UNIVERSITY of WASHINGTON

Introduction to Machine Learning

David Beck[1,2,3], Joseph Hellerstein[1,3], **Bernease Herman[1]**, Colin Lockard[3]

[1]eScience Institute
[2]Chemical Engineering
[3]Computer Science

November 20, 2018

Agenda

1. Team Standups
2. Conceptual intro to machine learning
3. Machine learning algorithms
4. If time, Python exercise in **`scikit-learn`** and **`keras`**

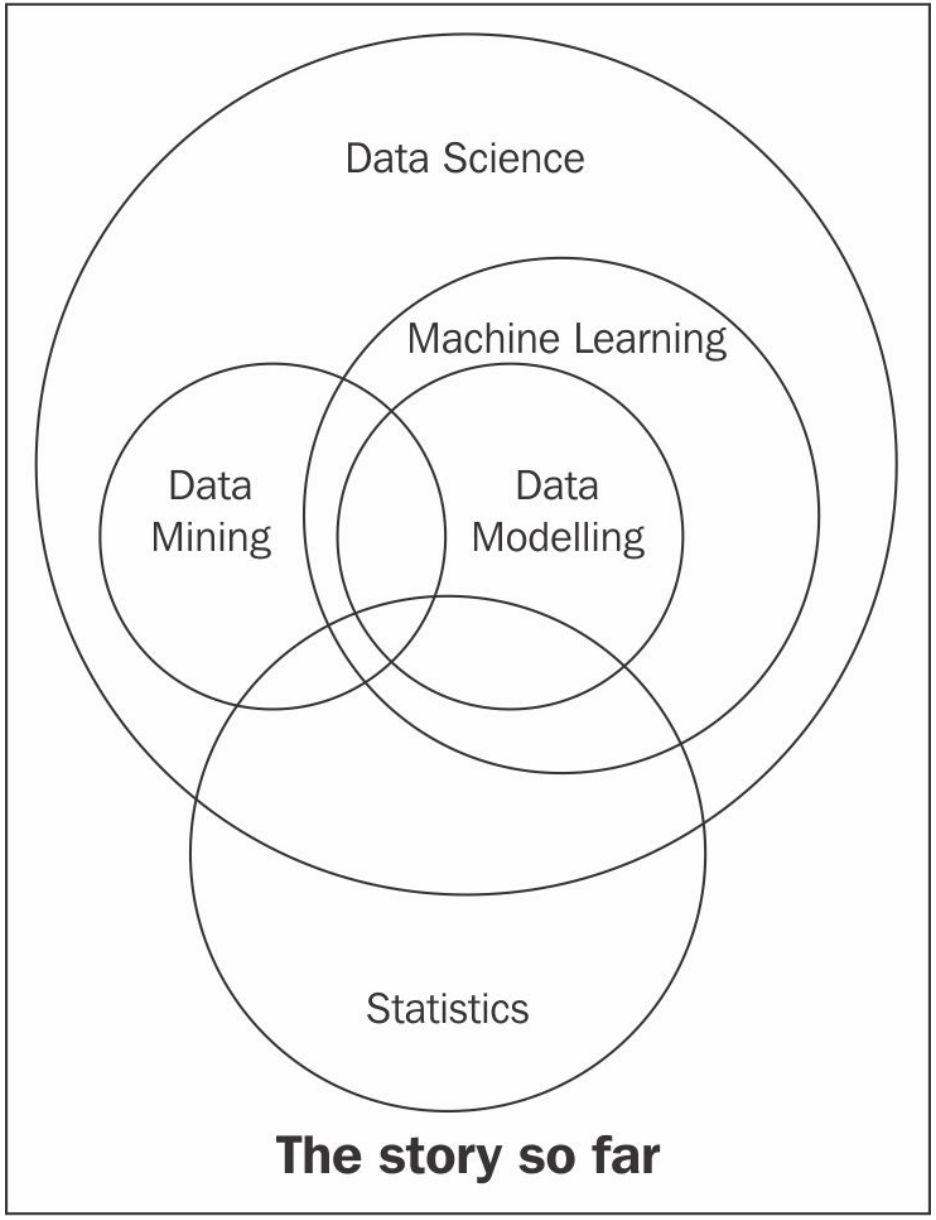
Team standups

1. What progress have you made in the last two weeks *(excluding technology reviews)*?
2. Anything cool learned worth sharing?
3. Anything you need help with?

Where does machine learning fit into data science?

Data science is the practice of translating data into insights.

Machine learning is a collection of methods to allow a computer to "learn" (improve performance) without explicit programming.



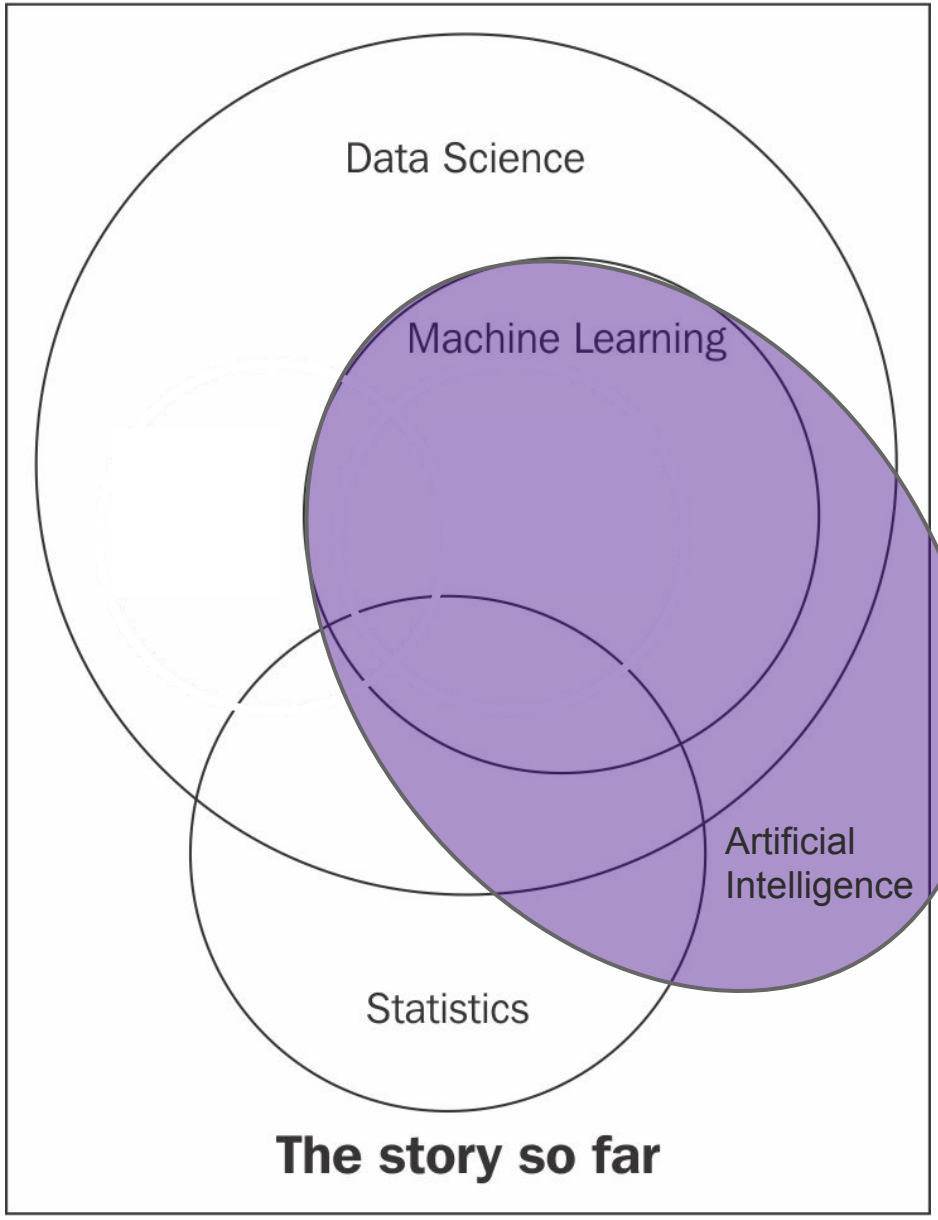
Credit: Sinan Ozdemir, Packt
https://www.packtpub.com/books/content/data-science-venn-diagram

Where does machine learning fit into data science?

Data science.

Machine learning.

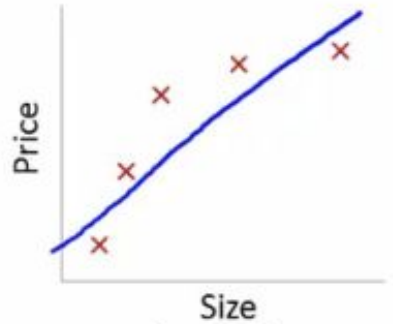
Artificial intelligence is the academic discipline of getting computers to demonstrate intelligence in contrast to that of humans and animals.



Credit: Sinan Ozdemir, Packt
https://www.packtpub.com/books/content/data-science-venn-diagram

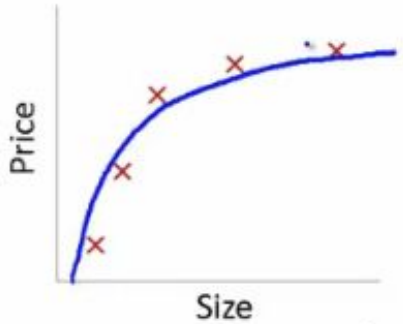
Overfitting and regularization

Models that are too specific will perform *worse* on new data.



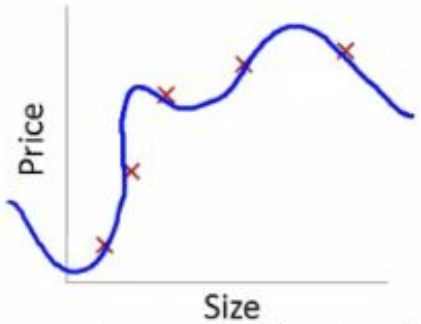
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



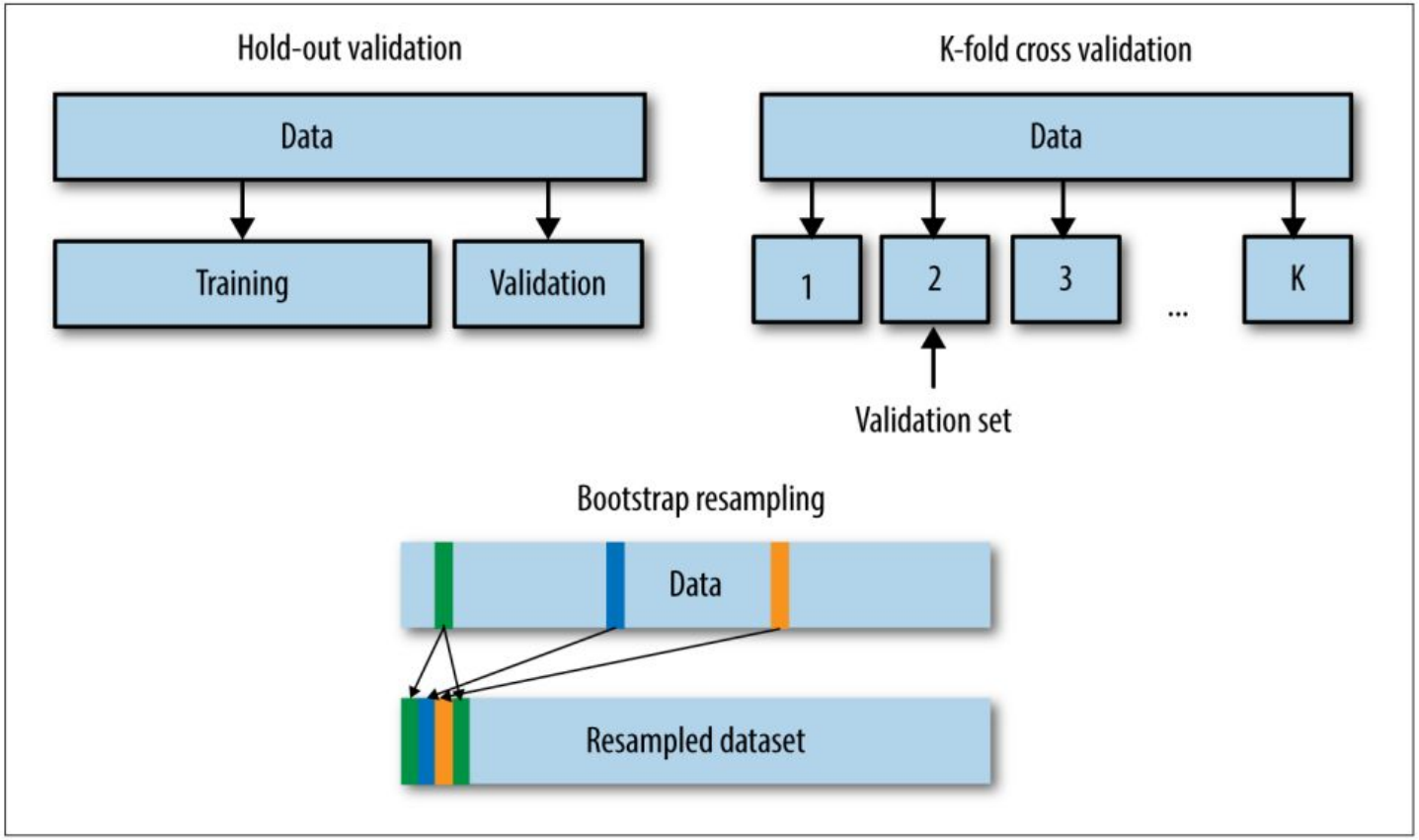
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

Credit: Andrew Ng, Coursera

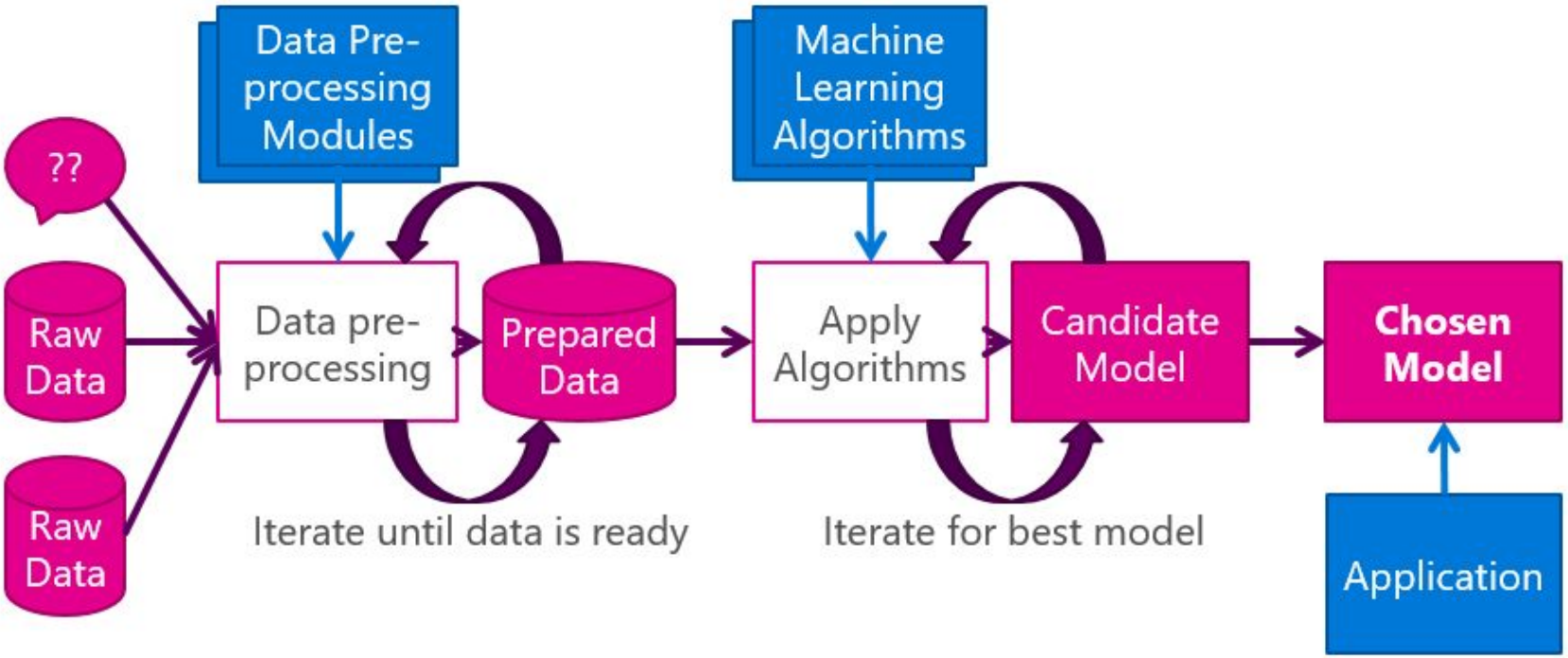
Model validation and bootstrapping

Rearranging and subsetting data to improve model performance.



Credit: Alice Zheng, Evaluating Machine Learning Models

Machine Learning Pipeline



Credit: Martin Kearn, Machine Learning is for Muggles too!
https://blogs.msdn.microsoft.com/martinkearn/2016/03/01/machine-learning-is-for-muggles-too/

```
from sklearn.cross_validation import
train_test_split


# Split the `digits` data into training
and test sets

X_train, X_test, y_train, y_test,
images_train, images_test =
train_test_split(data, digits.target,
digits.images, test_size=0.25,
random_state=42)
```

```
# Import the `cluster` module
from sklearn import cluster

# Create the KMeans model
clf = cluster.KMeans(init='k-means++',
n_clusters=10, random_state=42)

# Fit the training data to the model
clf.fit(X_train)
```

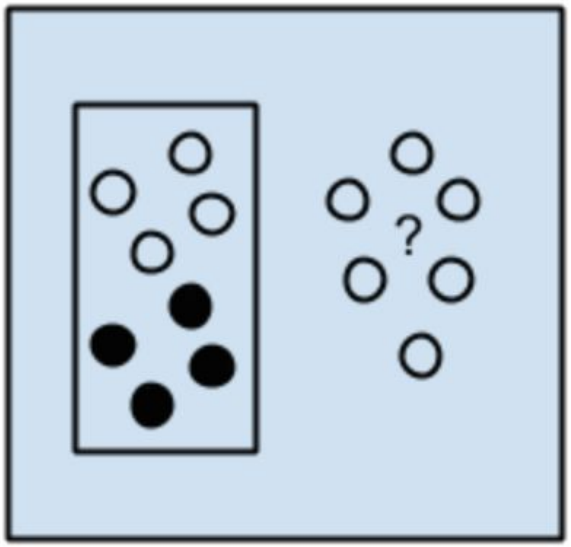
1. Supervised Learning

Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time.

A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.

Example problems are classification and regression.

Example algorithms include Logistic Regression and the Back Propagation Neural Network.



Supervised Learning Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

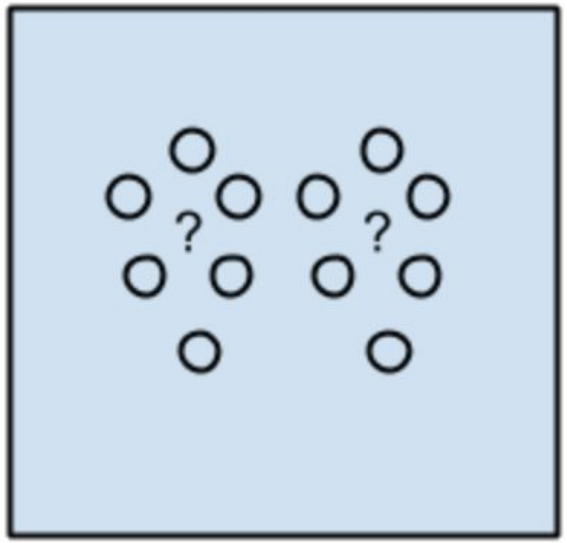
2. Unsupervised Learning

Input data is not labeled and does not have a known result.

A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

Example problems are clustering, dimensionality reduction and association rule learning.

Example algorithms include: the Apriori algorithm and k-Means.



Unsupervised Learning Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

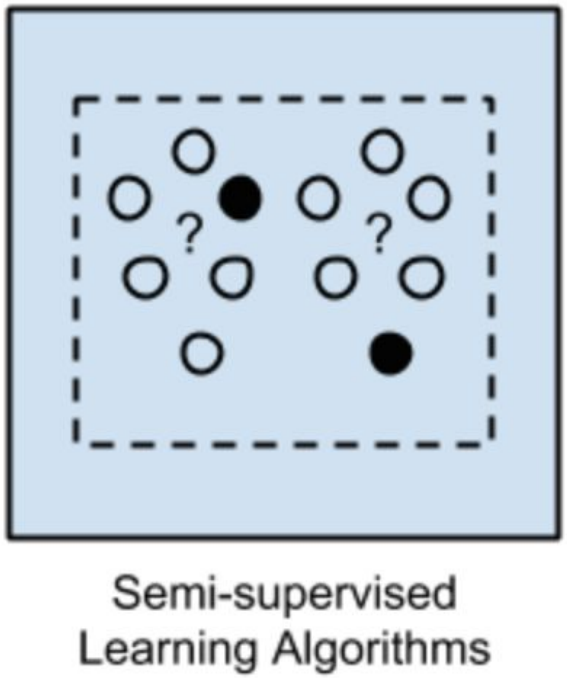
3. Semi-Supervised Learning

Input data is a mixture of labeled and unlabelled examples.

There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.

Example problems are classification and regression.

Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.



Semi-supervised Learning Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Other Algorithm Classes:

4. Reinforcement Learning

Input data is in the form of environment states with labels being rewards and punishments given as feedback to the programs actions in a dynamic environment.

Example problems include driving a vehicle or playing a game against an opponent.

5. Active Learning

This is a special task in labeled techniques (i.e., supervised, semi-supervised, reinforcement learning) that allows for the algorithm to interactively query for new labels on input data.

Machine Learning Algorithms

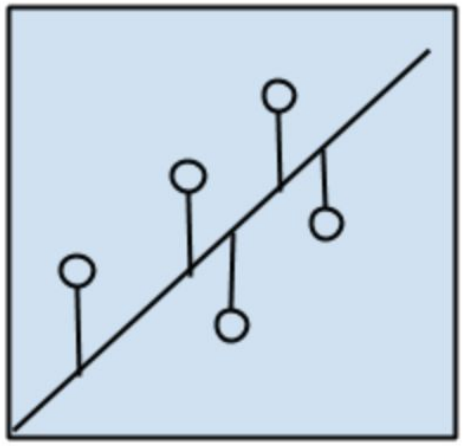
Regression Algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.

Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process.

The most popular regression algorithms are:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)



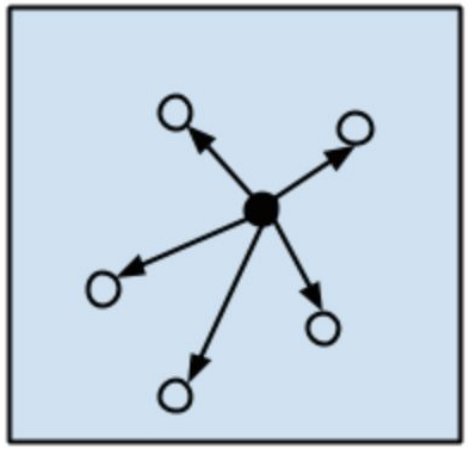
Regression Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Instance-based Algorithms

Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model.

Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances.



Instance-based Algorithms

The most popular instance-based algorithms are:

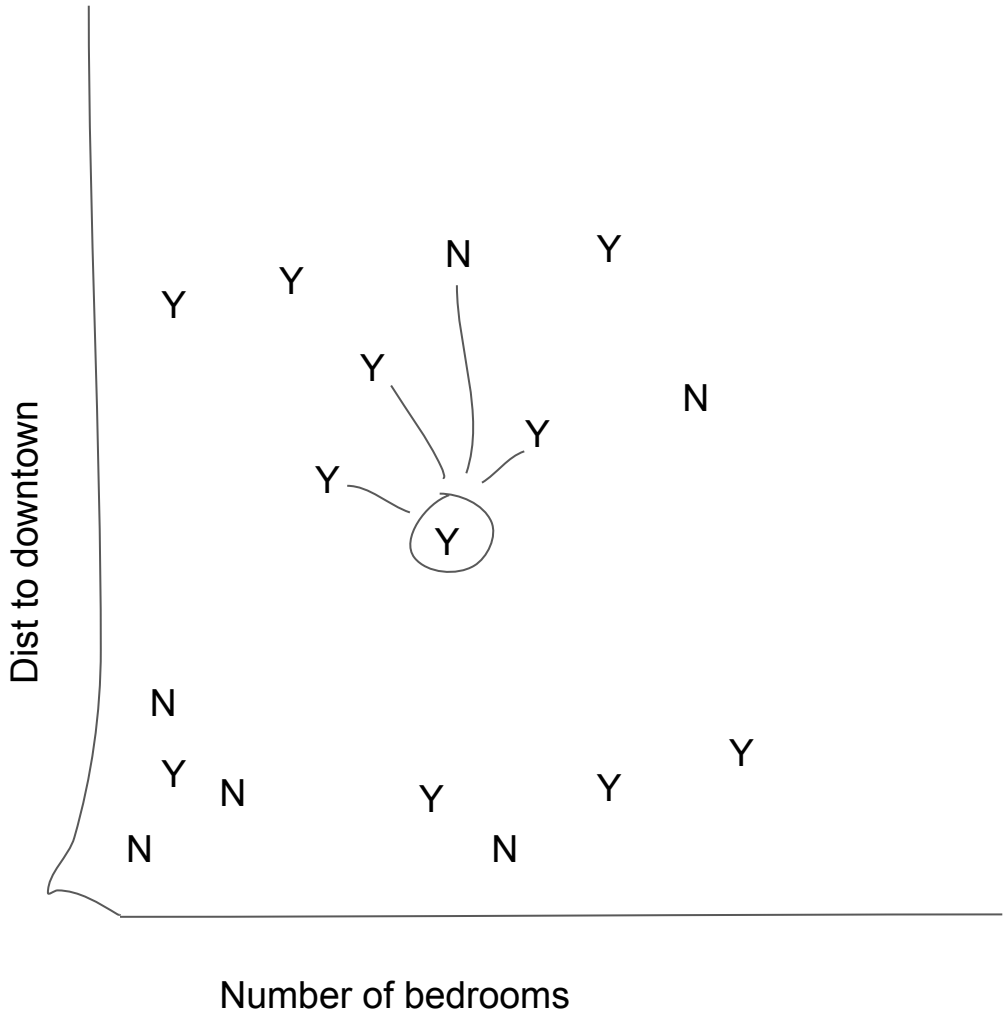
- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Apartment Pricing

Number of bedrooms
Distance to Downtown Seattle

Classification: Has dishwasher
Regression: Rent (monthly)



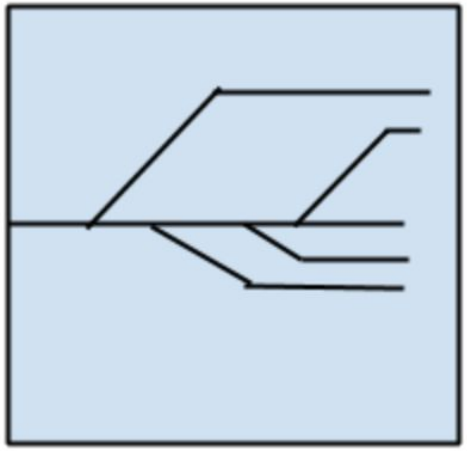
Regularization Algorithms

An extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing.

I have listed regularization algorithms separately here because they are popular, powerful and generally simple modifications made to other methods.

The most popular regularization algorithms are:

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)



Regularization Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

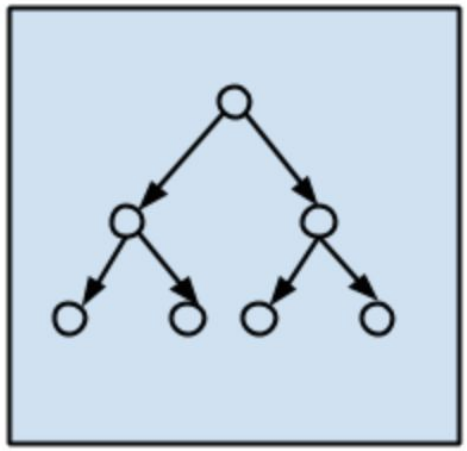
Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of attributes in the data.

Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees



Decision Tree Algorithms

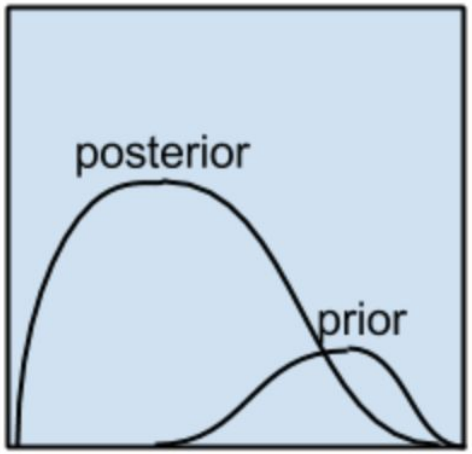
Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Bayesian Algorithms

Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression.

The most popular Bayesian algorithms are:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)



Bayesian Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

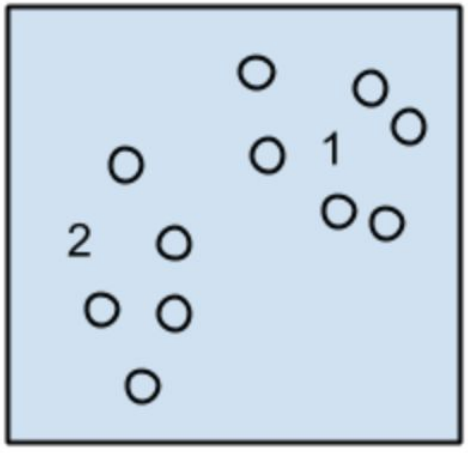
Clustering Algorithms

Clustering, like regression, describes the class of problem and the class of methods.

Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

The most popular clustering algorithms are:

- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering



Clustering Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

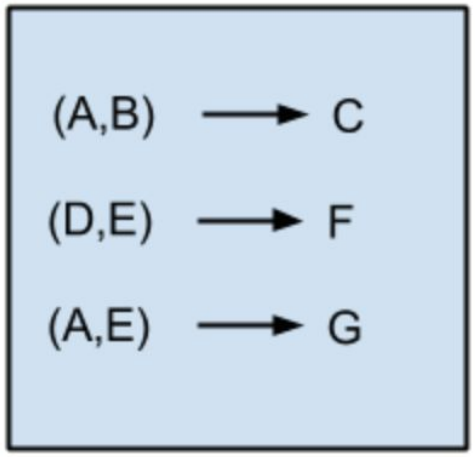
Association Rule Learning Algorithms

Association rule learning methods extract rules that best explain observed relationships between variables in data.

These rules can discover important and commercially useful associations in large multidimensional datasets that can be exploited by an organization.

The most popular association rule learning algorithms are:

- Apriori algorithm
- Eclat algorithm



Association Rule Learning Algorithms

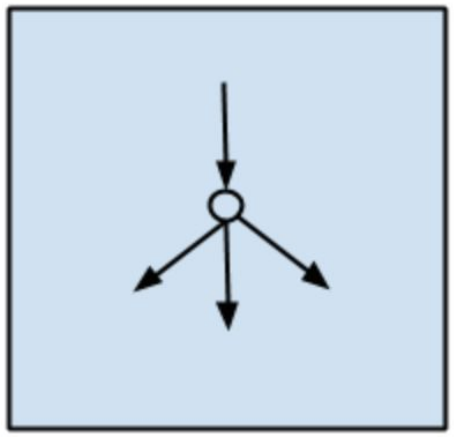
Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Artificial Neural Network Algorithms

Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks.

They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types.

Note that I have separated out Deep Learning from neural networks because of the massive growth and popularity in the field. Here we are concerned with the more classical methods.



Artificial Neural Network Algorithms

The most popular artificial neural network algorithms are:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

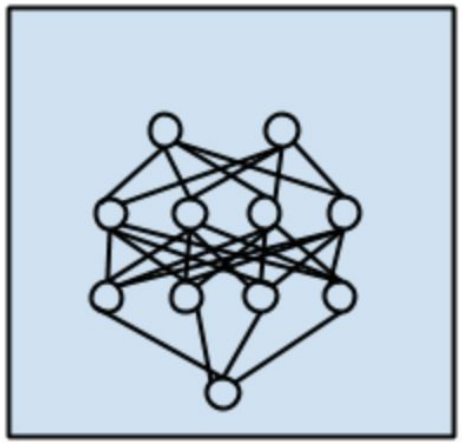
Deep Learning Algorithms

Deep Learning methods are a modern update to Artificial Neural Networks that exploit abundant cheap computation.

They are concerned with building much larger and more complex neural networks and, as commented on above, many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data.

The most popular deep learning algorithms are:

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders



Deep Learning Algorithms

Credit: Jason Brownlee, Machine Learning Mastery
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Popular in machine learning, not obvious in previous slides:

Support Vector Machines and Kernel Methods

These methods are discriminative models that seek a kernel, or high dimensional transformation of a space, for a decision boundary.

Graphical Models

These algorithms combine elements from probability theory and graph theory and utilize structural information encoded in the given graph. Home to a large number of Bayesian and causal models -- e.g., structural equation models, belief networks.

http://www.cis.upenn.edu/~mkearns/papers/barbados/jordan-tut.pdf

Generative Adversarial Networks

These algorithms utilize both a generative and discriminative model to ultimately create a model of the data's distribution, i.e., density estimation.

Questions?