What is Machine Learning?

- Stanford: Machine learning is the science of getting computers to act without being explicitly programmed.

What is Machine Learning?

- Wikipedia: Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed

What is Machine Learning?

- SAS: Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

What is Machine Learning?

- TechTarget: Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

Machine Learning = Data Mining?

NO

- Machine learning focuses on prediction, based on *known* properties learned from the training data.
- Data mining; which is the analysis step of Knowledge Discovery in Databases focuses on the discovery of (previously) *unknown* properties on the data.

So ML=AI?

YES

- Artificial intelligence is the general field of "intelligent-seeming algorithms" of which machine learning is the leading frontier at the moment.

Lets be practical !

- Machine learning in layman's words is Learning by a machine.
- How do we make machine learn by themselves ?
- By writing algorithms which can learn on their own by analysing the data they are fed with.
- Lets explore top 10 machine learning algorithms.

Machine Learning Terminologies

Accuracy (error rate)

- The rate of correct (incorrect) predictions made by the model over a data set (cf. coverage).
- Accuracy is usually estimated by using an independent test set that was not used at any time during the learning process.
- More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with data sets containing a small number of instances.

Source and Credits of ML Terminologies: http://robotics.stanford.edu/~ronnyk/glossary.html

Machine Learning Terminologies

Association learning

- Techniques that find conjunctive implication rules of the form ``X and Y implies A and B'' (associations) that satisfy given criteria.
- The conventional association algorithms are sound and complete methods for finding all associations that satisfy criteria for minimum support and minimum confidence

Machine Learning Terminologies

Categorical Values

- A finite number of discrete values. The type nominal denotes that there is no ordering between the values, such as last names and colors.
- The type ordinal denotes that there is an ordering, such as in an attribute taking on the values low, medium, or high.

Machine Learning Terminologies

Attribute (field, variable, feature)

- A quantity describing an instance. An attribute has a domain defined by the attribute type, which denotes the values that can be taken by an attribute. The following domain types are nominal, ordinal
- A feature is the specification of an attribute and its value. For example, color is an attribute. ``Color is blue'' is a feature of an example.

Feature vector (record, tuple)

- A list of features describing an instance.

Machine Learning Terminologies

Classifier

- A mapping from unlabelled instances to (discrete) classes.
- Examples:
  - K-Means
  - Nearest Neighbour
  - Decision Tree
  - And many more…

Machine Learning Terminologies

Confusion matrix

actual \ predicted	negative	positive
Negative	**a**	b
Positive	c	**d**

- The following terms are defined for a two by two confusion matrix:
  - Accuracy : (a+d)/(a+b+c+d).
  - True positive rate (Recall, Sensitivity): d/(c+d).
  - True negative rate (Specificity): a/(a+b).
  - Precision: d/(b+d).
  - False positive rate : b/(a+b).
  - False negative rate : c/(c+d).

Machine Learning Terminologies

Cost (utility/loss/payoff)

- A measurement of the cost to the performance task (and/or benefit) of making a prediction Y' when the actual label is y.

Cross-validation

- A method for estimating the accuracy (or error) of an inducer by dividing the data into k mutually exclusive subsets (the ``folds'') of approximately equal size.
- The inducer is trained and tested k times. Each time it is trained on the data set minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds.

Machine Learning Terminologies

Data set

- A schema and a set of instances matching the schema. Generally, no ordering on instances is assumed. Most machine learning work uses a single fixed-format table.

Dimension

- An attribute or several attributes that together describe a property.
- For example, a geographical dimension might consist of three attributes: country, state, city.
- A time dimension might include 5 attributes: year, month, day, hour, minute.

Inducer / induction algorithm

- An algorithm that takes as input specific instances and produces a model that generalizes beyond these instances.

Instance (example, case, record)

- A single object of the world from which a model will be learned, or on which a model will be used (e.g., for prediction).
- In most machine learning work, instances are described by feature vectors;
- some work uses more complex representations (e.g., containing relations between instances or between parts of instances).

Machine Learning Terminologies

Model

- A structure and corresponding interpretation that summarizes or partially summarizes a set of data, for description or prediction. Most inductive algorithms generate models that can then be used as classifiers, as regressors, as patterns for human consumption, and/or as input to subsequent stages of the KDD process.

Machine Learning Terminologies

Sensitivity

- True positive rate (see Confusion matrix).

Specificity

- True negative rate (see Confusion matrix).

Machine Learning Algorithms

Machine learning algorithms can be divided into 3 broad categories

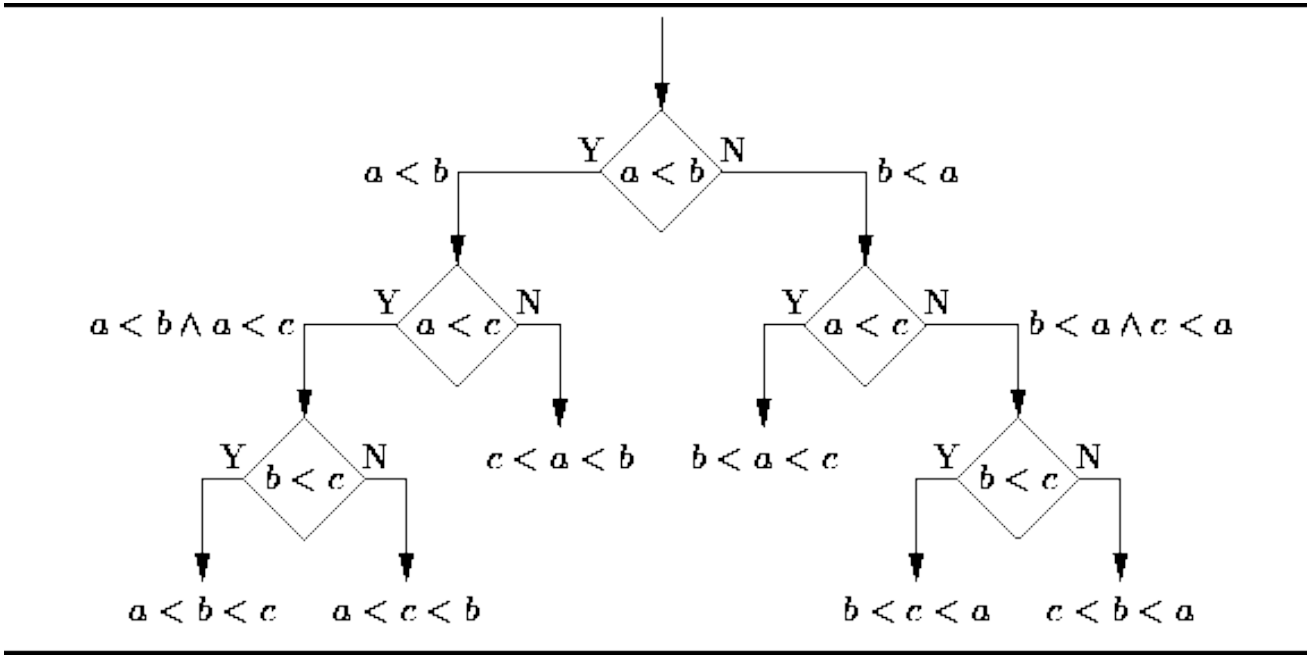
- supervised learning
- unsupervised learning
- reinforcement learning

Machine Learning Algorithms

- Supervised learning : Where a property (*label*) is available for a certain dataset (*training set*), but is missing and needs to be predicted for other instances.
- Unsupervised learning: Where the challenge is to discover implicit relationships in a given *unlabelled* dataset (items are not pre-assigned).
- Reinforcement learning: Concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward by learning in trial and error fashion.

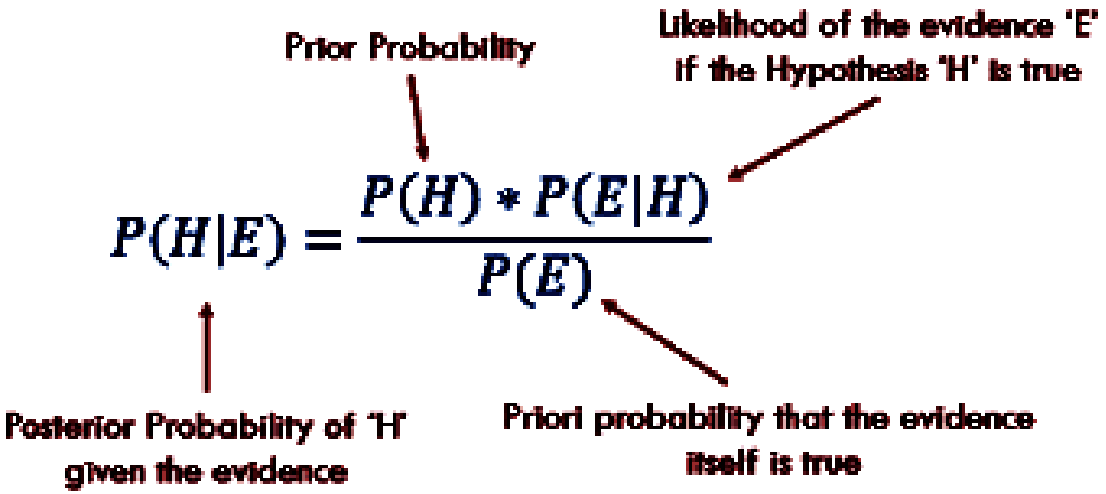
Algorithm-1 (Decision Trees)

A tree-like graph or model of decisions and their possible consequences.



Algorithm-2 (Naïve Bayes Classification)

Based on simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

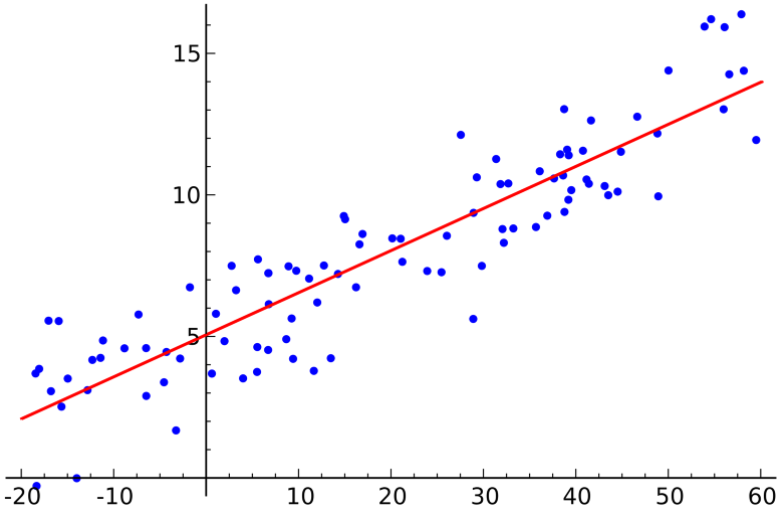


Some of real world examples:
1) To mark an email as spam or not spam
2) Classify a news article about technology, politics, or sports

Algorithm-3 (Ordinary Least Squares Regression)

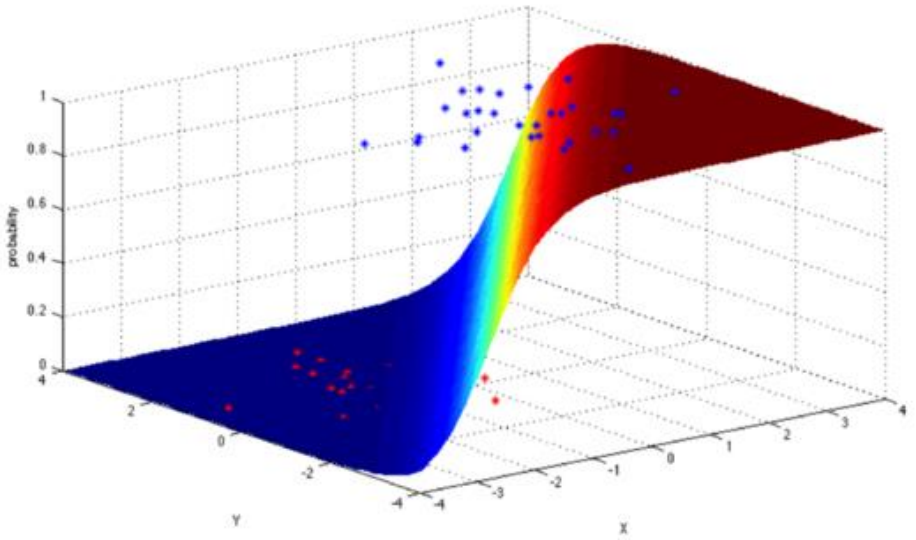
A method performing linear regression (task of fitting a straight line through a set of points).

Strategy is draw a line, and then for each of the data points, measure the vertical distance between the point and the line, and add these up; the fitted line would be the one where this sum of distances is as small as possible.



Algorithm-4 (Logistic Regression)

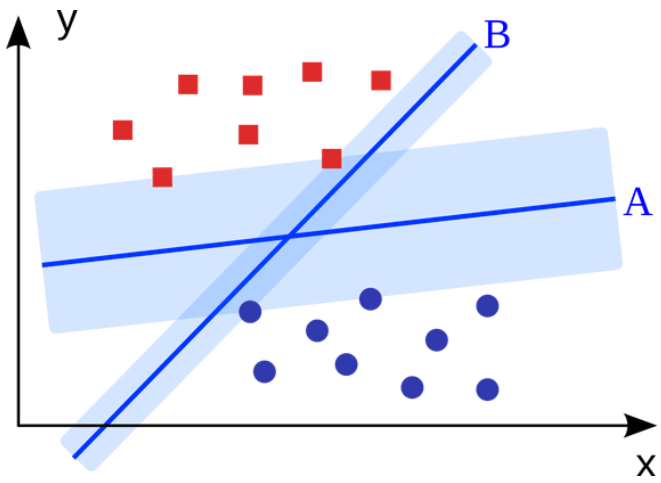
Modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.



E.g. Do body weight calorie intake, fat intake, and age have an influence on heart attacks (yes vs. no)?

Algorithm-5 (Support Vector Machines)

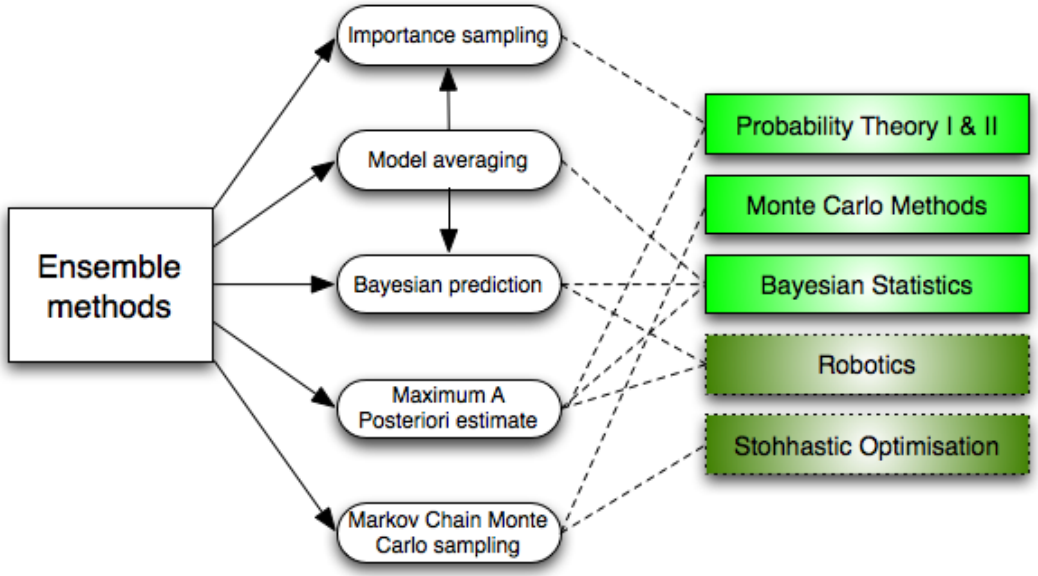
- A binary classification algorithm, given a set of points of 2 types in N dimensional place, SVM generates a (N — 1) dimensional hyperlane to separate those points into 2 groups.
- 2 set of points on a paper, linearly separable. SVM will find a straight line which separates those points into 2 types and situated as far as possible from all those points.



E.g. Human splice site recognition, image-based gender detection, large-scale image classification

Algorithm-6 (Ensemble Methods)

- Learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions.
- Ability to average out biases, reduce variance, unlikely to over-fit



Algorithm-7(Clustering Algorithms)

- Task of grouping a set of objects such that objects in the same group (*cluster*) are more similar to each other than to those in other groups.
- Different Clustering Algorithms

  Centroid-based algorithms

  Connectivity-based algorithms

  Density-based algorithms

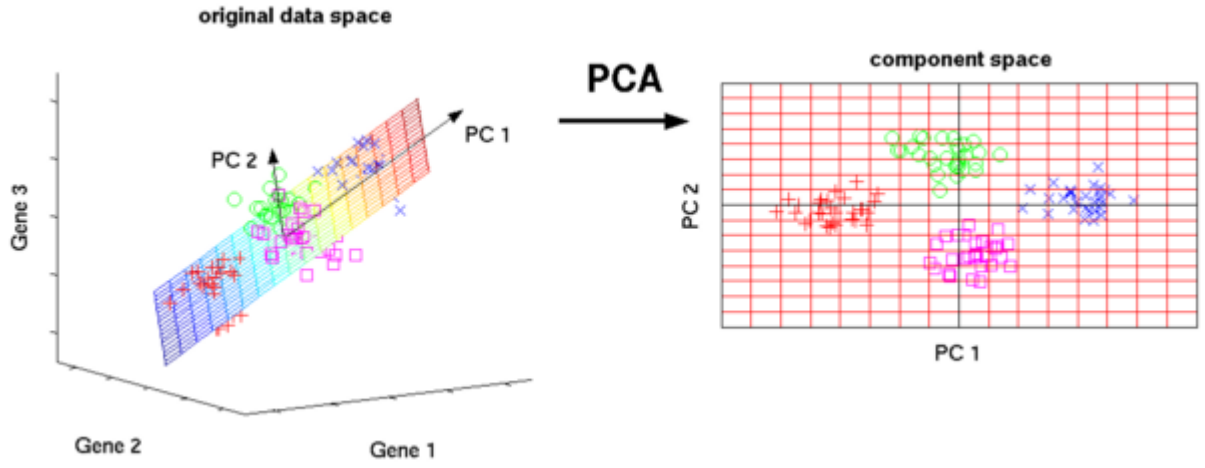
  Probabilistic

  Dimensionality Reduction

  Neural networks / Deep Learning

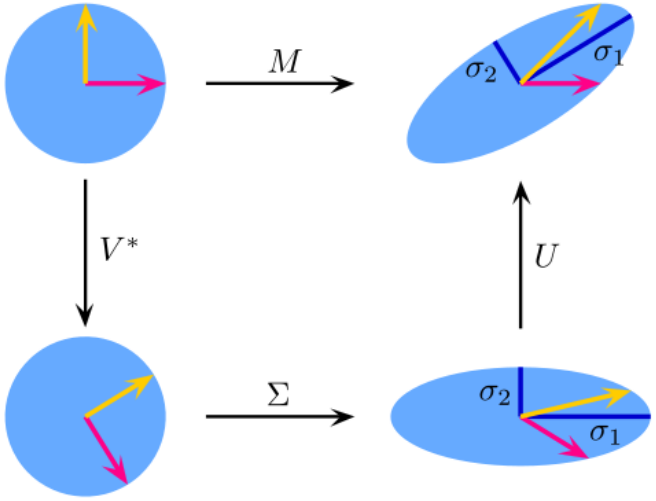
Algorithm-8(Principal Component Analysis)

- Statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Domain knowledge is very important while choosing whether to go forward with PCA or not. It is not suitable in case of noisy data.



Algorithm-9(Singular Value Decomposition)

- A linear algebra, a factorization of a real complex matrix.
- For a given *m* * *n* matrix M, there exists a decomposition such that M = UΣV, where U and V are unitary matrices and Σ is a diagonal matrix.



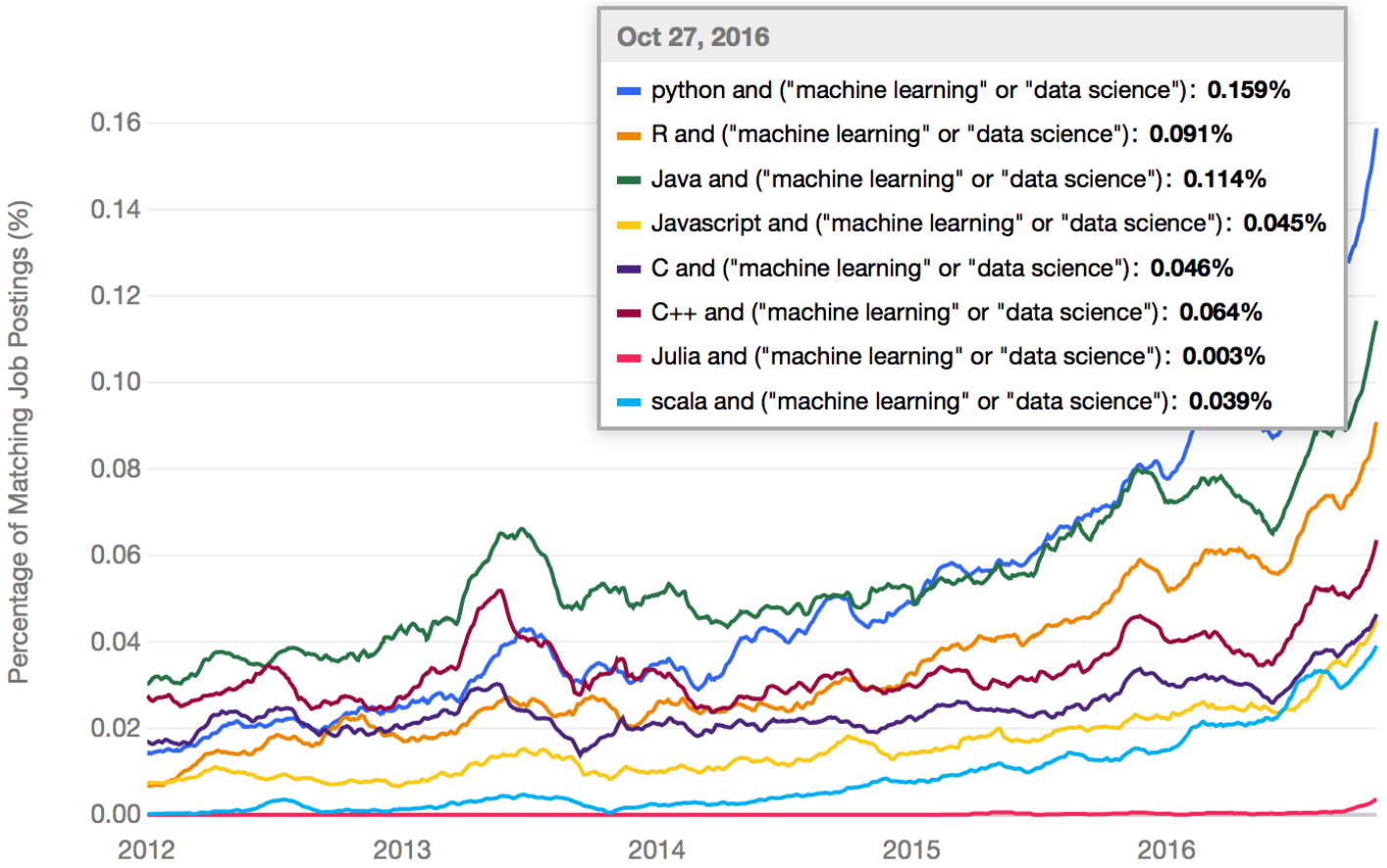
$$M = U \cdot \Sigma \cdot V^*$$

Algorithm-10 (Independent Component Analysis)

- Statistical technique for revealing hidden factors that underlie sets of random variables, measurements, or signals defining a generative model for the observed multivariate data, which is typically given as a large database of samples
- ICA is related to PCA, but it is a much more powerful technique that is capable of finding the underlying factors of sources when these classic methods fail completely.
- Applications include digital images, document databases, economic indicators and psychometric measurements.

Programming Language for ML?

#1 Python, #2 Java, #3 R, #4 C++, #5 C



Current trends in ML

- Neural Networks
- Deep Learning ! (incredibly successful cases!!)

Examples of Machine Learning:

Google News

Facebook Trending Topics

Google's AlphaGO (Deep learning based)

Do check this news article on TechTimes.com:
5 Ways Artificial Intelligence Freaked Us Out In 2016: WaveNet, AlphaGo, Interceptor And More

Sources of content in this PPT

1. https://www.quora.com/What-is-the-connection-between-data-science-and-artificial-intelligence-Is-it-machine-learning
2. https://www.quora.com/What-is-the-difference-between-Data-Analytics-Data-Analysis-Data-Mining-Data-Science-Machine-Learning-and-Big-Data-1
3. http://www.kdnuggets.com/2016/05/machine-learning-key-terms-explained.html/2
4. http://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html
5. http://www.kdnuggets.com/2016/12/machine-learning-artificial-intelligence-main-developments-2016-key-trends-2017.html
6. http://www.kdnuggets.com/2016/12/data-science-machine-learning-cheat-sheets-updated.html
7. http://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html
8. http://www.kdnuggets.com/2016/10/top-10-data-science-videos-youtube.html/2