Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

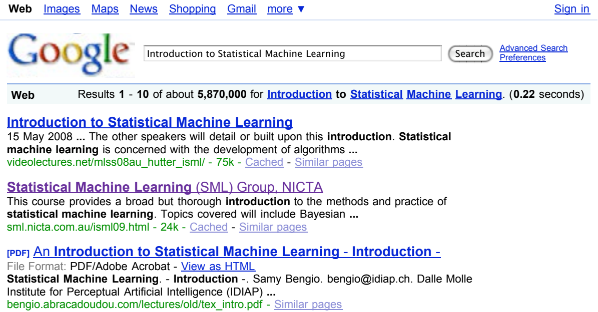
Outlines

Part I

Overview

Search Machine Ranking – 2009

- Given a number of web sites which match some search phrase:
  Learn which pages most of the users are looking for.


Web Images Maps News Shopping Gmail more ▼ Sign in

Introduction to Statistical Machine Learning | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **5,870,000** for **Introduction to Statistical Machine Learning**. (**0.22** seconds)

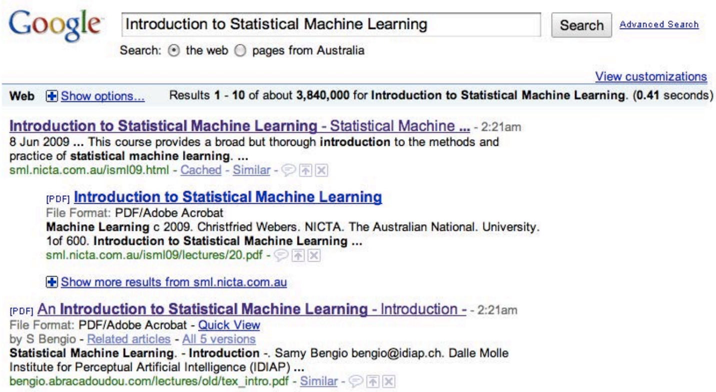
Introduction to Statistical Machine Learning
15 May 2008 **...** The other speakers will detail or built upon this **introduction**. **Statistical machine learning** is concerned with the development of algorithms **...**
videolectures.net/mlss08au_hutter_isml/ - 75k - Cached - Similar pages

Statistical Machine Learning (SML) Group, NICTA
This course provides a broad but thorough **introduction** to the methods and practice of **statistical machine learning**. Topics covered will include Bayesian **...**
sml.nicta.com.au/isml09.html - 24k - Cached - Similar pages

[PDF] An **Introduction to Statistical Machine Learning - Introduction -**
File Format: PDF/Adobe Acrobat - View as HTML
Statistical Machine Learning. - **Introduction** -. Samy Bengio. bengio@idiap.ch. Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) **...**
bengio.abracadoudou.com/lectures/old/tex_intro.pdf - Similar pages

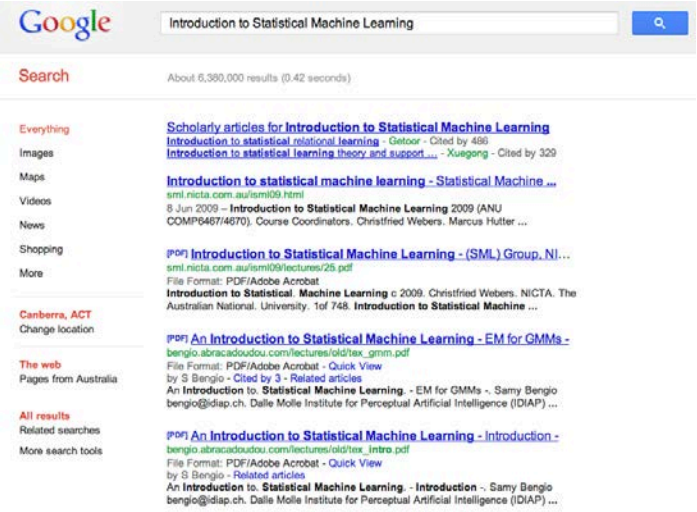
Search Machine Ranking – 2010

Google has learned



Search Machine Ranking – 2012

Has Google learned more?

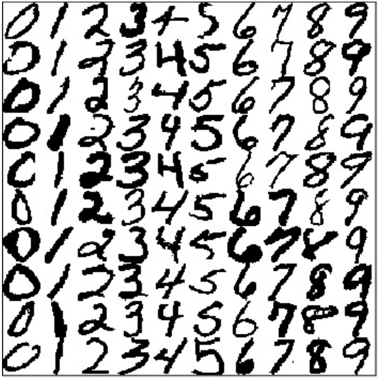


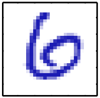
Junk Mail Filtering

	From / To	Date Sent	Thread
2949	Support	10/02/09 7:45 +0…	Message from eBay com au
2950	Ken.Johnston	10/02/09 14:12 +…	Fool them once, fool them twice, fool…
2951	christfried.web…	10/02/09 3:14 -0…	Assistance, Petersen
2952	Air Sep	9/02/09 4:53 -0800	Un negocio de por vida 1000% Renta…
2953	Osita John	10/02/09 17:33 +…	Now Contact my secretary ask him fo…
2954	Air Sep	9/02/09 0:38 -0800	Un negocio de por vida 1000% Renta…
2955	Air Sep	9/02/09 10:12 -0…	Un negocio de por vida 1000% Renta…
2956	MISS MERCY…	29/01/09 23:13 -…	Urgent Attention(YOUR FILE HAVE…
2957	PEPSI BOTTL…	25/07/08 11:23 -…	OEP00934/UK
2958	JOSEPH POON	11/02/09 12:04 +…	MY PROPOSAL!!!
2959	MADAM ERL…	11/02/09 13:41 +…	LOOKING FOR A TRUSTWORTHY…
2960	REBECA RO…	11/02/09 18:48 +…	Dear sir/madam:
2961	REBECA RO…	11/02/09 18:48 +…	Dear sir/madam:
2962	Elinor Shannon	11/02/09 22:37 +…	I shall look forward to hearing from you
2963	Air Sep	10/02/09 14:37 -…	Un negocio de por vida 1000% Renta…
2964	Foreign Payme…	1/02/09 16:13 +0…	Goodday,
2965	JANET KUEN	12/02/09 16:11 +…	Dear sir/madam:
2966	Abubakar Mar…	10/02/09 19:04 +…	OUR DEAR FRIEND
2967	JAMES ROBE…	12/02/09 23:12 -…	From James Roberts
2968	Bases de Email…	13/02/09 10:50 -…	Nuevas Bases de Datos de Mexico
2969	Barrister Willi…	15/02/09 1:23 +0…	WILL AND TESTAMENT
2970	Isolde	15/02/09 9:45 -0…	A Valentine's Day Ecard Special Deli…
2971	NTI eNews	15/02/09 12:25 -…	Super Sweet Deals From NTIus.com

- Given some examples what the user defined as junk mail.
- From these examples, learn to identify new incoming junk mail.

Handwritten Digit Recognition





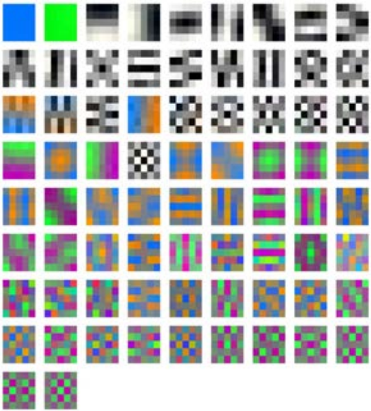
- Given handwritten ZIP codes on letters, money amounts on cheques etc.
- Learn to recognise the correct digit written by hand.

Backgammon

- World best computer program TD-GAMMON (Tesauro 1992, 1995) played over a million games against itself.
- Plays now on the level of human world champion.

Image Denoising

1 Learn a statistics over patches from many images of natural scenes



50 000 patches ($5 \times 5 \times 3$) from the Berkeley Segmentation Database.
McAuley et. al., "Learning High-Order MRF Priors of Color Images", ICML2006

Image Denoising

2 Use this knowledge to a denoise a yet unseen image

Original image

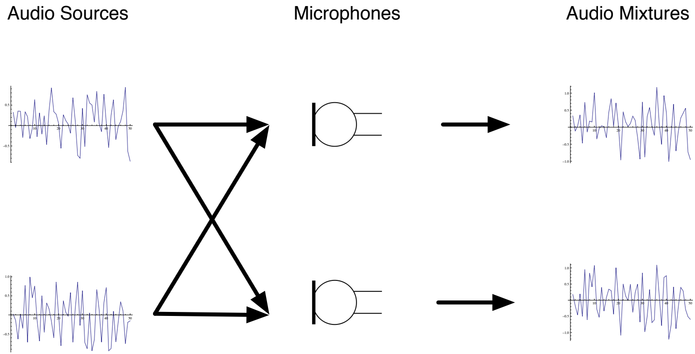
Noise added

Denoised

McAuley et. al., "Learning High-Order MRF Priors of Color Images", ICML2006

Separating Audio Sources

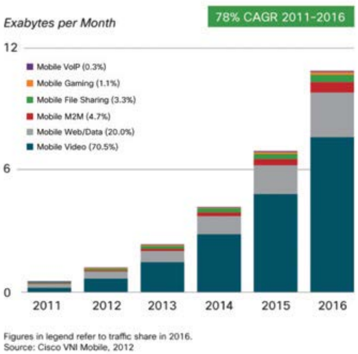
Cocktail Party Problem (human brains may do it differently ;–)



(J. Steinbauer et. al., `http://cnx.org/content/m15712/latest/`)

Smart Mobile Content Delivery

- 70% of all traffic on mobile networks is video
- Streaming of video is awful because of congestions
- Preload content with spare network capacity
- Play from local storage
- Learn the user behaviour on the device (privacy!)



Smart Mobile Content Delivery

- Trial on Android OS
- `http://watch.incoming.tv`

Predicting Solar Panel Output

- Photovoltaics now very close to grid electricity in price
- Distributed system of generators
- Energy market
- Great Machine Learning Problem: Predict the solar energy output (variability primarily due to clouds) for Australia
- Pilot project in Canberra : Use cheap cameras to take 360° sky photos in several location.
- Learn to predict 3-D model of cloud movement.
- Learn orientation and efficiency of solar panels for each house from time series of energy output.
- Predict output of each solar panel for 15 min to 1 hour from current snapshots.

Other applications of Machine Learning

- autonomous robotics,
- detecting credit card fraud,
- detecting network intrusion,
- bioinformatics,
- neuroscience,
- medical diagnosis,
- stock market analysis,
- playing games by self-play: Checker and Backgammon.

What is common to this examples?

1. Given some data (e.g. hand written digits).
2. Possibly some extra information (e.g. which digit does this number represent)
3. Goal: Built a machine which can learn from the given data utilising the extra information (if available).

NICTA
The Australian National University

ISML 2013

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

- machines? computers? HAL?
- to learn?

NICTA
The Australian National University

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

- machines? computers? HAL?
- to learn?
- need to quantify "learning"
- to improve their performance over time

Definition (Second Try)

Machine learning is concerned with the design and development of algorithms that allow computers (machines) to improve their performance over time.

What is Machine Learning?

- What is the source of the improved performance?
- New insights by the algorithm designer?

ISML 2013

What is Machine Learning?

- What is the source of the improved performance?
- New insights by the algorithm designer?

Definition (Final Version)

Machine learning is concerned with the design and development of algorithms that allow computers (machines) to improve their performance over time based on data.

Christfried Webers
NICTA
The Australian National University

What is Machine Learning?

NICTA
The Australian National University

Definition (Mitchell, 1998)

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 the challenge?

- Given only some examples.
- Need to derive a relation for many more (possibly infinite) unseen examples.
- Occam's Razor (William Ockham, circa 1285 – 1349): "Plurality must never be posited without necessity"
- "The simplest explanation or strategy tends to be the best one."
- By the way: Often cited "Entities should not be multiplied unnecessarily." can not be found literally in Ockham's works. Multiple versions ;–)

Christfried Webers
NICTA
The Australian National University

Statistical Machine Learning - Some History

- 1960's : symbolic AI; computers learn rules from data; analysis of the underlying statistics is seldom done.
- Perceptron (Rosenblatt, 1957), "Perceptrons" (Minsky and Papert, 1969)
-
- 1980's : artificial neural networks
- 1990's - 2000's : statistical machine learning (kernel methods, decision trees, graphical models)
- Why Statistical Machine Learning not earlier?
  - faster computers with larger memory to represent statistical models have become available
  - numerical methods on the desktop computer (BLAS, LAPACK, Optimisation)
  - found new interesting classes of algorithms (e.g. on graphs)
  - large amounts of data available which can be tapped into (flickr, social networks)
  - many data sets with partial/incomplete data (e.g. netflix)

Why Machine Learning?

Machine Learning is essential when

- humans are unable to explain their expertise (speech recognition).
- humans are not around for help (navigation on Mars, underwater robotics).
- large amount of data with possible hidden relationships and correlations
- environment changes (fast) in time (mobile phone network).
- solutions need to be adapted to many particular cases (junk mail).

Example: It is easier to write a program that learns to play checkers or backgammon well by self-play rather than converting the expertise of a master player to a program.

Christfried Webers
NICTA
The Australian National University

Related Fields

- Artificial Intelligence - AI
- Statistics
- Game Theory
- Neuroscience, Psychology
- Data Mining
- Computer Science
- Adaptive Control Theory
- ...

Artificial Intelligence - AI

- Artificial intelligence is the intelligence of machines and the branch of computer science which aims to create it.
- The field was founded on the claim that human intelligence can be so precisely described that it can be simulated by a machine.
- Central areas: reasoning, knowledge, planning, learning, communication, perception and the ability to move and manipulate objects (autonomous robotics).
- Philosophical questions: Can a machine have a mind and consciousness? Are there limits to how intelligent machines can be?

Christfried Webers
NICTA
The Australian National University

Statistics

- Descriptive Statistics: Summarize or describe a collection of data
- Inferential Statistics: Draw inferences about a process taking randomness and uncertainty into account
- What can be inferred from data and some modelling assumptions?
- How reliable are the results?

Neuroscience, Psychology

- How do animals learn?
- Modelling of human learning (e.g. Bayesian models of human inductive learning)
- increasing interaction between Statistical Machine Learning and Neuroscience, Psychology
  e.g. NIPS - Neural Information Processing Systems Conference 2009:
  - “Discriminative Network Models of Schizophrenia”,
  - “Functional network reorganization in motor cortex can be explained by reward-modulated Hebbian learning”,
  - “Canonical Time Warping for Alignment of Human Behavior”
- new technologies (e.g. functional magnetic imaging resonance [fMRI])

Data Mining

- searching through large data sets (databases)
- goal: extracting hidden patterns from data
- examples: bioinformatics, genetics, medicine
- genetics: how do differences in the DNA between humans relate to different risks of getting diseases such as cancer
- no magic: can not uncover patterns which are not already present in the data

ISML 2013

Computer Science

- "What can be (efficiently) automated?"
- Algorithms
- Data Structures
- Computational complexity theory

Adaptive Control Theory

- consider systems with parameters changing (slowly) over time or being uncertain
- Example: aircraft which changes its weight over time depending on fuel consumption (which in turn depends on the wind)
- how to control such a system?
- how to estimate the parameters?

Some Basic Notation - Data

- The set of all input data is denoted as $\mathcal{X}$. For instance, $\mathcal{X} = \{\, x \mid x \text{ is an image containing a handwritten digit} \,\}$.
- One data point with D elements :

$$\mathrm{x} = \begin{bmatrix} x_1 \\ \dots \\ x_D \end{bmatrix} = (x_1, \dots, x_D)^T.$$

- Data matrix : A set of N data points $\mathrm{x_i}$, where $i = 1 \dots N$,

$$\mathrm{X} = \begin{bmatrix} x_{1,1} \dots x_{1,D} \\ x_{2,1} \dots x_{2,D} \\ \dots \\ x_{N,1} \dots x_{N,D} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \dots \\ \mathrm{x}_{\mathrm{N}}^{\mathrm{T}} \end{bmatrix}.$$

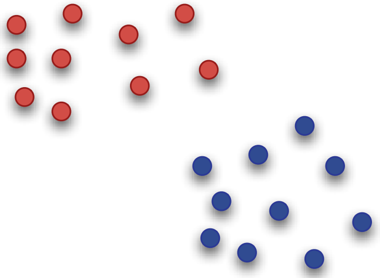
  (Note : Each data point $\mathrm{x_i}$ is a column vector, but appears as a row vector in X.)
- If $D = 1$, X is a vector of N scalar data points. We write

$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \dots \\ \mathrm{x}_{\mathrm{N}} \end{bmatrix}.$$

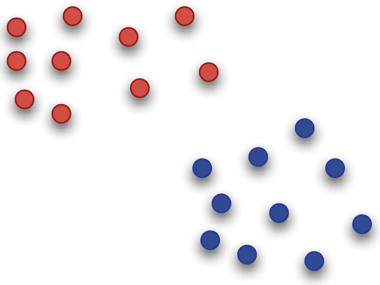
Some Basic Notation - Targets

- A target can be from a finite discrete set ('labels') or from $\mathbb{R}$.
  (Note: Can extend this idea to m-dimensional labels and $\mathbb{R}^m$.)
- Set of Targets $\mathcal{T}$, e.g.
  $\mathcal{T} = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}, \text{seven}, \text{eight}, \text{nine}, \text{zero}\}$.
- An ordered set of N scalar labels $\mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_N \end{bmatrix} = (t_1, \dots, t_N)^T$.

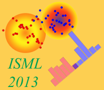
Supervised Learning



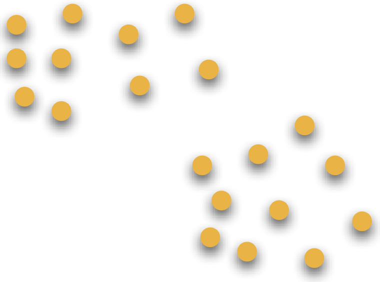
Supervised Learning



- Given are pairs of data $\mathrm{x_i} \in \mathcal{X}$ and targets $t_i \in \mathcal{T}$ in the form $(\mathrm{x_i}, t_i)$, where $i = 1..N$.
- Learn a mapping between the data X and the target $\mathbf{t}$ which generalises well to new data.

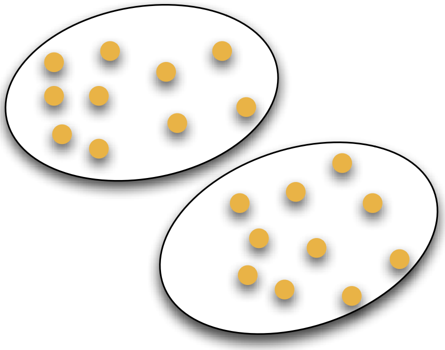


Unsupervised Learning



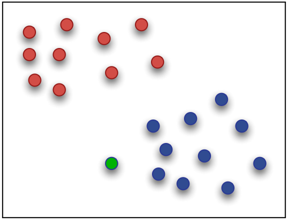
- Given only the data $\mathrm{x}_{\mathrm{i}} \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data X.

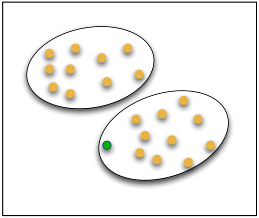
Unsupervised Learning



- Given only the data $x_i \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data.

Testing - Supervised versus Unsupervised Learning

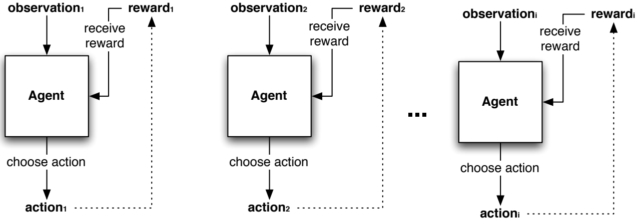




Reinforcement Learning

- Example: Game playing. There is one reward at the end of the game (negative or positive).
- Find suitable actions in a given environment with the goal of maximising some reward.
- correct input/output pairs never presented
- Reward might only come after many actions.
- Current action may not only influence the current reward, but future rewards too.

Reinforcement Learning



- Exploration versus Exploitation.
- Well suited for problems with a long-term versus short-term reward trade-off.
- Naturally focusing on online performance.

Other Machine Learning Types

- Active Learning
  - The algorithm may choose which data $\mathbf{x}_i \in \mathcal{X}$ to select next when building the model.
  - The order of the data is actively chosen by the algorithm at run-time.
- Transduction
  - The algorithms is allowed to use the test data (but of course not labels!) when building a model.
- Estimation with missing variables.
- Co-training with two different but related data sets.
- ... and others.

Training Regimes

- Batch Learning
  - All training data $\mathrm{X} = \{\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}\}$ and targets $\mathbf{t} = \{t_1, \dots, t_n\}$ are given.
  - Learn a mapping from x_i to t_i which can then be applied to yet unseen data $\mathrm{X}' = \{\mathrm{x}'_1, \dots, \mathrm{x}'_\mathrm{m}\}$ to find $\mathbf{t}' = \{t'_1, \dots, t'_m\}$.
- Online Processing
  - Pairs of $(\mathrm{x}_\mathrm{i}, t_i)$ become available one at a time.
  - At each step, learn and refine a mapping from x_i to t_i which can then be applied to yet unseen data x'_i.

Journals

- Journal of Machine Learning Research
- Machine Learning
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- Annals of Statistics
- Journal of the American Statistical Association
- SIAM Journal on Applied Mathematics (SIAP)

Conferences

- Algorithmic Learning Theory (ALT)
- Computational Learning Theory (COLT)
- Uncertainty in Artificial Intelligence (UAI)
- Neural Information Processing Systems (NIPS)
- European Conference on Machine Learning (ECML)
- International Conference on Machine Learning (ICML)
- International Joint Conference on Artificial Intelligence (IJCAI)
- International Conference on Artificial Neural Networks (ICANN)

Python

- dynamically typed programming language (no declarations for variables)
- supports object oriented, imperative, and functional programming style
- many built-in data types (str, tuple, list, set, dict, ...)
- packages for scientific programming (numpy, scipy)
- easily extensible to use code written in C and C++ (or FORTRAN for that matter)
- Python runs on Windows, Linux/Unix, Mac OS X, OS/2, Amiga, Palm Handhelds, and Nokia mobile phones
- OSI-approved Open Source License

Christfried Webers
NICTA
The Australian National University

Elefant

- Efficient Learning, Large-scale Inference, and Optimization Toolkit
- Mozilla Public License
- Two Layer
  - Functional Interface
  - Graphical User Interface

Christfried Webers
NICTA
The Australian National University

Elefant - System Diagram