Foundation of Intelligent Systems

Part I: Statistical Machine Learning

mcuturi@i.kyoto-u.ac.jp
Course Introduction

- The Promises of Big Data
- What kind of tools will we use?
- Do we have to program?
- For starters... a first assignment
- Why is this useful for me?
The Promises of Big Data
Data can help us predict when people will have to go to the hospital.
Small Businesses

- Data can help us predict the dynamics of restaurants’ popularity
• Data can help us predict who we can lend money to

www.lendingclub.com
Lending Money

- Data can help us predict who we can lend money to

LendingClub  Zopa  AQUSH

Download Loan Data
These files contain complete loan data, including the current loan status (Current, Late, F).
We have removed all personally identifiable information to protect our members’ privacy.

Download CSV  (44,533kb)

www.lendingclub.com
Movies

- Data can help us predict whether people will like a given movie.
All these problems have in common that...

Data is Available

all you have to do, is download it... and analyze it!
What we will do in 7 lectures

The graduate school has many courses on how to handle data. Check the course offerings.

In these 7 lectures, we will focus on 3 things:

• Present elementary tools: regression and classification

• Study the mathematical foundations of statistical learning theory:
  ○ Choose the right models, address computational issues,
  ○ Address the problem of overfitting.

• Introduce advanced topics: kernel methods, sparsity.
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.*
Programming

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

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

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.
Polynomial Interpolation

even if points are not evenly spaced...
Polynomial Interpolation

Graph showing data points and polynomial interpolations of different degrees.
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

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

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 15th 23:59 by email

• 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?