

Foundations of Data Science, Fall 2020

1. Introduction: Machine Learning

Prof. Dan Olteanu



Sept 15, 2020

<https://lms.uzh.ch/url/RepositoryEntry/16830890400>



<https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09>

Machine Learning in Action



(Using <https://www.betafaceapi.com/demo.html>)

Machine Learning in Action



(Using <https://www.betafaceapi.com/demo.html>)

Machine Learning in Action



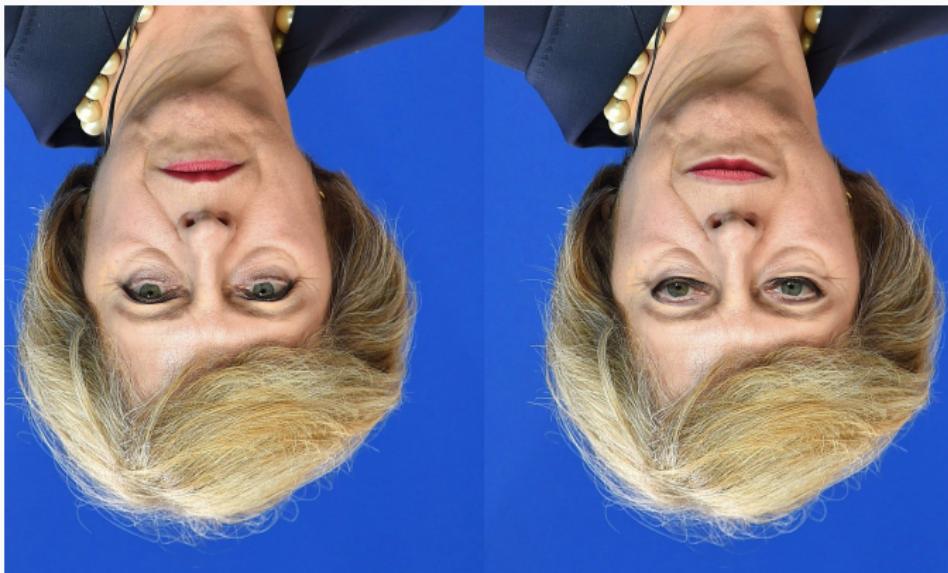
X
age: 19, beard: no, expression: other,
gender: female, glasses: no, mustache: no,
race: white,



X
age: 23, beard: no, expression: other,
gender: female, glasses: no, mustache: no,
race: white,

(Using <https://www.betafaceapi.com/demo.html>)

Is anything wrong?

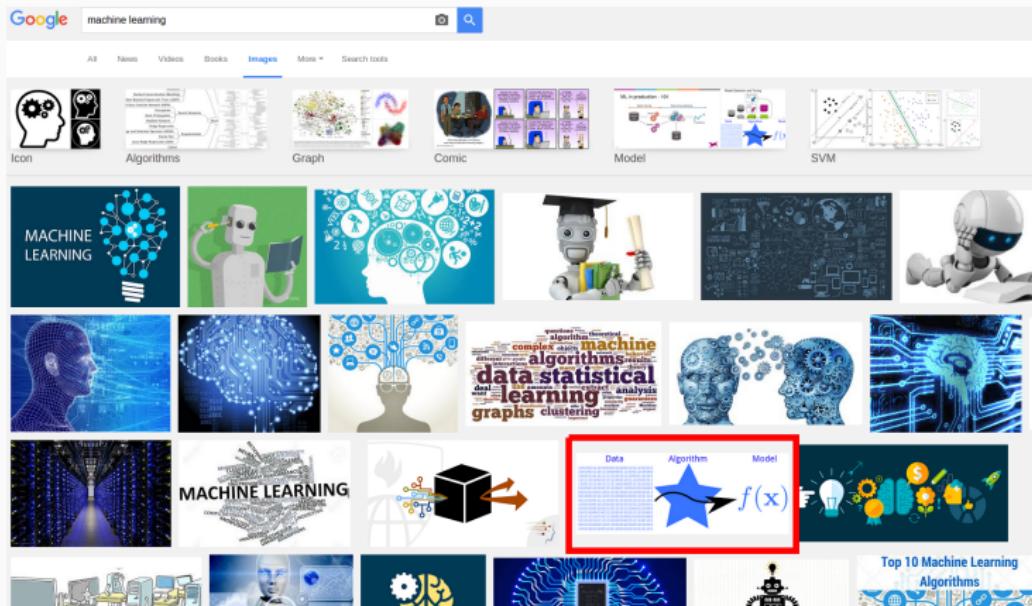


Is anything wrong?



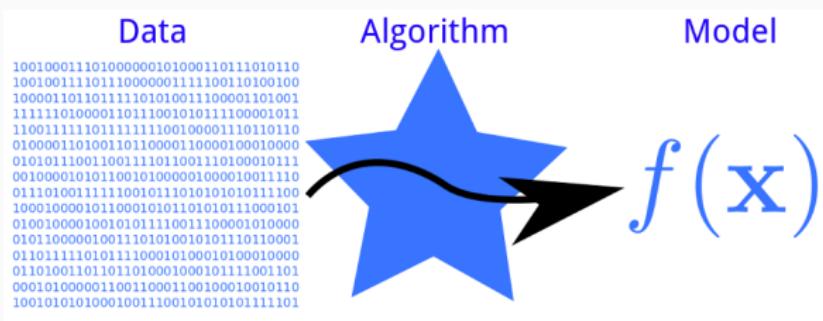
(See [Guardian article](#))

What is machine learning?



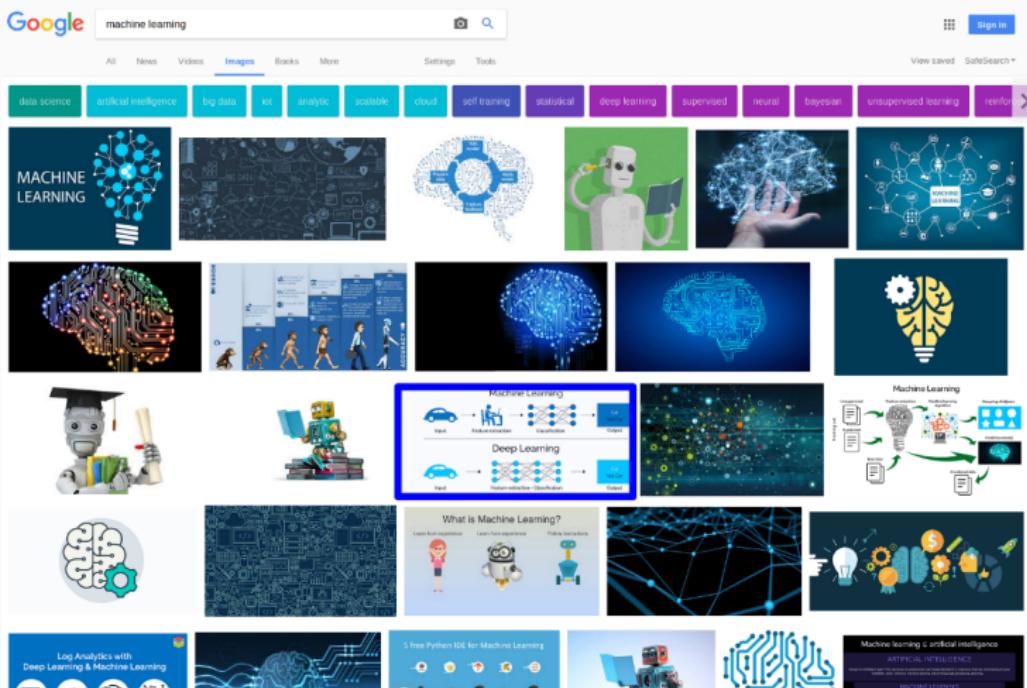
circa October 2016

What is machine learning?



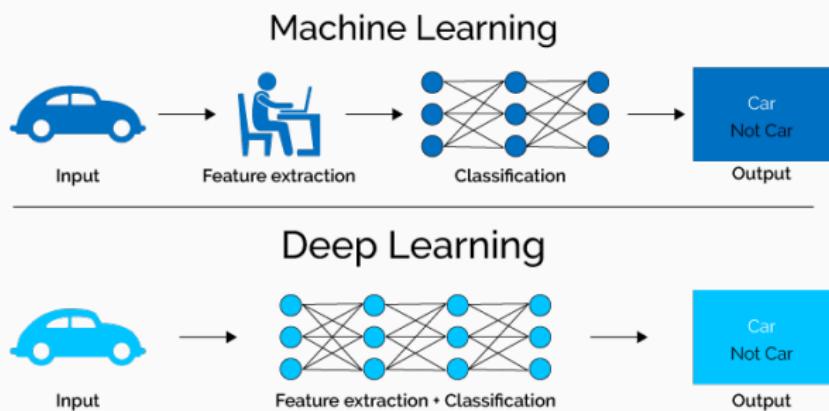
circa October 2016

What is machine learning?



circa October 2017 ([Le Cunn et al., Deep Learning, Nature \(2015\)](#))

What is machine learning?



circa October 2017 ([Le Cunn et al., Deep Learning, Nature \(2015\)](#))

What is machine learning?

What is artificial intelligence?

What is machine learning?

What is artificial intelligence?

“Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child’s? If this were then subjected to an appropriate course of education one would obtain the adult brain.”



Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.

Programming vs Learning

Programming, like all engineering, is a lot of work:

We have to build everything from scratch.

Learning is more like farming, which lets nature do most of the work. Farmers combine seeds with nutrients to grow crops.

Learners combine knowledge with data to grow programs.

What is machine learning?

Definition by Tom Mitchell (1997)

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

What is machine learning?

Definition by Tom Mitchell (1997)

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Face Detection

- E : images (with bounding boxes) around faces

What is machine learning?

Definition by Tom Mitchell (1997)

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Face Detection

- E : images (with bounding boxes) around faces
- T : given an image without boxes, put boxes around faces

What is machine learning?

Definition by Tom Mitchell (1997)

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Face Detection

- E : images (with bounding boxes) around faces
- T : given an image without boxes, put boxes around faces
- P : number of faces correctly identified

What is machine learning?

Learning = Representation + Evaluation + Optimisation

- **Representation:** Hypothesis space of the learner
 - Choose the set of classifiers that it can possibly learn
- **Evaluation:** Objective or scoring function
 - Distinguish good classifiers from bad ones
- **Optimisation:** Search method for the highest-scoring classifier
 - Key to the efficiency of the learner
 - Unlike in most optimisation problems, we do not have access to the function we want to optimise!
 - Use training error as a surrogate for test error :(
 - Objective function is only a proxy for the true goal \Rightarrow no need to fully optimise it, local optimum may be OK

An early (first?) example of automatic classification

Ronald Fisher: Iris Flowers (1936)

- Three types: setosa, versicolour, virginica
- Data: sepal width, sepal length, petal width, petal length



setosa

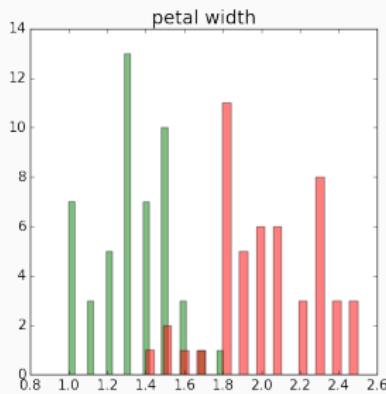
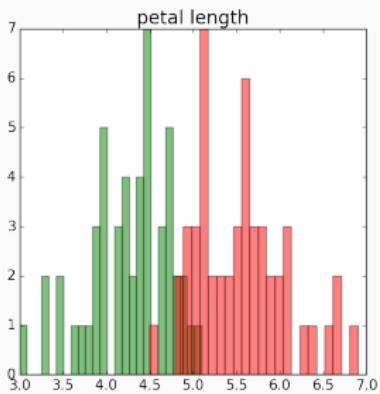
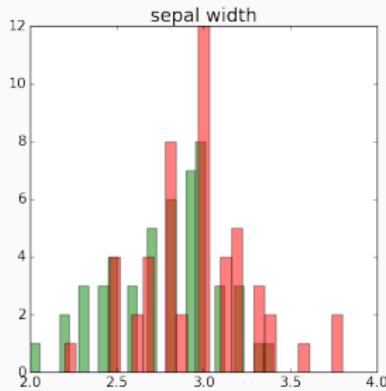
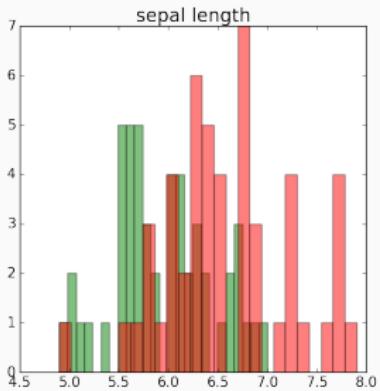


versicolour

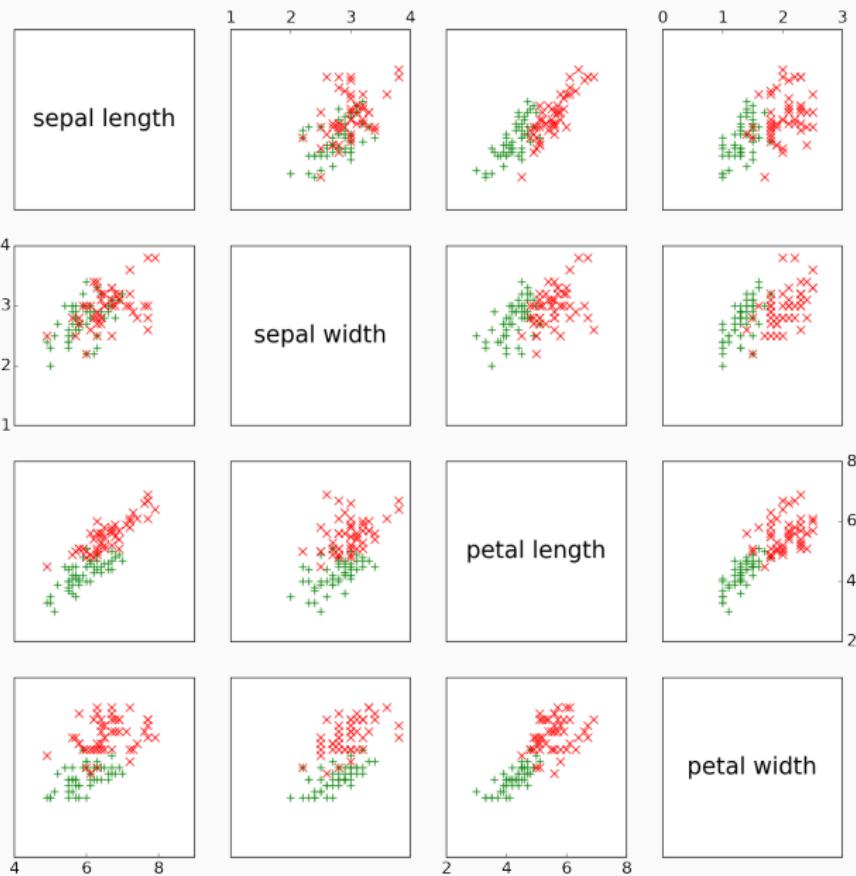


virginica

Histogram Plots for Different Measurements for Versicolour & Virginica



Scatter Plots of Pairwise Measurements for Versicolour & Virginica



An early (first?) example of automatic classification

Ronald Fisher: Iris Flowers (1936)

- Three types: setosa, versicolour, virginica
- Data: sepal width, sepal length, petal width, petal length
- Method: Find linear combinations of features that maximally differentiates the classes



setosa



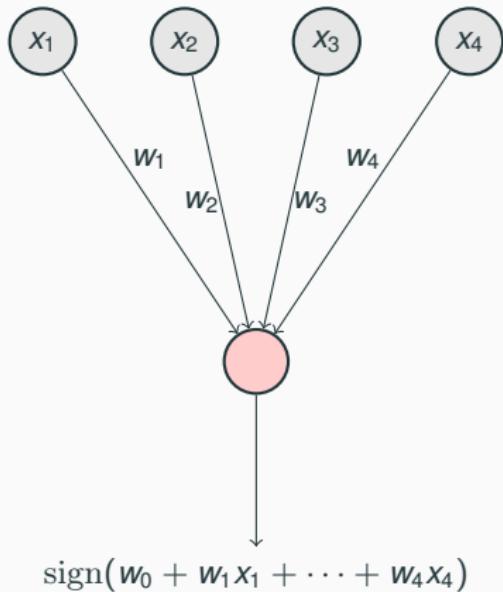
versicolour



virginica

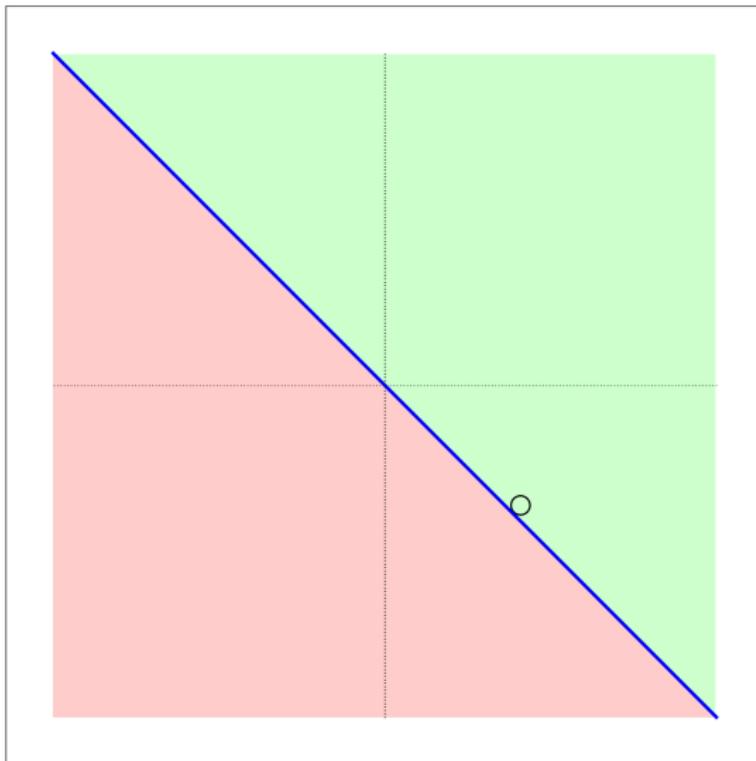
Frank Rosenblatt and the Perceptron

- Perceptron (1957) - inspired by neurons
- Simple learning algorithm: Adjust the weights if incorrect prediction on new input
- Built using specialised hardware

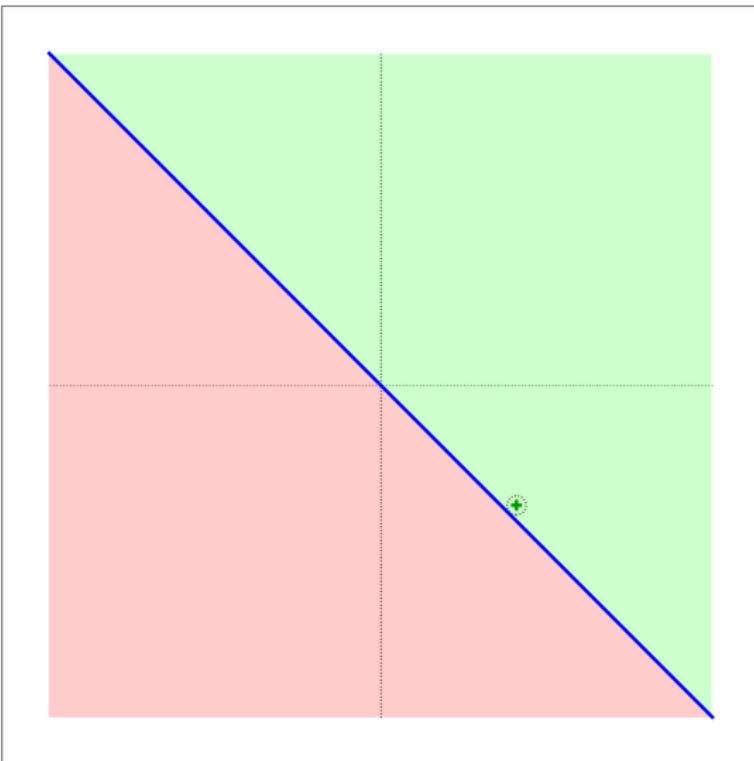


$$= \text{sign}(\mathbf{w} \cdot \mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{w} \cdot \mathbf{x} \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

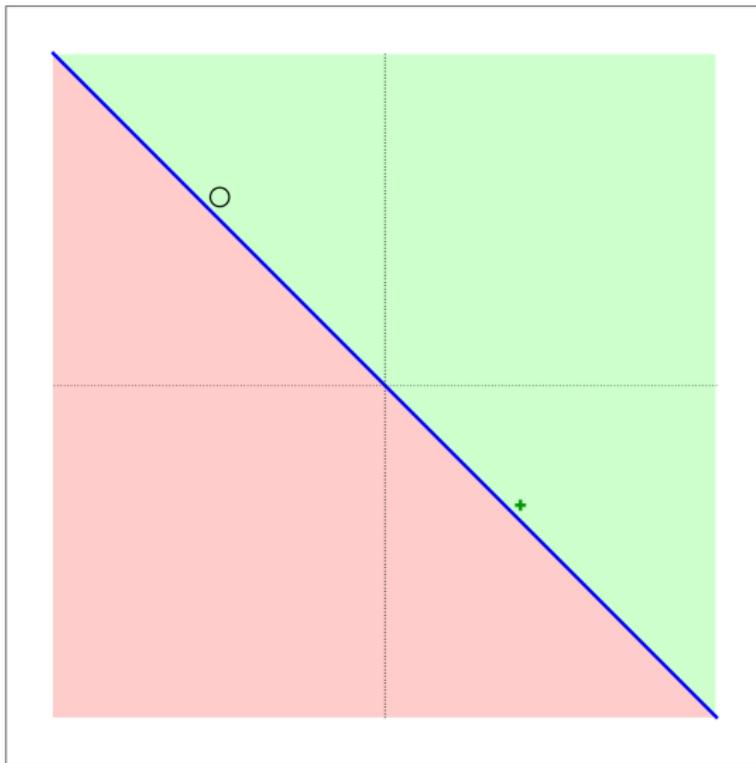
Perceptron Training Algorithm



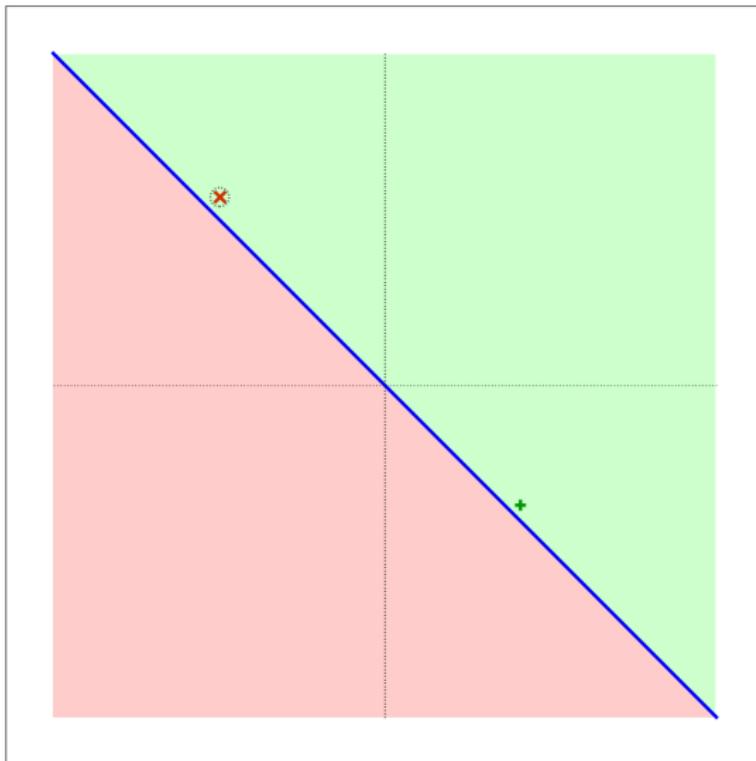
Perceptron Training Algorithm



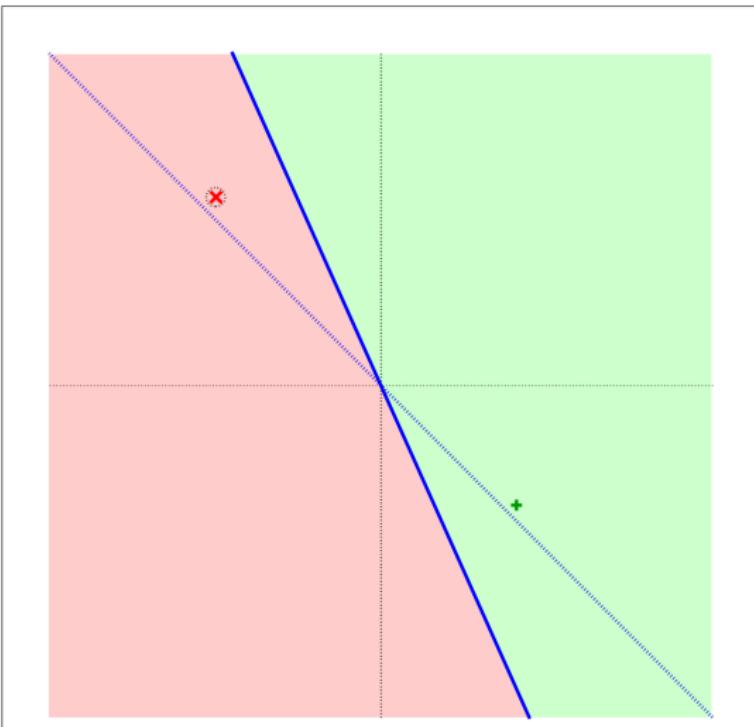
Perceptron Training Algorithm



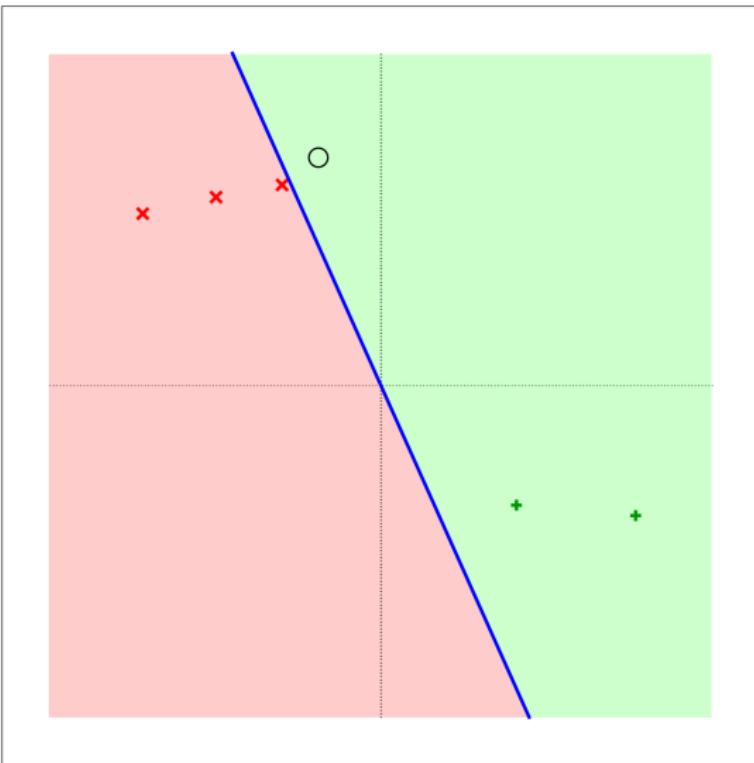
Perceptron Training Algorithm



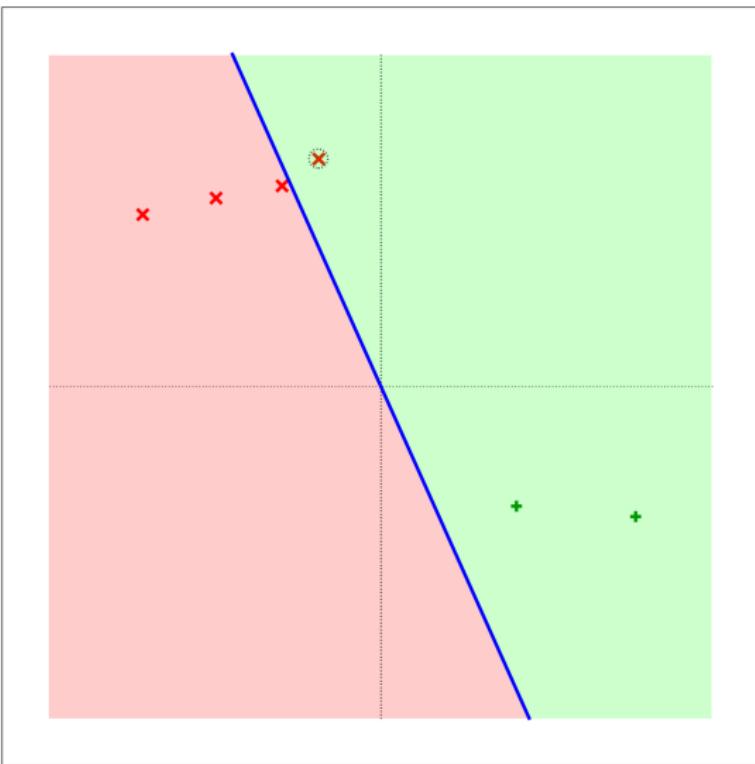
Perceptron Training Algorithm



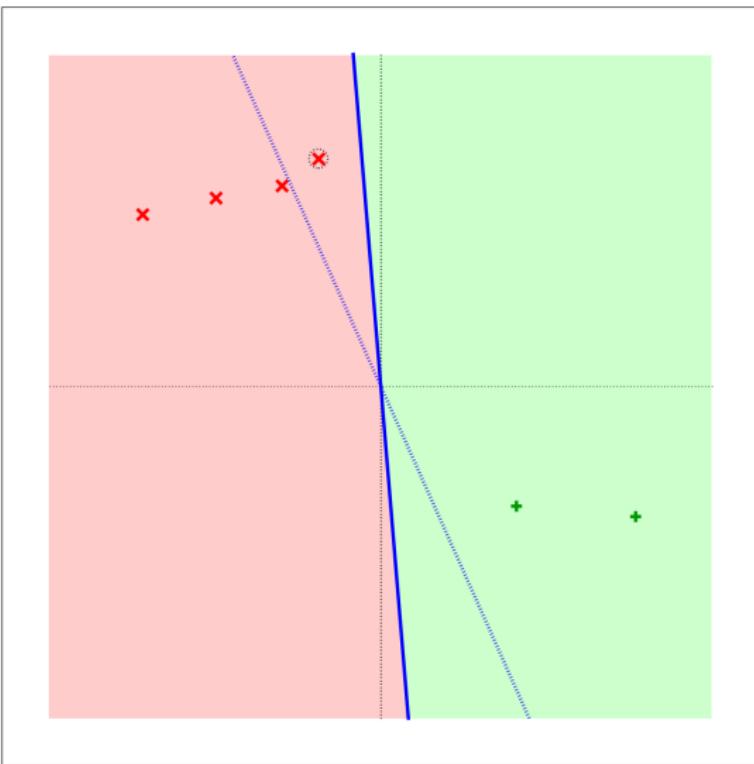
Perceptron Training Algorithm



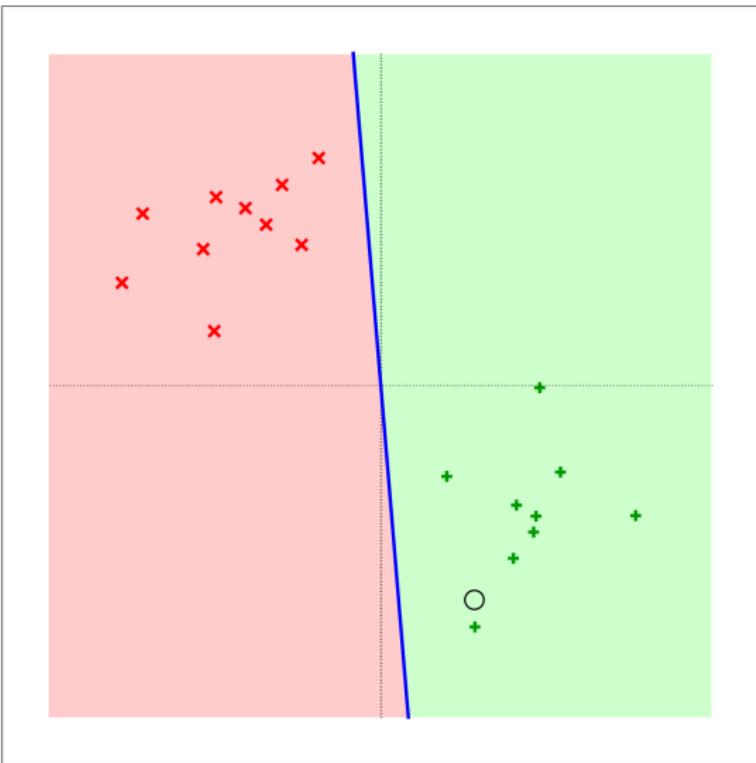
Perceptron Training Algorithm



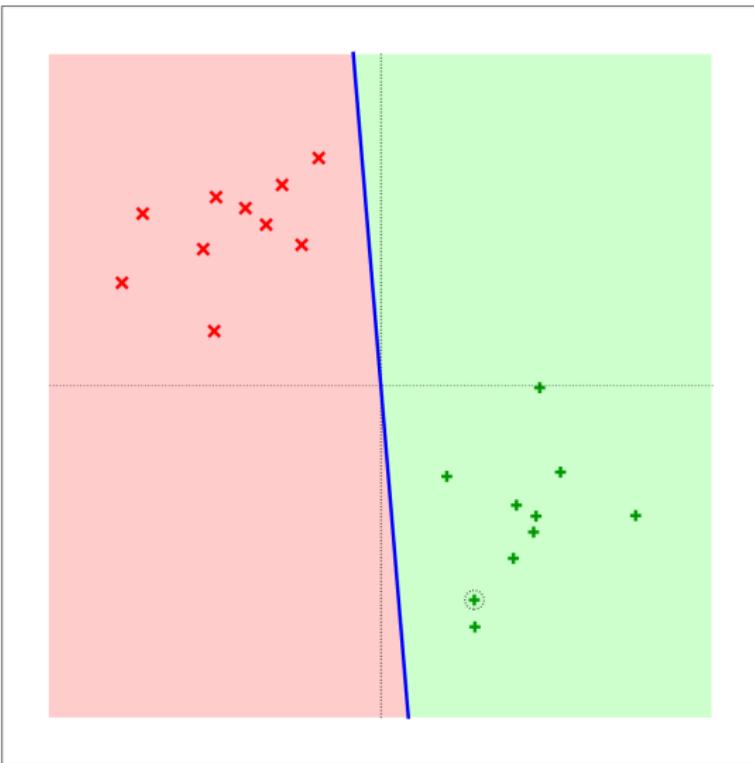
Perceptron Training Algorithm



Perceptron Training Algorithm



Perceptron Training Algorithm



The Perceptron Algorithm

```
Initialisation  $w = 0$                                 -- all params set to 0

Repeat Until Convergence:

    For  $t = 1 \dots n$                                 -- go over all examples

        1.  $y' = \text{sign}(\mathbf{x}_t \cdot w)$             -- compute the prediction

        2. If  $y' \neq y_t$  then  $w := w + y_t \mathbf{x}_t$       -- update the params
            else leave  $w$  unchanged
```

The Perceptron Algorithm

```
Initialisation  $w = 0$                                 -- all params set to 0

Repeat Until Convergence:

    For  $t = 1 \dots n$                                 -- go over all examples

        1.  $y' = \text{sign}(\mathbf{x}_t \cdot w)$             -- compute the prediction

        2. If  $y' \neq y_t$  then  $w := w + y_t \mathbf{x}_t$       -- update the params
            else leave  $w$  unchanged
```

Convergence: w remains unchanged for an entire pass over the training set.
Then, all training examples are classified correctly.

The Perceptron Convergence Theorem

Assume:

- $\exists \mathbf{w}', \gamma > 0$ such that $\forall t = 1 \dots n: y_t(\mathbf{x}_t \cdot \mathbf{w}) \geq \gamma$
- $\forall t = 1 \dots n: \|\mathbf{x}_t\| \leq R$ -- $\|\cdot\|$ is the Euclidean norm

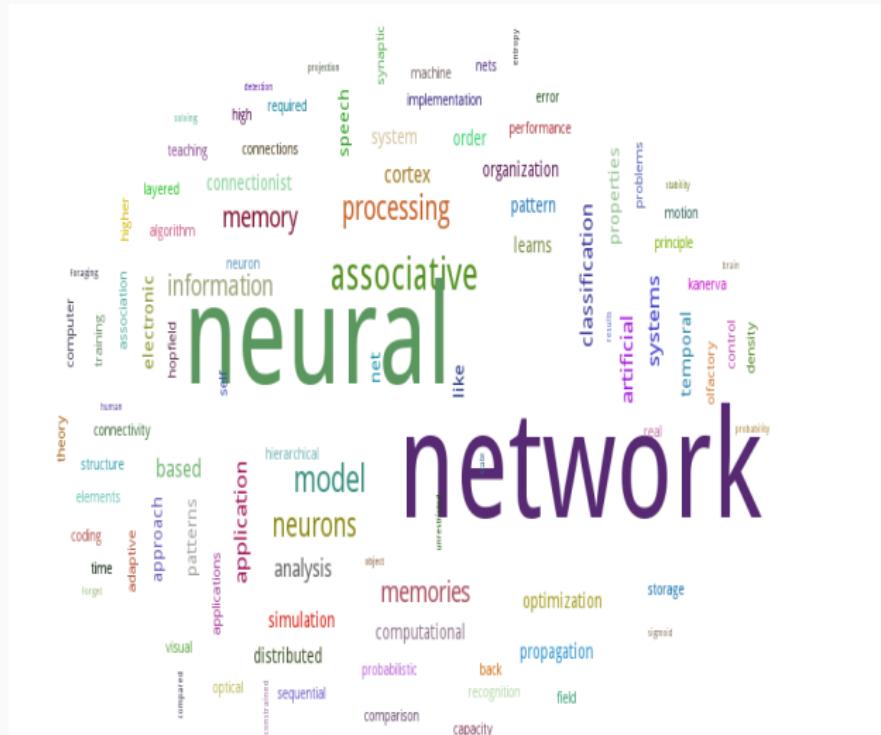
Then the perceptron algorithm makes at most $\frac{R^2}{\gamma^2}$ updates.

Simple proof by Michael Collins available online.

Machine Learning Models and Methods

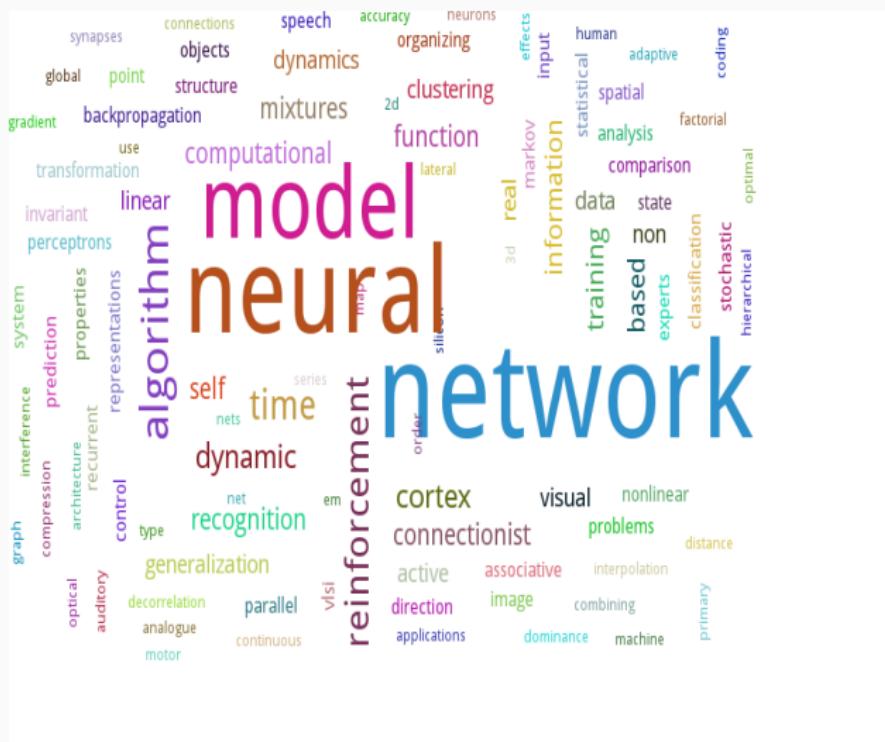
- k-Nearest Neighbours
- Linear Regression
- Logistic Regression
- Ridge Regression
- Hidden Markov Models
- Mixtures of Gaussian
- Principle Component Analysis
- Independent Component Analysis
 - Kernel Methods
 - Decision Trees
 - Boosting and Bagging
 - Belief Propagation
 - Variational Inference
 - EM Algorithm
 - Monte Carlo Methods
 - Spectral Clustering
 - Hierarchical Clustering
- Recurrent Neural Networks
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Perceptron Algorithm
- Naïve Bayes Classifier
- Hierarchical Bayes
- k -means Clustering
- Support Vector Machines
- Gaussian Processes
- Deep Neural Networks
- Convolutional Neural Networks
- Markov Random Fields
- Structural SVMs
- Conditional Random Fields
- Structure Learning
- Restricted Boltzmann Machines
- Multi-dimensional Scaling
- Reinforcement Learning
- ...

NIPS Papers!

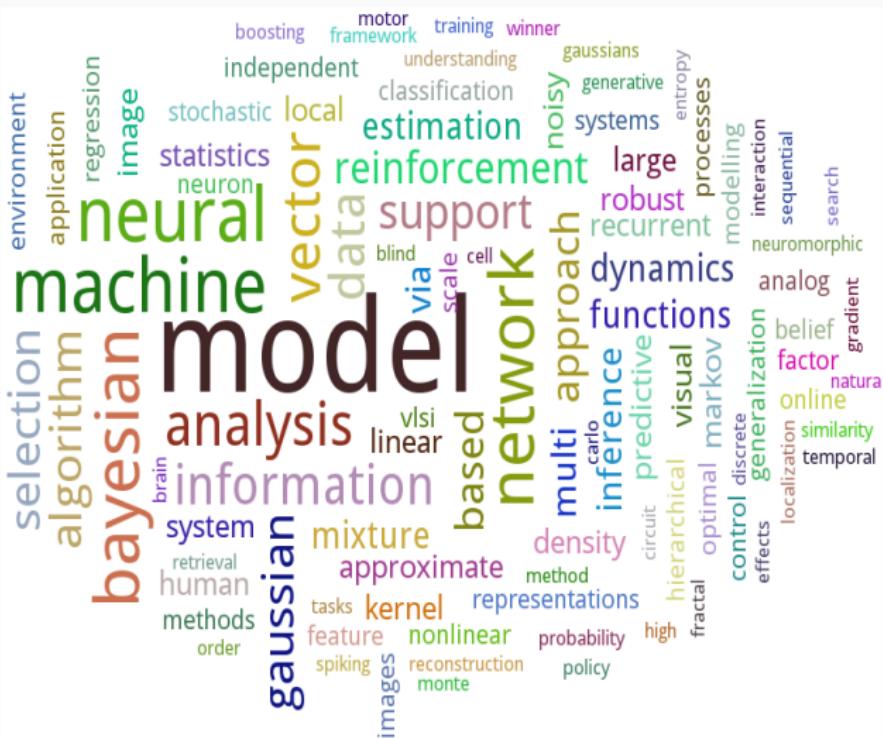


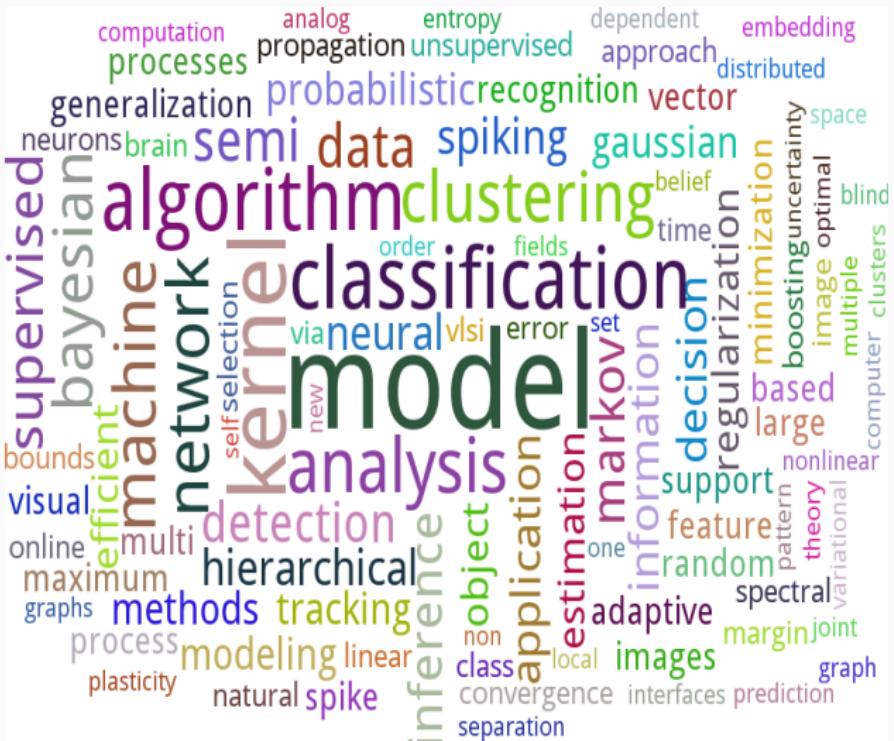
Advances in Neural Information Processing Systems 1988

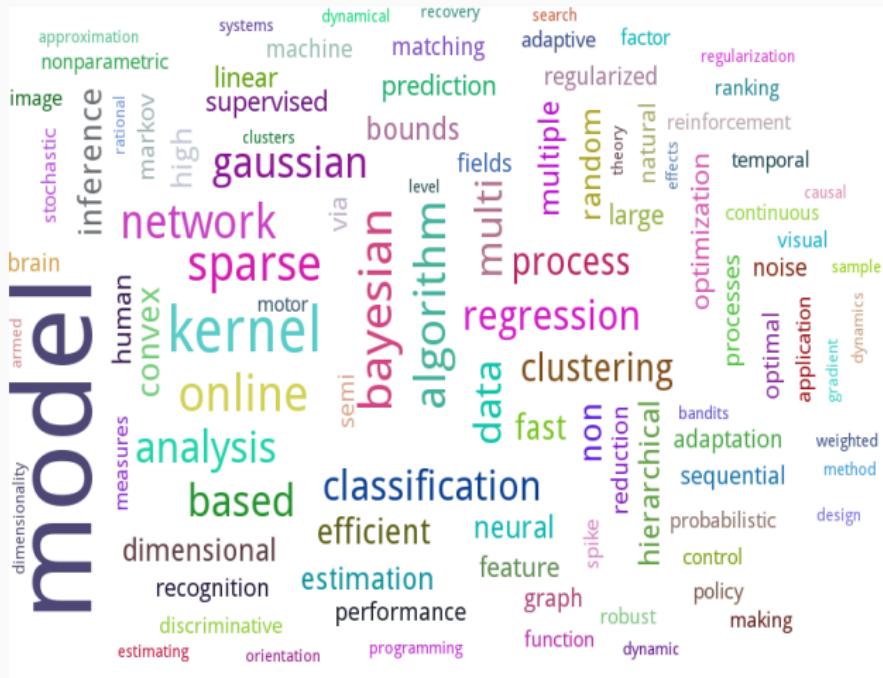
NIPS Papers!

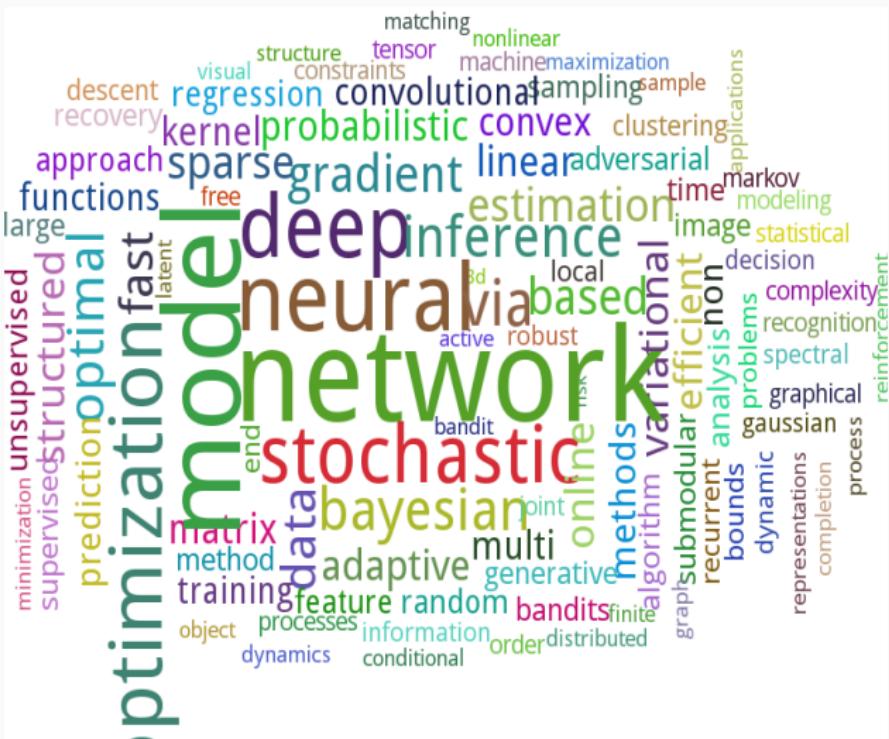


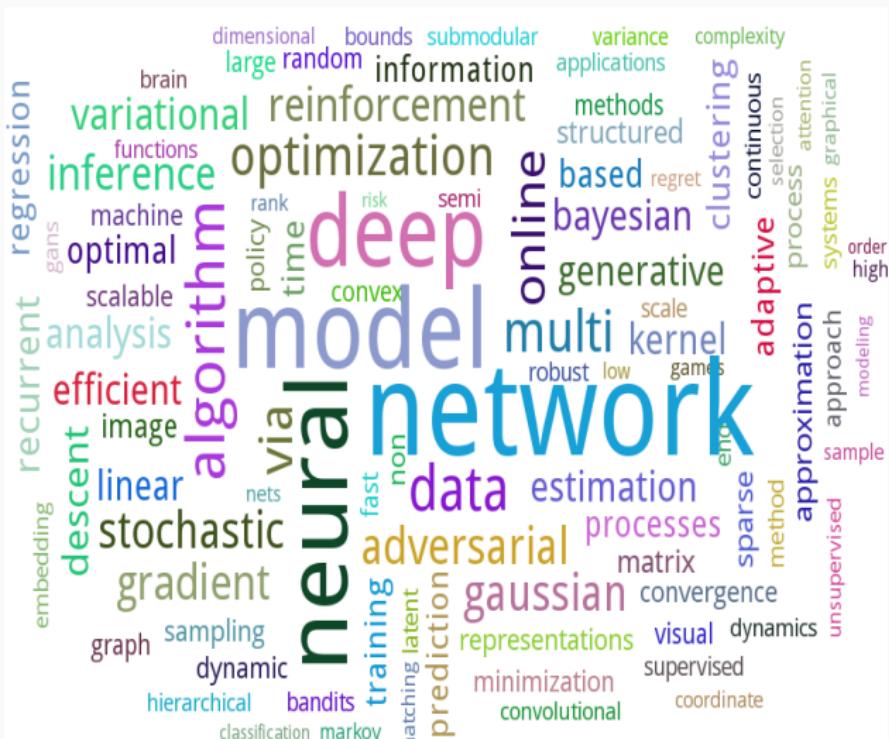
Advances in Neural Information Processing Systems 1995











Advances in Neural Information Processing Systems 2017

<https://www.youtube.com/watch?v=mlXzufEk-2E>

Application: Boston Housing Dataset

Numerical attributes

- Crime rate per capita
- Non-retail business fraction
- Nitric Oxide concentration
- Age of house
- Floor area
- Distance to city centre
- Number of rooms

Predict house cost

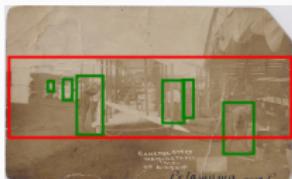


Categorical attributes

- On the Charles river?
- Index of highway access (1-5)

Source: [UCI repository](#)

Application: Object Detection and Localisation



- 200-basic level categories
- Here: Six pictures containing airplanes and people
- Dataset contains over 400,000 images
- Imagenet competition (2010-)
- All recent successes through very deep neural networks!

Supervised Learning

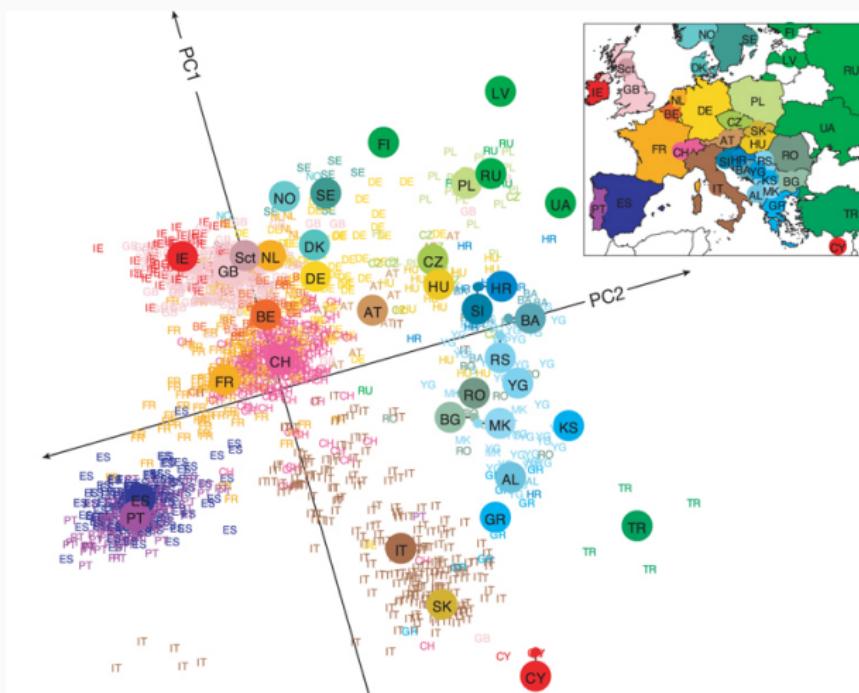
Training data has inputs \mathbf{x} (numerical, categorical) as well as outputs y (target)

Regression: When the output is real-valued, e.g., housing price

Classification: Output is a category

- Binary classification: only two classes e.g., spam
- Multi-class classification: several classes e.g., object detection

Unsupervised Learning : Genetic Data of European Populations



Source: Novembre *et al.*, Nature (2008)

Dimensionality reduction - Map high-dimensional data to low dimensions

Clustering - group together individuals with similar genomes

Unsupervised Learning : Group Similar News Articles

The screenshot shows the Google News homepage in Mozilla Firefox. The left sidebar lists categories: Top Stories, Donald Trump, Google, Florida, Nobel Prize, Brexit, Formula One, Samsung Electronics Limited, Wayne Rooney, Oculus Rift, PlayStation VR, Oxford, England, World, U.K., Business, Technology, Entertainment, Sports, Science, Health, and Spotlight. The main content area displays news articles under these categories:

- Top Stories:** US election: Donald Trump says he will not quit over video (Daily Mail), German city on lock down as police investigate bomb plot threat (Daily Mail), Trump vows to stay in race after calls for him to quit over lewd remarks (Daily Mail), A weakening Matthew rakes Atlantic coast; US death toll at 4 (Daily Mail), Derby County part company with Nigel Pearson by mutual agreement (SkySports).
- Donald Trump:** US presidential candidate Donald Trump has said he will not withdraw from the race in phone interviews with US media. Mr Trump has been under pressure after a tape of him making lewd sexual comments and bragging about groping and kissing women ...
- Google:** Viewers' GUIDE: Time for candidates to keep calm, debate on
- Florida:** US presidential candidate Donald Trump has said he will not withdraw from the race in phone interviews with US media. Mr Trump has been under pressure after a tape of him making lewd sexual comments and bragging about groping and kissing women ...
- Nobel Prize:** Related: Donald Trump + Hillary Clinton +
- Brexit:** Live Updating: Second presidential debate 2016: What time, how to watch and live stream
- Formula One:** Live stream coverage
- Samsung Electronics Limited:** Samsung Electronics
- Wayne Rooney:** Wayne Rooney
- Oculus Rift:** Oculus Rift
- PlayStation VR:** PlayStation VR
- Oxford, England:** Oxford, England
- World:** Oxford, England
- U.K.:** Oxford, England
- Business:** Oxford, England
- Technology:** Oxford, England
- Entertainment:** Oxford, England
- Sports:** Oxford, England
- Science:** Oxford, England
- Health:** Oxford, England
- Spotlight:** Oxford, England

On the right side, there is a weather forecast for Oxford, England, showing sun icons for the next five days and temperatures ranging from 14° to 19°. Below the weather is a link to "The Weather Channel - Weather Underground - AccuWeather". There is also a section titled "Editors' Picks" featuring a "Mirror" logo.

Group similar articles into categories such as politics, music, sport, etc.

In the dataset, there are no labels for the articles

Active and Semi-Supervised Learning

Active Learning

- Initially all data is unlabelled
- Learning algorithm can ask a human to label some data



Semi-supervised Learning

- Limited labelled data, lots of unlabelled data
- How to use the two together to improve learning?

000000000000000000
111111111111111111
222222222222222222
333333333333333333
444444444444444444
555555555555555555
666666666666666666
777777777777777777
888888888888888888
999999999999999999

7210414959
0690159784
9665407401
3134727121
1742351244

Collaborative Filtering : Recommender Systems

Movie / User	Alice	Bob	Charlie	Dean	Eve
The Shawshank Redemption	7	9	9	5	2
The Godfather	3	?	10	4	3
The Dark Knight	5	9	?	6	?
Pulp Fiction	?	5	?	?	10
Schindler's List	?	6	?	9	?

Netflix competition to predict user-ratings (2008-09)

Any individual user will not have used most products

Most products will have been used by some individual



Reinforcement Learning

- Automatic flying helicopter; self-driving cars
- Cannot conceivably program by hand
- Uncertain (stochastic) environment
- Must take **sequential decisions**
- Can define **reward functions**
- Fun: Playing Atari breakout!

<https://www.youtube.com/watch?v=V1eYniJ0Rnk>



Cleaning up data

Spam Classification

- Look for words such as Nigeria, millions, Viagra, etc.
- Features such as the IP, other metadata
- If email addressed by to user personally

Getting Features

- Often hand-crafted features by domain experts
- In this course, we mainly assume that we already have features
- Feature learning using deep networks

Some pitfalls

Sample Email

"To build a spam classifier, we check if at least two words such as Nigeria, millions, etc. appear in the message. If that is the case, we mark the email as spam."

Training vs Test Data

- Future data should look like past data
- Not true for spam classification. Spammers will try adversarially to break the learning algorithm.

Cats vs Dogs

