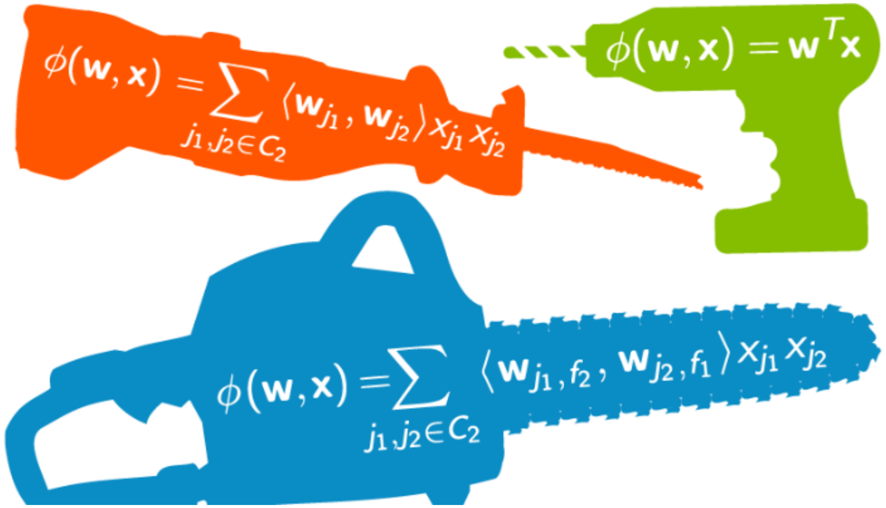
Machine Learning in the Scientific World

By:
Sanchit Alekh
Hilfswissenschaftler
Lehrstühl Informatik 5

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Contents


$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1} \rangle x_{j_1} x_{j_2}$

Sanchit Alekh

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

“A BREAKTHROUGH in MACHINE LEARNING would be worth TEN MICROSOFTS”

-Bill Gates

What is Machine Learning ?

*"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."*

*"Machine learning explores the **study and construction** of algorithms that can **learn from** and **make predictions** on data."*

OWN INSIGHT:

"Machine Learning is an extremely close abstraction of human learning, but is pursued in a planned, algorithmic and machine-representable manner."

Why is it so relevant today ?

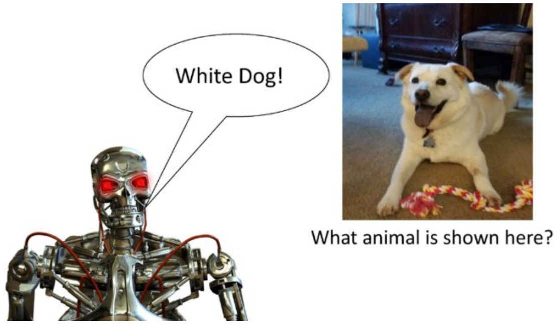
- **Data Explosion**
  Due to the sheer explosion in the volume and velocity of data, manual methods for data analysis are infeasible in today's world

- **Availability of Computing Power**
  Modern computers are far more capable than they used to be. This has made complex calculations possible within times that would not have been possible before

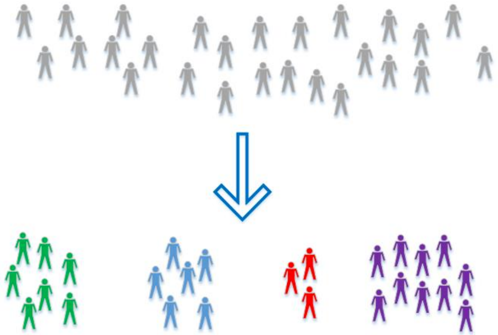
- **Man's Quest to Build Intelligent Machines**
  From self-driving cars to intelligent robotics, Machine Learning has brought Artificial Intelligence closer to Human Cognition

Types of Learning Methods

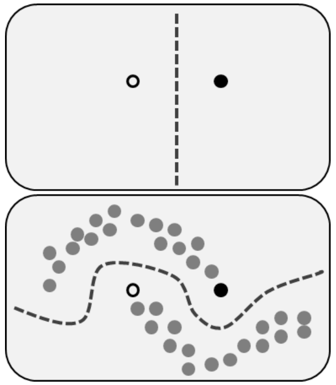
1. **Supervised learning**
   - trained using labeled examples (inputs where the desired output is known)
   - Compares actual output with correct outputs to find errors
   - modifies the model accordingly to reduce the errors



2. **Unsupervised learning**
   - used against data that has no historical labels
   - algorithms must figure out what is being shown
   - goal is to explore the data and find some structure within.



3. **Semi-supervised learning**
   - uses both labeled and unlabeled data for training
   - typically a small amount of labeled data with a large amount of unlabeled data
   - careful analysis needed to understand how the unlabeled data can help



Some pointers while using ML Algorithms

RWTH AACHEN UNIVERSITY

1. **ML Algorithms are not a 'Black Box'**
   - Most applications that use ML have to modify the parameters/bias values/error functions etc. to meet the requirements of the problem
   - Algorithms need to be very carefully selected based on type of task

2. **Experimentation is the norm**
   - Even experts in the field can not handpick a specific algorithm for a task.
   - Human intuition is inaccurate as it deals with high-dimensional data
   - Co-relation between features not always visible

3. **Feature Engineering is of Paramount Importance**
   - Features used in the ML Algorithm are make-or-break for the implementation
   - Considerable amount of time must be spent
   - Domain knowledge must be applied to get the best possible features
   - Know your data well
   - Feature Engineering is an Art

mi-Mapa

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Sanchit Alekh

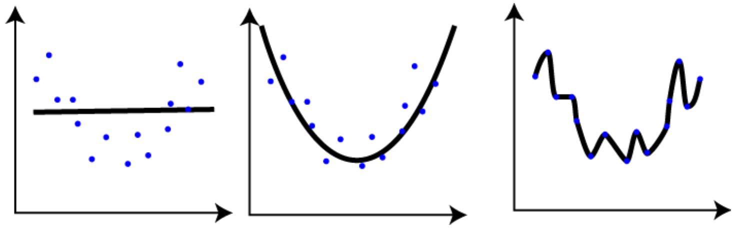
Coming up with features is **DIFFICULT**, **TIME-CONSUMING**, requires **EXPERT KNOWLEDGE**. Applied machine learning is basically **FEATURE ENGINEERING**.

— *Andrew Ng*

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Some pointers while using ML Algorithms

4. **Overfitting is a Major Problem**
   - What if the knowledge and data we have are not sufficient to completely determine the correct classifier?
   - We run the risk of hallucinating a classifier (or parts of it) that is not grounded in reality, and is simply encoding random quirks in the data.
   - Error Analysis should be performed to check for high bias (under-fitting) and high variance (over-fitting).
   - In most scenarios, we would prefer Classifier #2 over Classifier #1
   - Aim of Machine Learning is 'Generalization'



	Accuracy(Training Data)	Accuracy(Testing Data)
Classifier #1	100%	50%
Classifier #2	75%	75%

Some pointers while using ML Algorithms

5. **Generalization is Difficult**
   - Generalizing correctly becomes exponentially harder as the dimensionality of the examples grows
   - a moderate dimension of 100 and a huge training set of a trillion examples, the latter covers only a fraction of about 10^{-18} of the input space.
   - Sometimes the benefits of extra features are outweighed by the 'Curse of Dimensionality'
   - FEATURE ENGINEERING IS PARAMOUNT!

6. **More Data is better than a Powerful Algorithm**
   - Pragmatically the quickest path to success is often to just get more data.
   - As a rule of thumb, a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it.

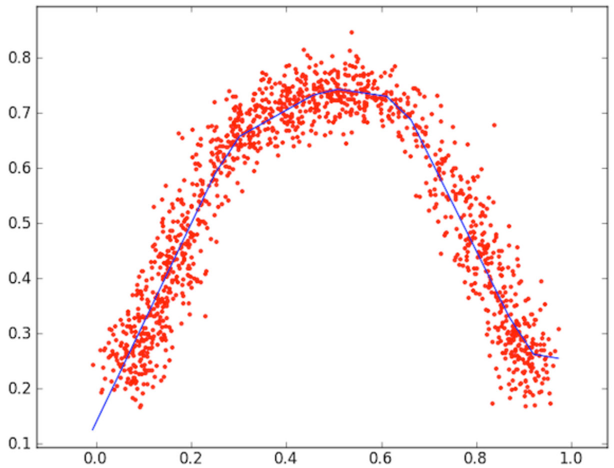
7. **Ensemble Methods are becoming the Standard**

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

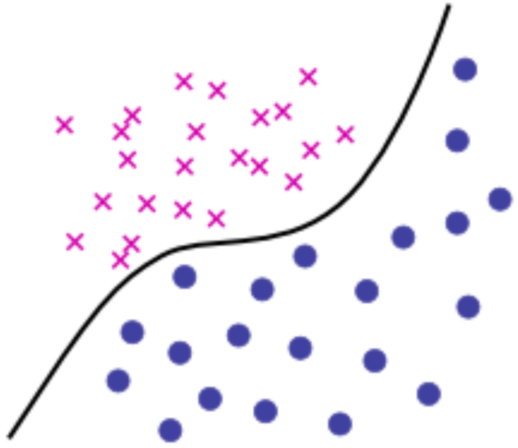
Major Classes of Learning

Regression



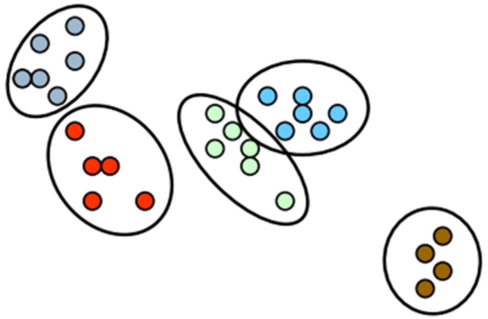
- Target variable is continuous or ordered whole values
- Typically supervised
- E.g. Stock-market predictions

Classification



- Target variable is discrete and categorical
- Typically supervised
- E.g. Labelling tweets as positive, negative or neutral

Clustering



- There are no target values
- Data is aggregated into groups without any labels
- Typically unsupervised or semi-supervised
- E.g. Author Disambiguation

Sanchit Alekh

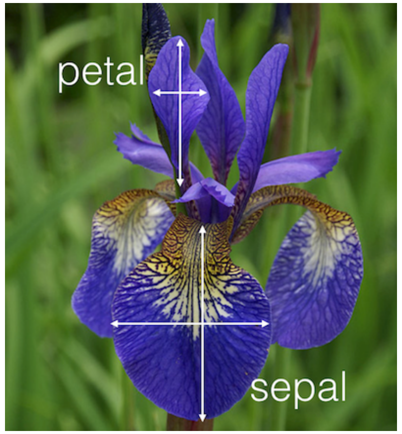
Popular ML-Algorithms

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Naïve Bayes Classification

Bayes Theorem: $$P(\omega_j \mid \mathbf{x}_i) = \frac{P(\mathbf{x}_i \mid \omega_j)\ P(\omega_j)}{P(\mathbf{x}_i)}$$

- Based on the Bayes' Theorem of Conditional Probabilities
- One of the most used classification algorithms
- Is easy to understand and implement
- Assumes conditional independence (naïve) between features
- Has proven to work well in practice for small datasets even when the features are correlated.



- E.g., on the UCI Iris Dataset
- Problem formulated like, $P(\text{Setosa}|x_i)$, where $x_i = [4.5 \text{ cm}, 7.4 \text{ cm}]$
- the decision rule is:

  class label $w_j \longleftarrow \text{argmax}_{i=1,2..m} P(w_j \mid x_i)$, where j {Setosa, Versicolor, Virginica}

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

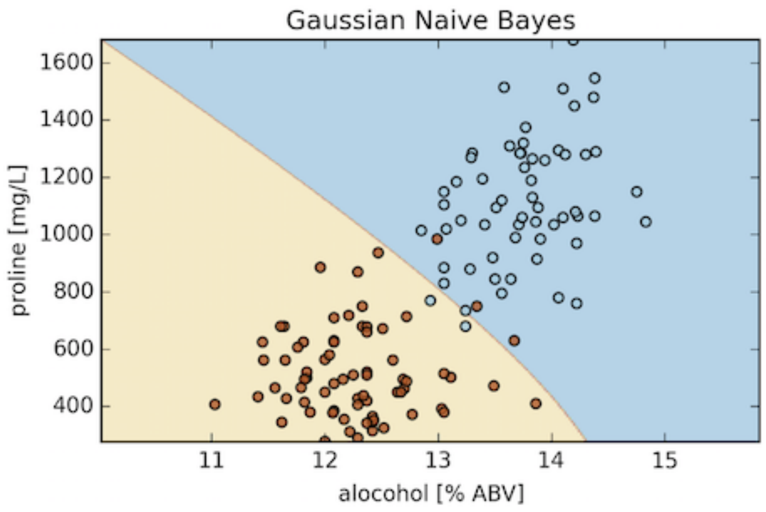
Naïve Bayes Classification

ADVANTAGES

- performs well when the input variables are categorical
- converges faster, requiring relatively little training data than other discriminative models like logistic regression
- easier to predict class of the test data set.
- Though it requires conditional independence assumption, Naïve Bayes Classifier has presented good performance in various application domains.

USES

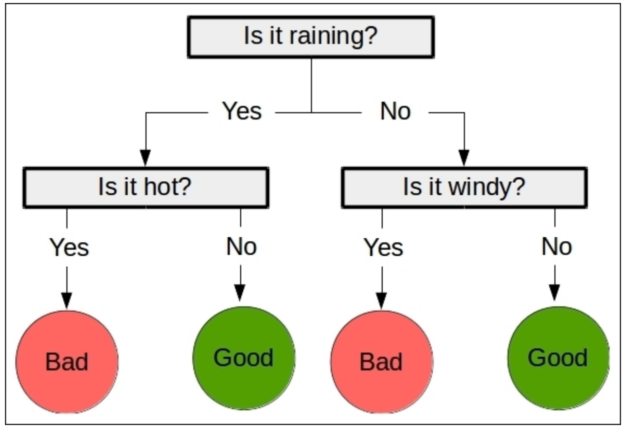
- Document Categorization
- Spam Filtering
- Sentiment Analysis



Sanchit Alekh

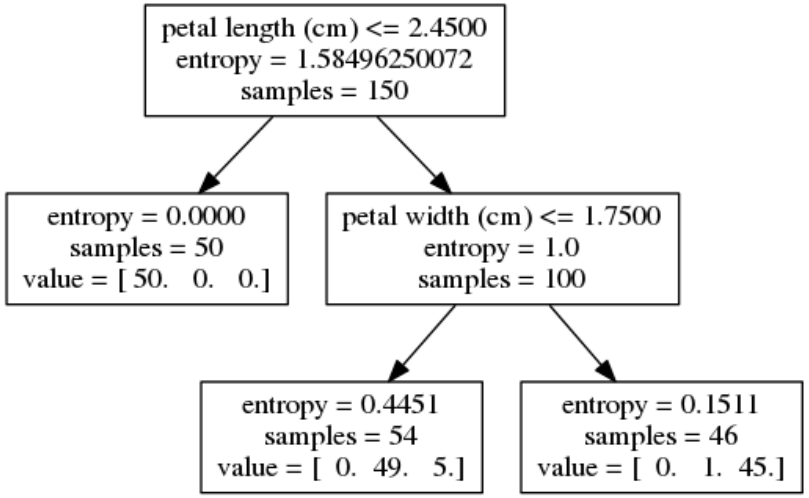
Decision Trees

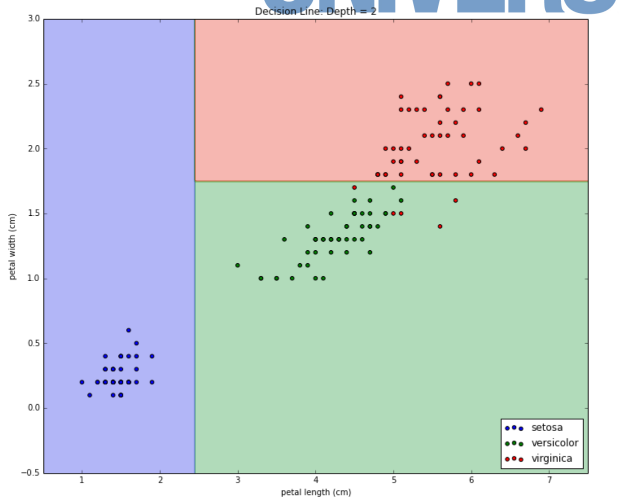
- Decision Trees try to segment the data iteratively to get a tree structure
- Searches through each independent variable to find the single variable that best splits the data into two or more groups
- Typically, the best split minimizes the impurity of the outcome in the resulting datasets
- Split criterion is based on Information Theoretic Models such as Entropy & Information Gain
- The split is performed repeatedly until a stopping criteria is invoked
- Have a geometric decision boundary
- Most popular algorithms: C4.5, C5.0, ID3, CART

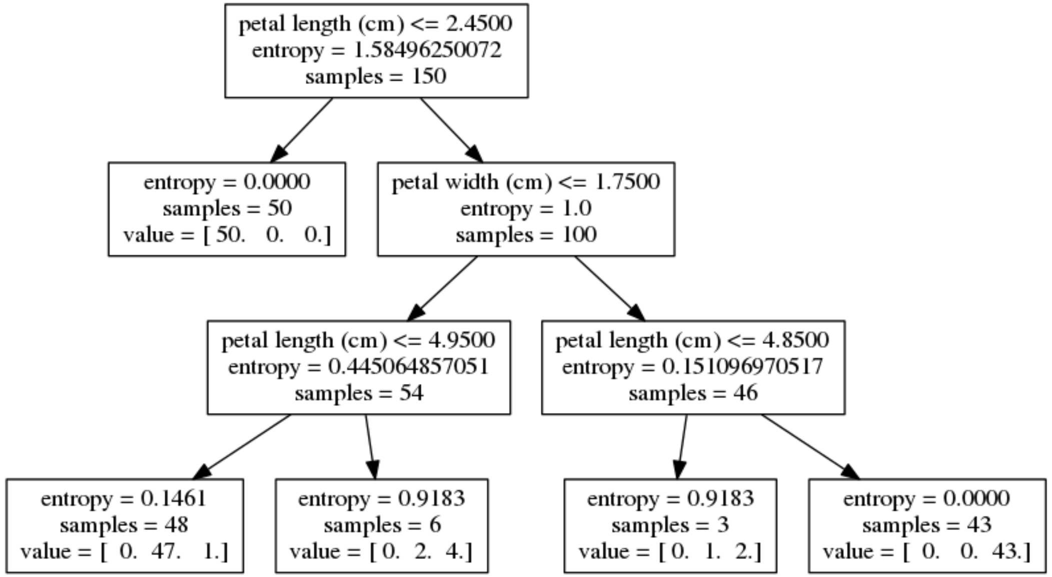


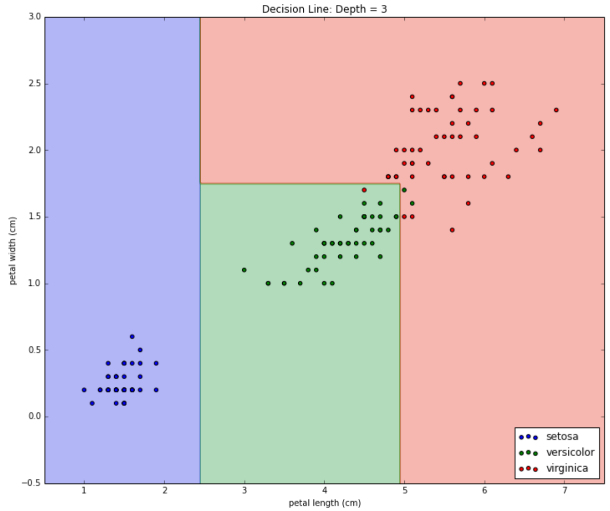
Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Decision Trees on the Iris Dataset



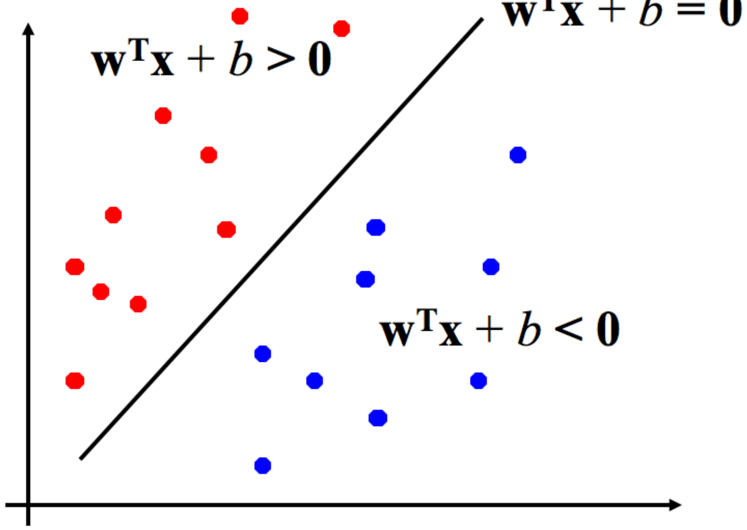


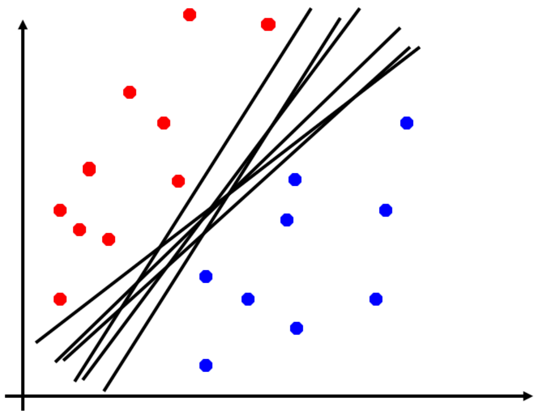


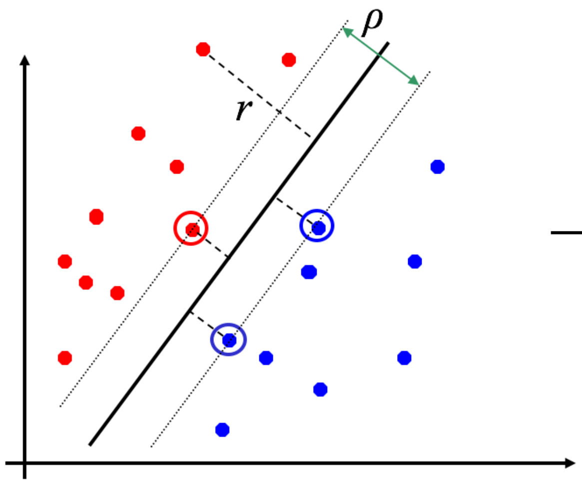


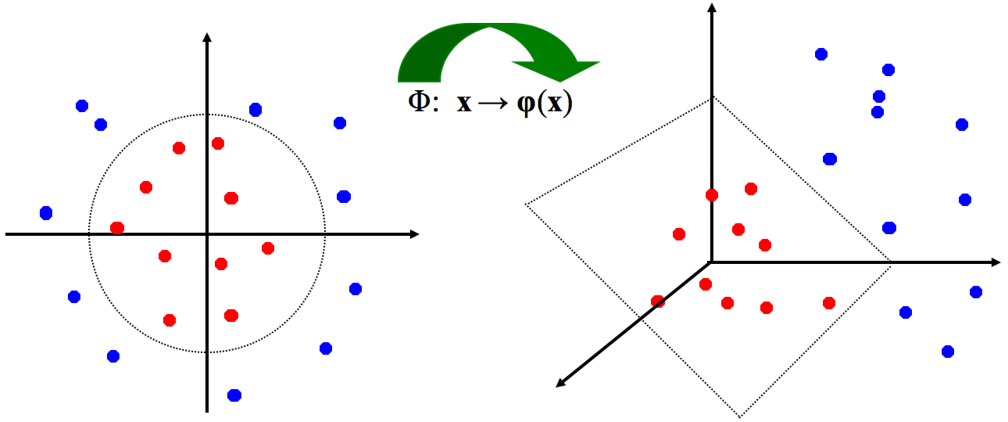
Support Vector Machines

- Is a class of Linear Separators
- It is a supervised learning technique which computes a separating hyperplane in n-dimensional space
- Is based on the concept of 'Support Vectors', i.e. points closest to the hyperplane, and margin, i.e. the distance between support vectors
- Highly effective when the number of dimensions is large
- Is not prone to overfitting of training data



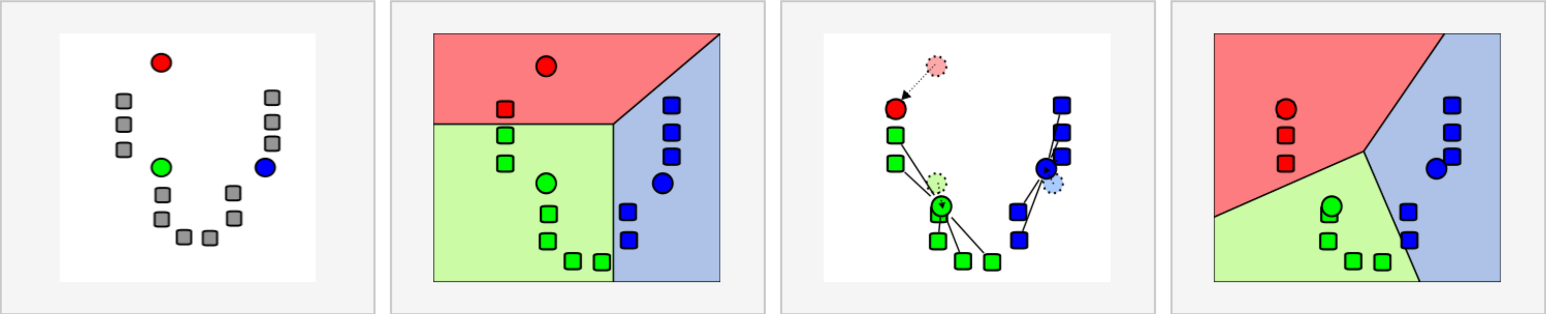






Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

k-means Clustering



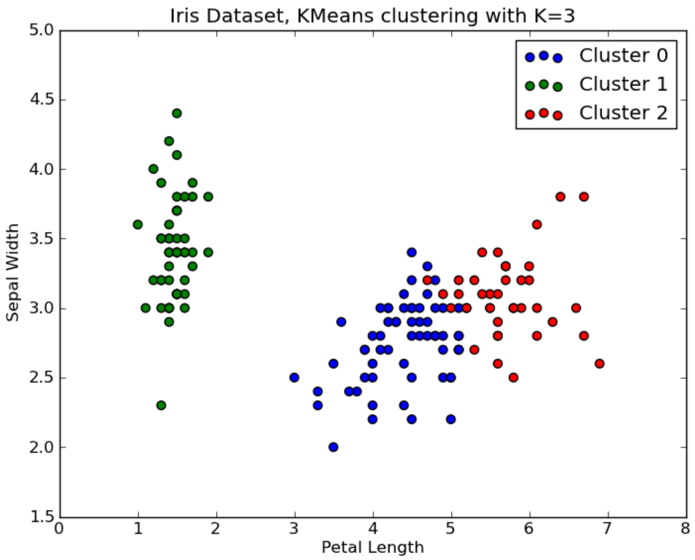
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

- One of the simplest unsupervised learning algorithms to solve the clustering problem.
- Number of clusters has to be fixed apriori
- Main idea is to define k centroids, one for each cluster.
- Extremely useful when the number of clusters are known
- Starting 'means' have a huge impact on the accuracy

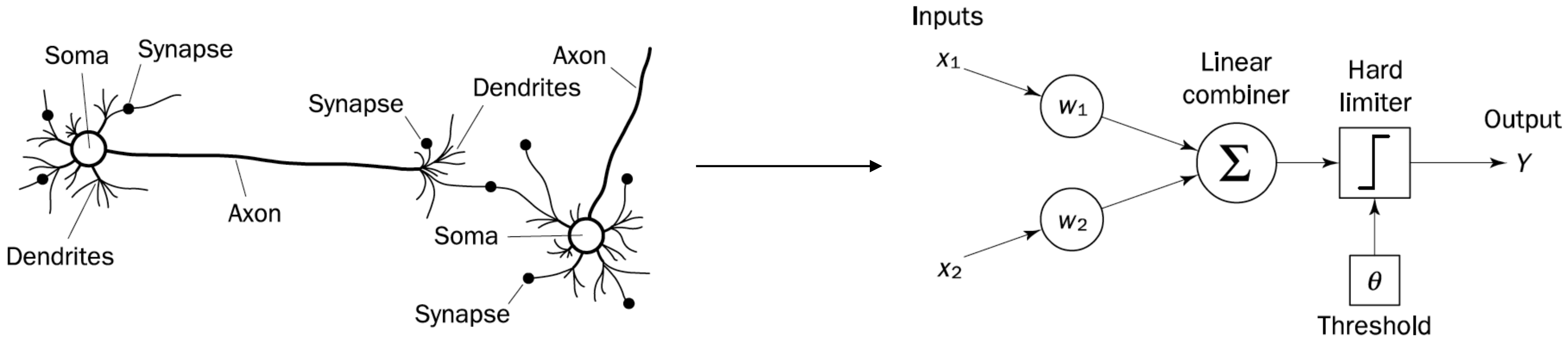


Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Sanchit Alekh

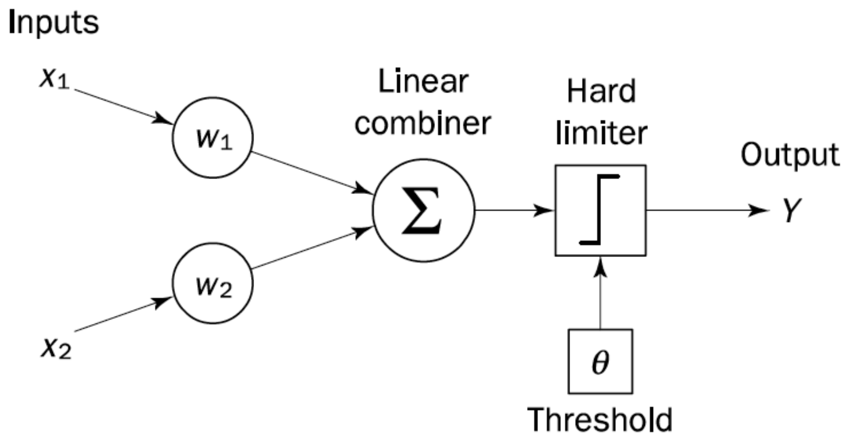
Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Artificial Neural Networks



- ANNs consist of a number of very simple and highly interconnected processors (neurons), analogous to the biological neurons in the brain
- Each neuron receives a number of input signals through its connections; however, it never produces more than a single output signal.
- The output signal is transmitted through the neuron's outgoing connection
- The outgoing branches terminate at the incoming connections of other neurons in the network.

ANN: McCulloch-Pitts' Model



Biological neural network	Artificial neural network
Soma	Neuron
Dendrite	Input
Axon	Output
Synapse	Weight

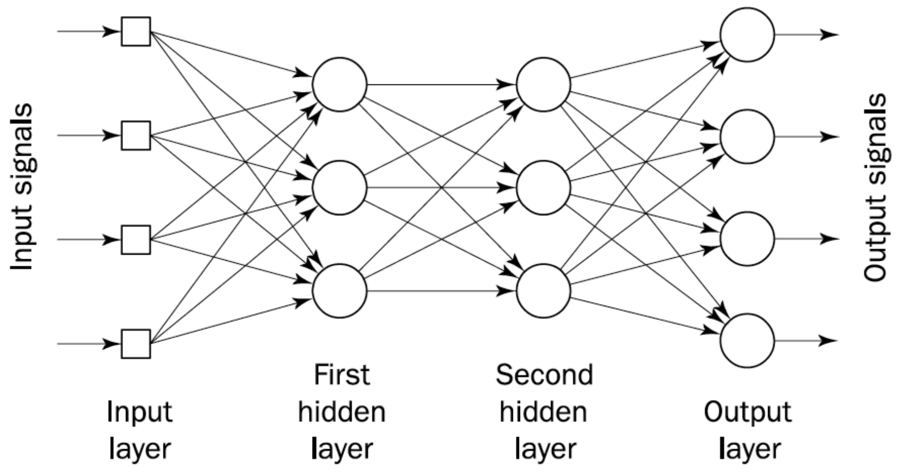
- In the McCulloch-Pitts' Model, the transfer function is typically a linear combiner
- The Activation function is the sign/step function of the linear sum minus threshold
- It uses the perceptron learning rule

Activation: $Y(p) = step\left[\sum_{i=1}^{n} x_i(p)w_i(p) - \theta\right],$ where n is the number of perceptron units

Weight Update: $w_i(p+1) = w_i(p) + \Delta w_i(p),$ where $\Delta w_i(p) = \alpha \times x_i(p) \times e(p)$

ANN: Multi-layer Perceptron with Backpropagation

- The McCulloch-Pitts' model is useful only when decision boundary is linear
- It can not learn more complex polynomial or other functions
- To learn more complicated functions, we introduce 'Hidden Layers'
- Neural Networks with >5 hidden layers can learn (in theory) any function

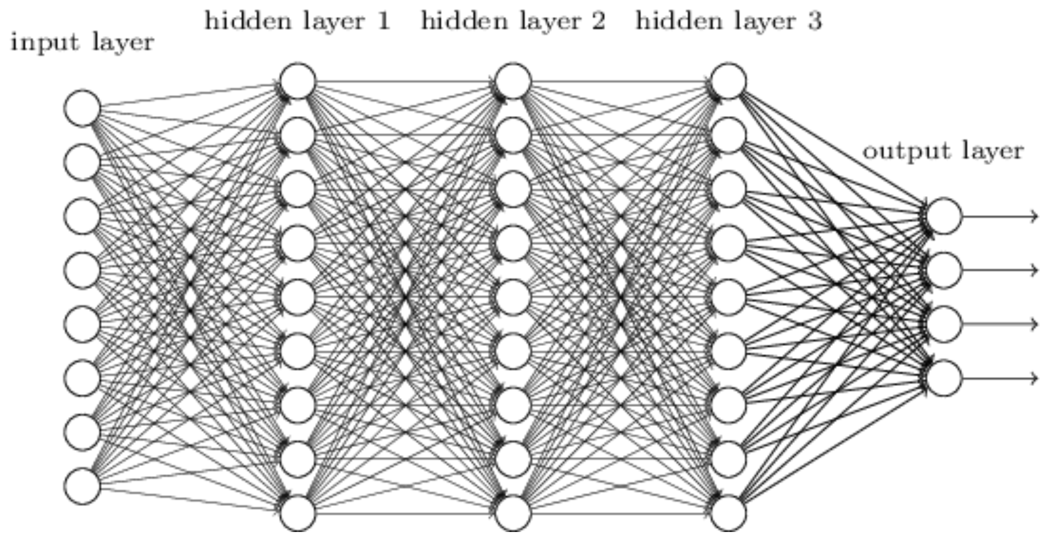


- Activation Function: Sigmoid/Tan Hyperbolic
- Learning Rule : Error Backpropagation with gradient
- Has been used successfully for a lot of complex learning tasks
- Convergence is difficult when the number of features rise
- VANISHING GRADIENT PROBLEM!

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Vanishing Gradient Problem

- Backpropagation becomes slower and slower as the neural networks become dense.
- This is because of the Vanishing Gradient Problem
- The error gradients are generally <1, and successively multiplying them for backpropagation yields a miniscule delta value for the input layers in the beginning



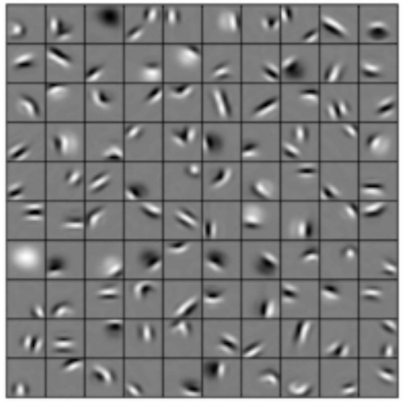
- This means that the layers in the beginning take an extremely long time to adjust their weights

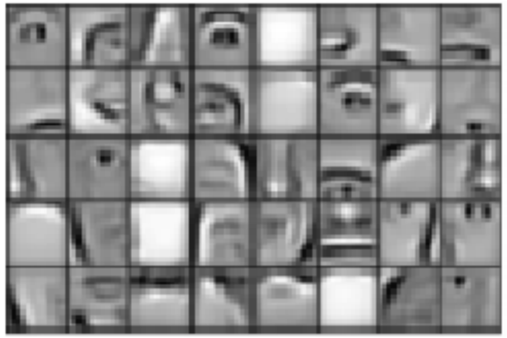
SOLUTION?? DEEP NEURAL NETWORKS

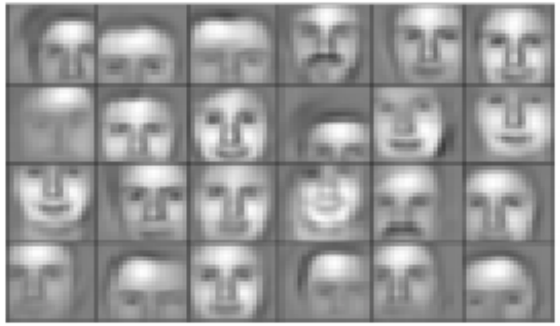
Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

Why are Deep Networks necessary?

- Deep Learning is the new revolution in Machine Learning
- It marks a paradigm shift from other ML Algorithms
- It solves the Vanishing Gradient Problem
- It is completely unsupervised
- It learns in steps, e.g. to recognize a face, it first learns to recognize edges, then local features, and then subsequently the entire face
- It is making things like 'Driverless Cars' a real phenomenon
- Deep Learning Algorithms: In the next presentation ☺



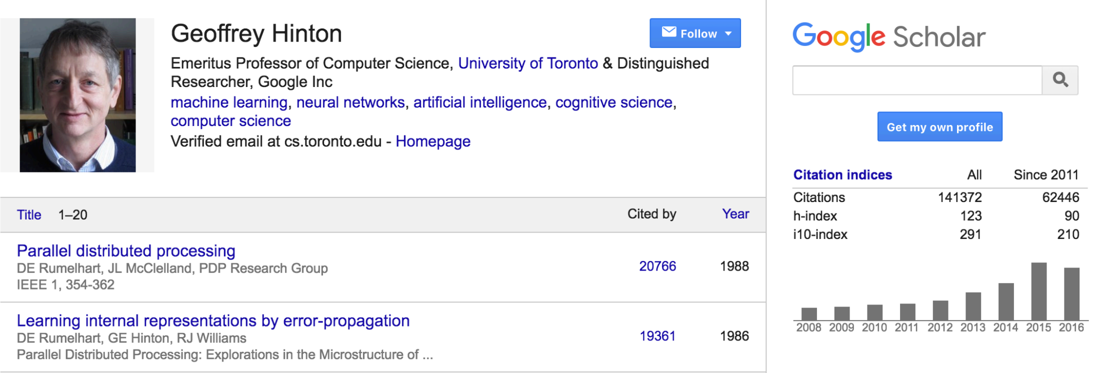




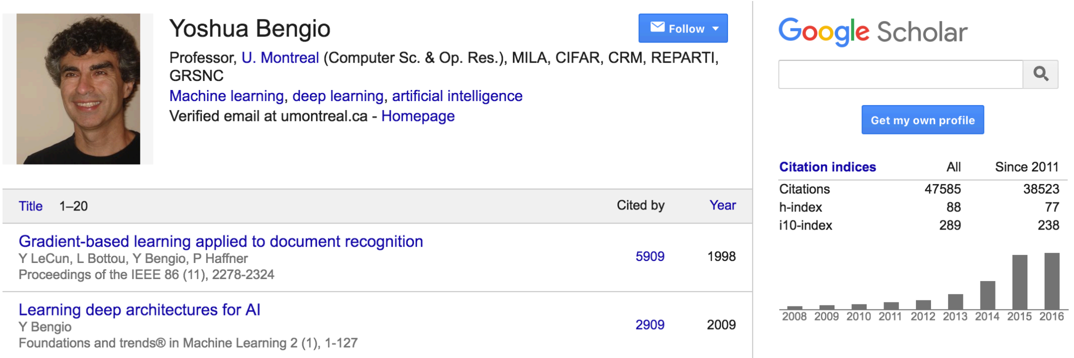
Sanchit Alekh

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen

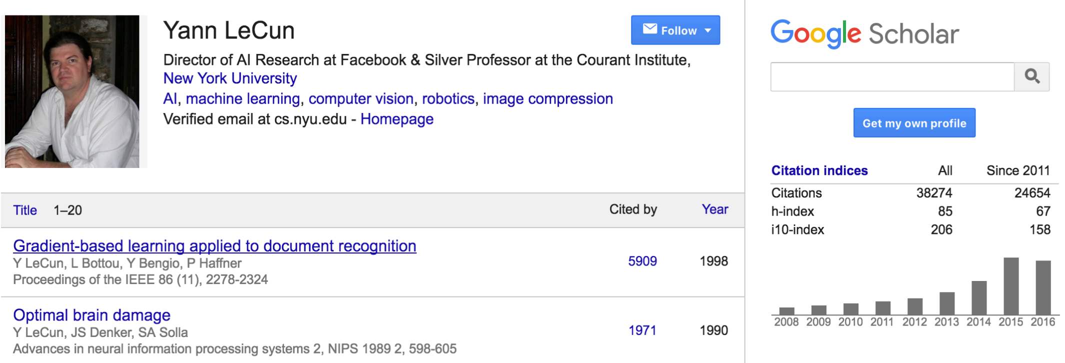
Scientists working on Deep Learning



141372 total citations
62446 since 2011



47585 total citations
38523 since 2011



38274 total citations
24654 since 2011

Machine Learning is Everywhere

RWTH AACHEN UNIVERSITY

- Web Search
- Recommender Systems
- Credit Scoring
- Fraud Detection
- Stock Trading
- Drug Design
- Driverless Cars
- Text Analysis
- Text/Media Labeling
- .... and many more

Sanchit Alekh

mi-Mapa

Prof. M. Jarke
Lehrstühl Informatik 5
RWTH Aachen