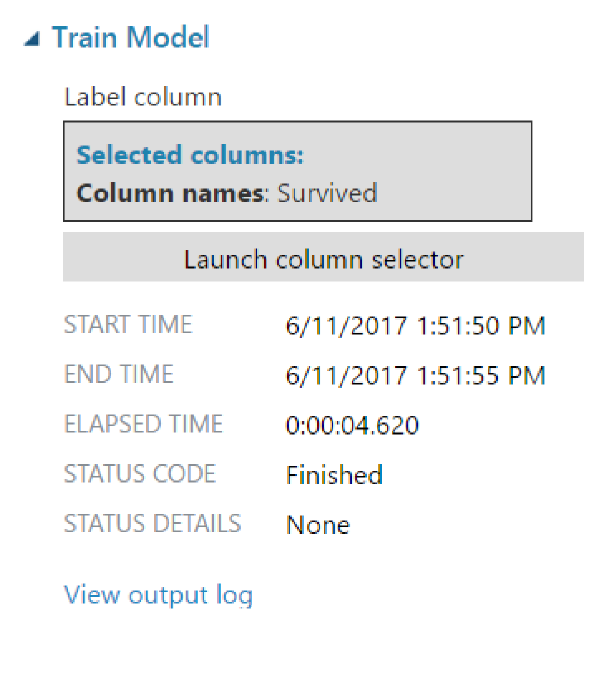
Number of random splits per node

128

Minimum number of samples per le...

1

☑ Allow unknown values for categ...

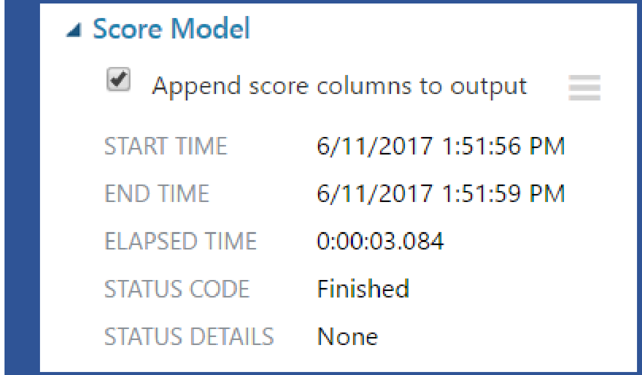


◢ Train Model

Label column

Selected columns:
Column names: Survived

START TIME	6/11/2017 1:51:50 PM
END TIME	6/11/2017 1:51:55 PM
ELAPSED TIME	0:00:04.620
STATUS CODE	Finished
STATUS DETAILS	None



◢ Score Model

☑ Append score columns to output

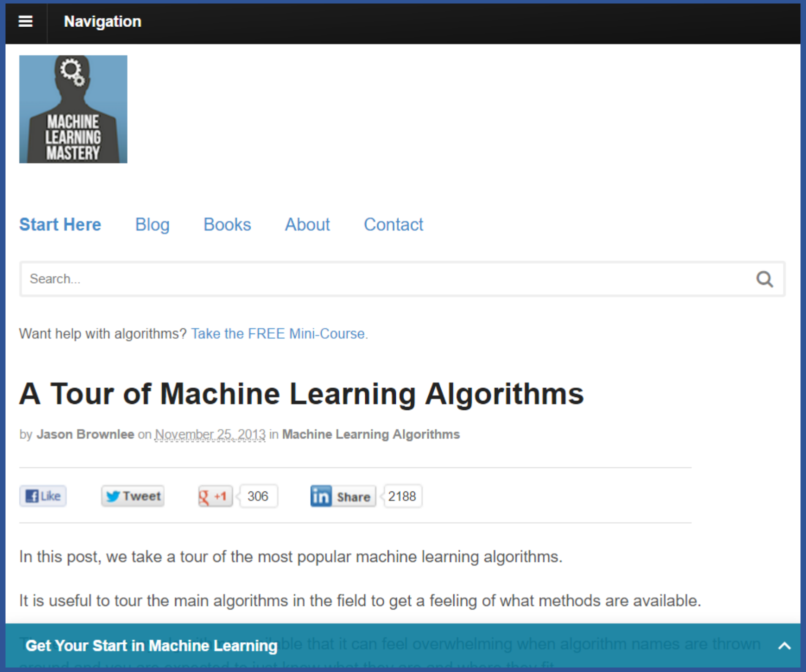
START TIME	6/11/2017 1:51:56 PM
END TIME	6/11/2017 1:51:59 PM
ELAPSED TIME	0:00:03.084
STATUS CODE	Finished
STATUS DETAILS	None

ML Algorithm

More information

A Tour of Machine Learning Algorithms

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



Want help with algorithms? Take the FREE Mini-Course.

A Tour of Machine Learning Algorithms

by Jason Brownlee on November 25, 2013 in Machine Learning Algorithms

In this post, we take a tour of the most popular machine learning algorithms.

It is useful to tour the main algorithms in the field to get a feeling of what methods are available.