Introduction to Machine Learning

Lecturer: Teemu Roos
Assistant: Ville Hyvönen

Department of Computer Science
University of Helsinki

material created by Patrik Hoyer, Jyrki Kivinen and others

November 1st–December 16th 2016

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: tic-tac-toe

- ▶ How to program the computer to play tic-tac-toe?
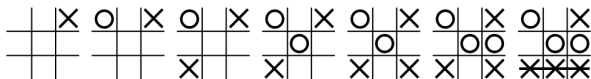
- ▶ Option A: The programmer writes explicit rules, e.g. 'if the opponent has two in a row, and the third is free, stop it by placing your mark there', etc (lots of work, difficult, not at all scalable!)

- ▶ 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.
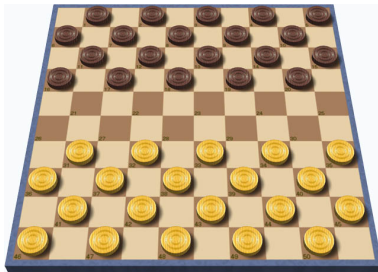
Example 2: spam filter

- ~~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   ···   ?
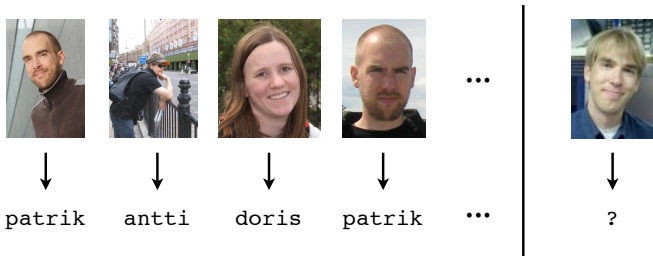
Problem setup

- 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!

Big Data

- one definition: data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges (Oxford English Dictionary)

- 3V: volume, velocity, and variety (Doug Laney, 2001)

- a database may be able to handle a lot of data, but you can't implement a machine learning algorithm as an SQL query

- on this course we do not consider technical issues relating to extremely large data sets

- basic principles of machine learning still apply, but many algorithms may be difficult to implement efficiently

Related scientific disciplines (2)

- Data mining
  - Trying to identify interesting and useful associations and patterns in huge datasets
  - Focus on scalable algorithms
  - Example: shopping basket analysis

- Statistics
  - historically, introductory courses on statistics tend to focus on hypothesis testing and some other basic problems
  - however there's a lot more to statistics than hypothesis testing
  - there is a lot of interaction between research in machine learning, data mining and statistics

Example 4

- Prediction of search queries

  - ~~The programmer provides a standard dictionary~~ (words and expressions change!)

  - Previous search queries are used as examples!
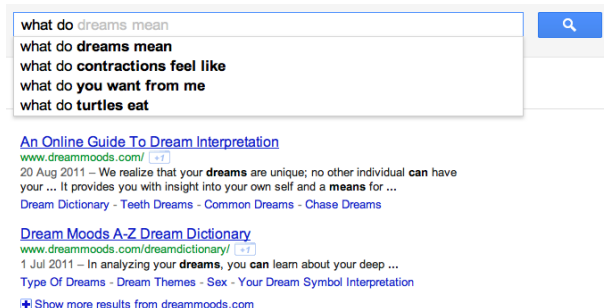
Example 5

- 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.

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

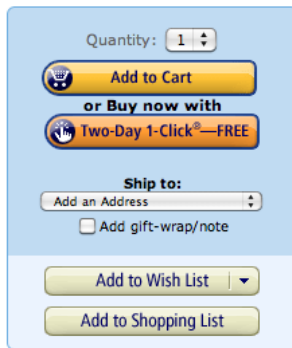Kielestä: suomi ▾ ⇄ Kielelle: englanti ▾   Käännä

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)

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 14

▶ Prediction of friends in Facebook, or prediction of who you'd like to follow on Twitter.
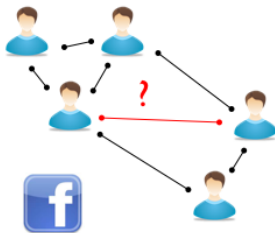
What about privacy?

- 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

- Ingredients of machine learning
    - task
    - models
    - data

- Supervised learning
    - classification
    - regression
    - evaluation and model selection

- Unsupervised learning
    - clustering
    - matrix decompositions

Related courses

- Various advaced CS courses (Spring 2017):
    - Probabilistic Models (period III, plus optional project)
    - Project in Practical Machine Learning (period III)
    - Advanced Course in Machine Learning (period IV)
    - Data Mining (self study, plus optional project)
    - Big Data Frameworks (period IV)
    - Seminar of Reinforcement Learning and Information Retrieval (period III)

- 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:
  - November 1st (today) – December 16th
  - Tuesdays and Fridays at 10:15–12:00 in Exactum CK112
  - Lecturer: Teemu Roos
    (Exactum A322, teemu.roos@cs.helsinki.fi)
  - Language: English
  - Based on parts of the course textbook (next slide)
  - (previous instances of this course have used different textbooks)

Practical details (2)

- Textbook:
    - authors: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
    - title: An Introduction to Statistical Learning – with Applications in R
    - publisher: Springer (2013, first edition)
    - web page: `www-bcf.usc.edu/~gareth/ISL/`

- we'll probably cover the whole book except splines and generalized additive models (GAMs) – and include some additional Bayesian stuff
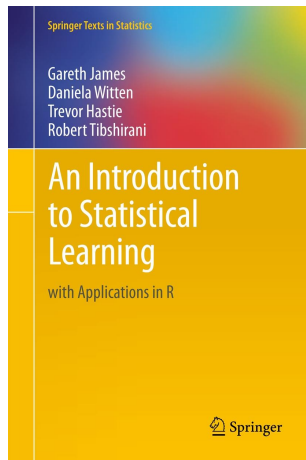
Practical details (3)

- Lecture material
  - this set of slides (by Hoyer/Kivinen/Roos) is intended for use as part of the actual lectures, together with the blackboard etc.
  - we will cover some topics in more detail than the textbook (and some less)
  - in particular some additional detail is needed for homework problems
  - both the selected parts of the textbook as well as additional material indicated on the course homepage are required material for the exam

Practical details (4)

- ▶ Exercises (Thu 14:15, Fri 12:15):
    - ▶ Two kinds:
        - ▶ mathematical exercises (pen-and-paper)
        - ▶ computer exercises (support given in R but Python is a good choice too)
    - ▶ Problem set handed out every Friday, focusing on topics from that week's lectures
    - ▶ Solutions returned at the exercise sessions
    - ▶ For those without an exercise group: crash the party in either group until we either create a new group or vacancy is created otherwise
    - ▶ If necessary, solutions can be turned in by email (ville.o.hyvonen@helsinki.fi) but pity the poor TA who checks piles of exercises weekly)
    - ▶ 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.

Practical details (5)

- Exercises *this week*:
    - Exercise session this week are voluntary R "tutorials".
    - Instruction on R and its features used on this course
    - Voluntary, no points awarded. Recommended for everyone not previously familiar with R.
    - **Bring you own laptop, with R (and possible RStudio) installed.**

Practical details (6)

- Course exam:
    - December 20th at 8:00am (sorry, not my choice)
    - 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. However since you are here at the lecture, this probably does not concern you.)

- You may answer exam problems also in Finnish or Swedish.

Practical details (8)

- ► Prerequisites:
  - ► Mathematics: Basics of probability theory and statistics, linear algebra (i.e., vectors and matrices) and real analysis (i.e., derivatives, etc.)
  - ► Computer science: Good programming skills (but no previous familiarity with R necessary)

Practical details (9)

- ▶ Course material:
  - ▶ Webpage (public information about the course):
    `www.cs.helsinki.fi/en/courses/582631/2016/s/k/1`

- ▶ NB: You should have signed up on the department registration system

- ▶ Help?
  - ▶ Ask the assistants/lecturer at exercises/lectures
  - ▶ Contact assistants/lecturer separately

We're in this together. Let's do it!