Foundation of Intelligent Systems

Part I: Statistical Machine Learning

mcuturi@i.kyoto-u.ac.jp
Answer a Few Questions

• What is this course about?
• What kind of tools will we use?
• Do we have to program?
• For starters... a first assignment
• Why is this useful for me?
What is this course about?

Doing Science & Engineering
Using Computers + Large databases
What is this course about?

• Science/engineering in 19th century

![Image of scientists in a laboratory](image1)

• Science/engineering in the 21st century

![Image of modern laboratory and computer equipment](image2)

Goal: Use **Fast Computers** to detect and exploit patterns in **Large databases**
A typical Machine Learning Task

The goal of the prize is to develop a predictive algorithm that can identify patients who will be admitted to the hospital within the next year, using historical claims data.

More than 71 million individuals in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association. Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions. Is there a better way? Can we identify earlier those most at risk and ensure they get the treatment they need? The Heritage Provider Network (HPN) believes that the answer is "yes".

To achieve its goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations, HPN is sponsoring the...

Started: 5:03 pm, Monday 4 April 2011 UTC  
Ends: 6:59 am, Wednesday 3 April 2013 UTC (729 total days)
This course is about adaptive machines that can learn

Before...

DATA ⇒ **Expert** (Doctor) ⇒ **Rule-based hard-coded not reusable** program

```
if age>36 then
  if cholesterol>105mg/L then
    ...
  else
    ...
```

... now, with machine learning

DATA ⇒ (**Meta-expert** (you!)→Algorithm ) ⇒ **Adaptive** Program
What kind of mathematical tools?

We will adopt a mathematical formalism to propose and study algorithms.

Probability & Statistics, Linear Algebra, Optimization
Mathematical Tools

- **Probability & Statistics** (*to handle uncertainty & randomness*)
  - Probability Spaces, Random variables
  - Expectation, variance, inequalities
  - Central limit theorem, convergence in probability

- **Linear Algebra** (*to handle high-dimensional problems*)
  - Matrix inverse, eigenvalues/vectors
  - Positive-definiteness.

- **Optimization** (*to give the best possible answer*)
  - convex programs,
  - lagrangean, Lagrange multipliers *etc.*
This is not a course about programming, but we will implement algorithms

I encourage you to use MATLAB
but you can use any other program (R, Python, etc...)

I do not recommend using C/C++ or other compiled languages.
For Starters...

Some simple ideas and a 1st assignment.
A function

\[(x-1)^4 - (x-3)^2 + (x-2)^3\]

a polynomial plotted between 0 and 4...
A function

\[(x-1)^4-(x-3)^2+(x-2)^3\]

... can be seen as a very detailed scatter plot.
A function

Yet, when less points are available...
A function

\((x-1)^4-(x-3)^2+(x-2)^3\)

can we still guess the whole blue line?
A partially observed function

Assume we only have the red points.
We can guess by using interpolating polynomials

Curve fitting tools can help us get back the original function. We can actually reconstruct it perfectly.
even if points are not evenly spaced...
Uncertainty in measurements

sometimes, we do not have access to the correct information...
but rather an information corrupted by “noise”.
Things become a lot more difficult

If we use standard tools...
Things become a lot more difficult

we might be very far from the original function.
Things become a lot more difficult

Can we handle **uncertainty** in a better way?
Quantify **how far** we might be from the true function?
**How many points** do we need to reconstruct a more **general** curve?
Does this work for surfaces in **higher dimensions**?
Things become a lot more difficult

First assignment - due Monday 18th 23:59

• Look for a definition of interpolation, e.g. check the wikipedia page.
• Do what I just did with Matlab and send me an email with the results:
  ○ Choose a function.. you can use fancier functions (\(\sin, \cos, \exp\ etc.\))
  ○ Plot it. Scatter plot a few points.
  ○ Use these points with the curve fitting tool. Interpolate & Compare.
• Finally: give me a hint of what might go wrong in higher dimensions?
To close this introduction...

Machine Learning will help you!
12 IT skills that employers can't say no to

Job hunters with these IT skills are assured of employment, now and in the future

By Mary Brandle
July 11, 2007 12:00 PM ET

Computerworld - Have you spoken with a high-tech recruiter or professor of computer science lately? According to observers across the country, the technology skills shortage that pundits were talking about a year ago is real (see "Workforce crisis: Preparing for the coming IT crunch").

"Everything I see in Silicon Valley is completely contrary to the assumption that programmers are a dying breed and being offshoreed," says Kevin...
...and the winner is

(See also "The top 10 dead (or dying) computer skills".)

1) Machine learning

As companies work to build software such as collaborative filtering, spam filtering and fraud-detection applications that seek patterns in jumbo-size data sets, some observers are seeing a rapid increase in the need for people with machine-learning knowledge, or the ability to design and develop algorithms and techniques to improve computers' performance, Scott says.

"It's not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes, which creates a fundamental problem of how you organize the data and present it to users."

Demand for these applications is expanding the need for data mining, statistical modeling and data structure skills, among others, Scott says. "You can't just wave your hand at some of these problems -- there are subtle differences in how the data structures or algorithms you choose impacts whether you get a reasonable solution or not," he explains.