

582631 – 4 credits

Introduction to Machine Learning

Lecturer: Jyrki Kivinen

Assistant: Yuan Zou

Department of Computer Science

University of Helsinki

based on material created by Patrik Hoyer and others

29 October–6 December 2013

Introduction

- ▶ What is machine learning? Motivation & examples
 - ▶ Definition
 - ▶ Relation to other fields
 - ▶ Examples
- ▶ Course outline and related courses
- ▶ Practical details of the course
 - ▶ Lectures
 - ▶ Exercises
 - ▶ Exam
 - ▶ Grading

What is machine learning?

- ▶ Definition:

machine = computer, computer program (in this course)

learning = improving performance on a given task, based on experience / examples

- ▶ In other words

- ▶ instead of the programmer writing explicit rules for how to solve a given problem, the programmer instructs the computer how to learn from examples
- ▶ in many cases the computer program can even become better at the task than the programmer is!

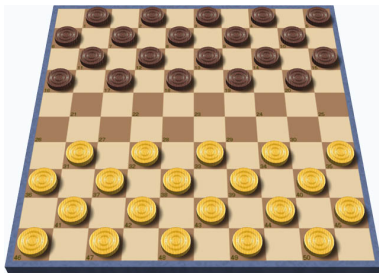
Example 1:

- ▶ How to program the computer to play tic-tac-toe?

		X		O		X		O		X		O		X		O		X		O		X		O		X
						X				X				X		X		X		X		X		X		X

- ▶ Option A: The programmer writes explicit rules, e.g. 'if the opponent has two in a row, and the third is free, stop it by placing your mark there', etc (lots of work, difficult, not at all scalable!)
- ▶ Option B: Go through the game tree, choose optimally (for non-trivial games, must be combined with some heuristics to restrict tree size)
- ▶ Option C: Let the computer try out various strategies by playing against itself and others, and noting which strategies lead to winning and which to losing (= 'machine learning')

- ▶ Arthur Samuel (50's and 60's):
 - ▶ Computer program that learns to play checkers
 - ▶ Program plays against itself thousands of times, learns which positions are good and which are bad (i.e. which lead to winning and which to losing)
 - ▶ The computer program eventually becomes much better than the programmer.



Example 2: spam filter

- ▶ Programmer writes rules: “If it contains ‘viagra’ then it is spam.” (difficult, not user-adaptive)
- ▶ The user marks which mails are spam, which are legit, and the computer learns itself what words are predictive

From: medshop@spam.com Subject: viagra cheap meds...	spam
From: my.professor@helsinki.fi Subject: important information here's how to ace the test...	non-spam
⋮	⋮
From: mike@example.org Subject: you need to see this how to win \$1,000,000...	?

Example 3: face recognition

- ▶ Face recognition is hot (facebook, apple; security; ...)
- ▶ Programmer writes rules: “If short dark hair, big nose, then it is Mikko” (impossible! how do we judge the size of the nose?!)
- ▶ The computer is shown many (image, name) example pairs, and the computer learns which features of the images are predictive (difficult, but not impossible)



patrik



antti



doris



patrik

...



?

...

Problem setup

- ▶ One definition of machine learning: A computer program improves its performance on a given task with experience (i.e. examples, data).
- ▶ So we need to separate
 - ▶ **Task:** What is the problem that the program is solving?
 - ▶ **Performance measure:** How is the performance of the program (when solving the given task) evaluated?
 - ▶ **Experience:** What is the data (examples) that the program is using to improve its performance?

Related scientific disciplines (1)

- ▶ Artificial Intelligence (AI)
 - ▶ Machine learning can be seen as 'one approach' towards implementing 'intelligent' machines (or at least machines that behave in a seemingly intelligent way).
- ▶ Artificial neural networks, computational neuroscience
 - ▶ Inspired by and trying to mimic the function of biological brains, in order to make computers that learn from experience. Modern machine learning really grew out of the neural networks boom in the 1980's and early 1990's.
- ▶ Pattern recognition
 - ▶ Recognizing objects and identifying people in controlled or uncontrolled settings, from images, audio, etc. Such tasks typically require machine learning techniques.

Availability of data

- ▶ These days it is very easy to
 - ▶ collect data (sensors are cheap, much information digital)
 - ▶ store data (hard drives are big and cheap)
 - ▶ transmit data (essentially free on the internet).
- ▶ The result? *Everybody* is collecting large quantities of data.
 - ▶ Businesses: shops (market-basket data), search engines (web pages and user queries), financial sector (stocks, bonds, currencies etc), manufacturing (sensors of all kinds), social networking sites (facebook, twitter), anybody with a web server (hits, user activity)
 - ▶ Science: genomes sequenced, gene expression data, experiments in high-energy physics, images of remote galaxies, global ecosystem monitoring data, drug research and development, public health data
- ▶ But how to benefit from it? Analysis is becoming key!

Related scientific disciplines (2)

- ▶ Data mining

- ▶ Trying to identify interesting and useful associations and patterns in huge datasets
- ▶ Focus on scalable algorithms

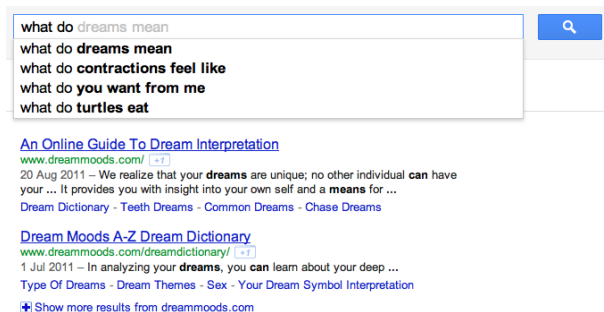
Example: On the order of 3 million people grocery shopping twice a week in just two main chains in Finland \Rightarrow each chain would collect hundreds of thousands of transaction receipts per day!

- ▶ Statistics

- ▶ Traditionally: focus on testing hypotheses based on theory
- ▶ Has contributed a lot to data mining and machine learning, and has also evolved by incorporating ideas derived from these fields

Example 4

- ▶ Prediction of search queries
 - ▶ ~~The programmer provides a standard dictionary~~ (words and expressions change!)
 - ▶ Previous search queries are used as examples!



Example 5

- ▶ Ranking search results:
 - ▶ Various criteria for ranking results
 - ▶ What do users click on after a given search? Search engines can learn what users are looking for by collecting queries and the resulting clicks.

nokia

Noin 186 000 000 tulosta (0,08 sekuntia)

[Mukautettu](#) >

[Nokia Online Kauppa](#)
[Nokia.fi/kauppa](#) Helppoa ja sujuvaa - osta puhelin ja lisälaitteet Nokian kaupasta.
Ilmainen autonavigointi ja teline - Ilmaiset karttapalvelut - [Lisälaitteet](#) - [Puhelimet](#)

[Nokia, Finland - Wikipedia, the free encyclopedia](#) ☆ - [[Käännä tämä sivu](#)]
Nokia is a town and a municipality on the banks of the Nokianvirta River (Kokemäenjoki) in the region of Pirkanmaa, some 15 kilometres (9 mi) west of ...
[en.wikipedia.org/wiki/Nokia,_Finland](#) - [Välimuistissa](#) - [Samankaltaisia](#)

[Nokia - Wikipedia, the free encyclopedia](#) ☆ - [[Käännä tämä sivu](#)]
Nokia Corporation OMX: NOK1V, NYSE: NOK, FWB: NOA3) is a Finnish ...
[en.wikipedia.org/wiki/Nokia](#) - [Välimuistissa](#) - [Samankaltaisia](#)

[Nokia 5700 XpressMusic – Wikipedia](#) ☆
Nokia 5700 XpressMusic on vuonna 2007 julkaistu nuorten musiikkipuhelin ...
[fi.wikipedia.org/wiki/Nokia_5700_XpressMusic](#) - [Välimuistissa](#) - [Samankaltaisia](#)
✚ Näytä lisää tuloksia kohteesta wikipedia.org

[Nokia \(nokia\) on Twitter](#) ☆ - [[Käännä tämä sivu](#)]
News and updates from **Nokia**. The main tweeps at the channels are @jussipekka & @JGallo02.
[twitter.com/nokia](#) - [Välimuistissa](#) - [Samankaltaisia](#)

[Ovi Musiikki - porttisi musiikin maailmaan](#) ☆
Aloitussivu · **Nokia** Ovi Player · Ovi Musiikki Unlimited **Nokia.com**; Copyright ©2010 **Nokia**. Kaikki oikeudet pidätetään.
[music.ovi.com/fi/fi/pc](#) - [Välimuistissa](#)

[YouTube - Lex Nokia anti-ad 2A: "Perustuslaki"](#) ☆
29. tammikuu 2009 ... Urkintalaki.fi:n masinoima Lex **Nokia** -lakiehdotuksen vastainen mainos 2a, " Perustuslaki".
[www.youtube.com/watch?v=0tDhemyzB3k](#) - [Välimuistissa](#) - [Samankaltaisia](#)

Example 6

- ▶ Detecting credit card fraud
 - ▶ Credit card companies typically end up paying for fraud (stolen cards, stolen card numbers)
 - ▶ Useful to try to detect fraud, for instance large transactions
 - ▶ Important to be adaptive to the behaviors of customers, i.e. learn from existing data how users normally behave, and try to detect 'unusual' transactions



Example 7

- ▶ Self-driving cars:
 - ▶ Sensors (radars, cameras) superior to humans
 - ▶ How to make the computer react appropriately to the sensor data?

SMARTER THAN YOU THINK

Google Cars Drive Themselves, in Traffic



Ramin Rahimian for The New York Times

Example 8

- ▶ Character recognition:
 - ▶ Automatically sorting mail (handwritten characters)
 - ▶ Digitizing old books and newspapers into easily searchable format (printed characters)



Example 9

- ▶ Recommendation systems ('collaborative filtering'):
 - ▶ Amazon: "Customers who bought X also bought Y" ...
 - ▶ Netflix: "Based on your movie ratings, you might enjoy..."

Challenge: One million dollars (\$1,000,000) prize money recently awarded!



		Seven		Fargo	Aliens	Leon		Avatar
Linda		4		5	5	1		2
			3		4	3		
Jack	1			4		1	5	1
Bill				?	4	1		?
Lucy			2	1	1		5	
John	1				1	4		5
		4				5		5
	2		3				3	

Example 10

- ▶ Machine translation:
 - ▶ Traditional approach: Dictionary and explicit grammar
 - ▶ More recently, *statistical* machine translation based on example data is increasingly being used

Google kääntäjä

Kielestä: suomi ▼



Kielelle: englanti ▼

Käännä


Tietojenkäsittelytieteen opinnot antavat erinomaisen pohjan työskentelylle kaikkialla, missä kehitetään tai sovelletaan tietotekniikkaa.

Käännös (suomi > englanti)


Computer studies provide an excellent foundation for the work, wherever applicable, or to develop information technology.

Example 11


- ▶ Online store website optimization:
 - ▶ What items to present, what layout?
 - ▶ What colors to use?
 - ▶ Can significantly affect sales volume
 - ▶ Experiment, and analyze the results! (lots of decisions on how exactly to experiment and how to ensure meaningful results)



Quantity: 1

 **Add to Cart**

or Buy now with

 **Two-Day 1-Click®—FREE**

Ship to:

☐ Add gift-wrap/note

Add to Wish List

Add to Shopping List

Example 12

- ▶ Mining chat and discussion forums
 - ▶ Breaking news
 - ▶ Detecting outbreaks of infectious disease
 - ▶ Tracking consumer sentiment about companies / products



Example 13

- ▶ Real-time sales and inventory management
 - ▶ Picking up quickly on new trends (what's *hot* at the moment?)
 - ▶ Deciding on what to produce or order (example: Jopo production moved from Taiwan to Finland for a quicker response to incoming sales data – YLE 10.6.2010)

Walmart 
Save money. Live better.

 PRISMA

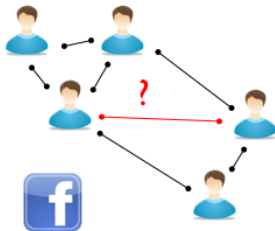
 K SUPERMARKET


OSTA VIISAAMMIN-OSTA NOPEAMMIN



Example 14

- Prediction of friends in Facebook, or prediction of who you'd like to follow on Twitter.



What about privacy?

- ▶ Users are surprisingly willing to sacrifice privacy to obtain useful services and benefits
- ▶ Regardless of what position you take on this issue, it is important to know what *can* and what *cannot* be done with various types information (i.e. what the dangers are)
- ▶ 'Privacy-preserving data mining'
 - ▶ What type of statistics/data can be released without exposing sensitive personal information? (e.g. government statistics)
 - ▶ Developing data mining algorithms that limit exposure of user data (e.g. 'Collaborative filtering with privacy', Canny 2002)

Course outline

- ▶ Introduction
- ▶ Data
 - ▶ data types and quality, preprocessing
 - ▶ similarity/distance measures, visualization
- ▶ Supervised learning
 - ▶ classification
 - ▶ regression
 - ▶ evaluation and model selection
- ▶ Unsupervised learning
 - ▶ clustering
 - ▶ anomaly detection

Related courses

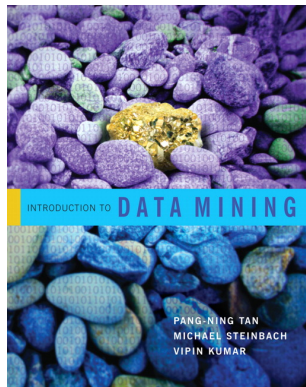
- ▶ Various continuation courses at CS (spring 2014):
 - ▶ Probabilistic Models (period III)
 - ▶ Supervised Machine Learning (period III)
 - ▶ Unsupervised Machine Learning (period IV)
 - ▶ Data Mining (period IV)
 - ▶ Seminar: Reinforcement Learning and Its Applications
- ▶ A number of other specialized courses at CS department
- ▶ A number of courses at maths+stats
- ▶ Lots of courses at Aalto as well

Practical details (1)

- ▶ Lectures:
 - ▶ 29 October (today) – 2 December
 - ▶ Tuesdays and Fridays at 10:15–12:00 in Exactum D122
 - ▶ Lecturer: Jyrki Kivinen
(Exactum B229a, jyrki.kivinen@cs.helsinki.fi)
 - ▶ Language: English
 - ▶ Based on parts of the course textbook (next slide)
 - ▶ Lecture slides available online soon after each lecture

Practical details (2)

- ▶ Textbook:
 - ▶ Tan, Steinbach, Kumar (2005): Introduction to Data Mining
 - ▶ This course covers (much of) chapters 1–5 and 8–10. There will be assigned reading each week
 - ▶ Although lectures and assigned reading from the textbook mostly overlap, the course requirements consist of the *union* of the two
 - ▶ Kumpula science library has a number of copies that can be borrowed and one reading room copy



Practical details (3)

- ▶ Exercises:
 - ▶ Learning by doing:
 - ▶ mathematical exercises (pen-and-paper)
 - ▶ computer exercises (with Matlab, Octave or R)
 - ▶ Problem set handed out every Wednesday (but problems may depend on material from the Friday lecture)
 - ▶ Deadline for handing in your solutions is Wednesday at 9:00am.
 - ▶ Exercise session (in which students discuss and present their solutions, credit awarded for indicating willingness present the solution): Fridays at 12:15–13.45 in B222 (starting 8 October)
 - ▶ Language of exercise sessions: English
 - ▶ Exercise points make up 40% of your total grade, must get at least half the points to be eligible for the course exam
 - ▶ Details will appear on the course web page

Practical details (4)

- ▶ Exercises *this week*:

- ▶ No regular exercise session this week.
- ▶ Instead: instruction on Matlab, Octave, and R. Choose either of the following:
 - ▶ Tuesday 29 October (today) at 12:15 in B221, or
 - ▶ Friday 2 November at 12:15 in B221

(roughly how many students are attending each session?)

- ▶ Voluntary, no points awarded. Recommended for everyone not previously familiar with Matlab, Octave, nor R.
 - ▶ You may instead use Python for your homework if you wish, but there will be no support from teachers
 - ▶ If you wish to use some other language, discuss it with the lecturer
- ▶ Tuesday session has a slight Matlab focus, and Friday a slight R focus, but either OK for either language.

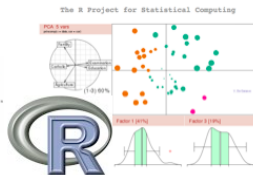
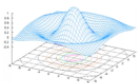
Practical details (5)

- ▶ Computer exercises: Choose one of
 - ▶ Matlab (dominant in computer science and engineering, commercial software)
 - ▶ Octave (free clone of Matlab, mainly compatible)
 - ▶ R (dominant in statistics, free software)

MATLAB
The Language of Technical Computing



Octave



Matlab, Octave, and R

- ▶ Common features:
 - ▶ Environments for numerical/statistical calculations
 - ▶ Scripts to automate (matlab/octave: .m files, R: .R files)
 - ▶ Native representations for matrices and vectors
 - ▶ Allow standard programming constructs: variables, functions, loops, conditional statements
 - ▶ Optimized for matrix and vector operations. Avoid explicit loops whenever possible!
- ▶ As always:
 - ▶ Use descriptive variable and function names
 - ▶ Indent your code to show the structure
 - ▶ Comment your code!
 - ▶ Write functions for any code snippets that you re-use

Practical details (6)

- ▶ Course exam:
 - ▶ 11 December at 9:00 (double-check a few days before the exam)
 - ▶ Constitutes 60% of your course grade
 - ▶ Must get a minimum of half the points of the exam to pass the course
 - ▶ Pen-and-paper problems, similar style as in exercises (also 'essay' or 'explain' problems)
- ▶ Note: To be eligible to take a 'separate exam' you need to first complete some programming assignments. These will be available on the course web page a bit later.
- ▶ Answering the exam problems in Finnish (or Swedish) is OK.
 - ▶ We will try to work out some kind of an *unofficial* machine learning dictionary during the course

Practical details (7)

► Grading:

- Exercises: (typically: 3 pen-and-paper and 1 programming problem per week)
 - Programming problem graded to 0–10 points
 - Pen-and-paper problems graded to 0–3 points
 - An extra point for being present in the homework session and willing to present the solution
 - Attendance in first week Matlab/Octave/R exercises: Voluntary, no points
- Exam: (4–5 problems)
 - Pen-and-paper: 0–6 points/problem (tentative)
- Rescaling done so that 40% of total points come from exercises, 60% from exam
- Half of all total points required for lowest grade, close to maximum total points for highest grade
- Note: Must get at least half the points of the exam, and must get at least half the points available from the exercises

Practical details (8)

- ▶ Prerequisites:
 - ▶ Mathematics: Basics of probability theory and statistics, linear algebra basics, real analysis
 - ▶ Computer science: Basics of programming (but no previous familiarity with Matlab, Octave, or R necessary)
- ▶ Prerequisites quiz!
 - ▶ For you to get a sense of how well you know the prerequisites
 - ▶ For me to get a sense of how well you (in aggregate!) know the prerequisites. Fully anonymous!

Practical details (9)

- ▶ Course material:
 - ▶ Webpage (public information about the course):
<http://www.cs.helsinki.fi/en/courses/582631/2013/s/k/1>
- ▶ Sign up in Ilmo (department registration system)
- ▶ Help?
 - ▶ Ask the assistants/lecturer at exercises/lectures
 - ▶ Contact assistants/lecturer separately

Coursera machine learning course

- ▶ In addition to the 'regular' and 'separate' exam tracks, a fully separate option is to take the **Coursera** online course on machine learning.

<https://www.coursera.org/course/ml>

This option will additionally require writing a short report and taking a small exam. Details will appear on the course web page.

Questions?