Computational Mathematics for Learning and Data Analysis: Introduction to the course

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

A.Y. 2022/23

Outline

Logistic

Motivation

Contents

Wrap up

- 1 course (9 CFU/ECTS)
- 1 program
- 1 exam
- 2 related but $\neq$ areas of computational mathematics $\Longrightarrow$ 2 lecturers:

Federico Poloni (Numerical methods)
Dipartimento di Informatica, room 343
050 2213143, mailto:federico.poloni@unipi.it
https://www.di.unipi.it/~fpoloni
Office hours (ricevimento): upon request

Antonio Frangioni (Optimization)
Dipartimento di Informatica, room 327
050 2212789, mailto:frangio@di.unipi.it
https://www.di.unipi.it/~frangio
Office hours (ricevimento): Tuesday 9:00 – 11:00

- Course Schedule
  - Wed 16:15 – 18:00 (Fib. C)
  - Thu 9:00 – 11:00 (Fib. C)
  - Fri 11:00 – 13:00 (Fib. C)

- Web page: https://elearning.di.unipi.it/course/view.php?id=307

- Team for lectures: https://teams.microsoft.com/l/team/19%3aRWlQjgwH67w-hqttkLXv6nqXSMoIsciPP9nqJDxD_Hg1%40thread.tacv2/conversations?groupId=5d8c2945-bf87-4ecb-97c8-6ce69f73b692&tenantId=c7456b31-a220-47f5-be52-473828670aa1

- Exam: project (groups of 2) + oral exam
  Projects either "ML" or "no-ML", but no difference in work and grading

Outline

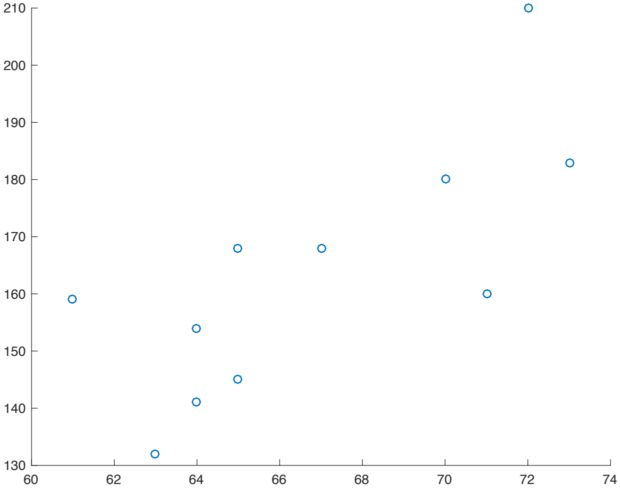
- Huge amounts of data is generated and collected, but one has to make sense of it in order to use it: learn (from) it
- Take something big (data) and therefore unwieldy and produce something small and nimble that can be used in its stead ("actionable")
- That's a (mathematical) model
- Word comes from "modulus", diminutive from "modus" = "measure": "small measure", "measure in the small" (small is good)
- Known uses in architecture: proving beforehand that the real building won't collapse (e.g., Filippo Brunelleschi for the Cupola of the Cathedral of Florence)
- Countless many physical models afterwards (planes, cars, ...), but mathematics is cheaper than bricks / wood / iron ...
- Yet, mathematical problems can be difficult, too, for various reasons (and, of course, only truly viable after computers)
- And most of them remain (likely) difficult for quantum computers, too, `https://www.smbc-comics.com/comic/the-talk-3`

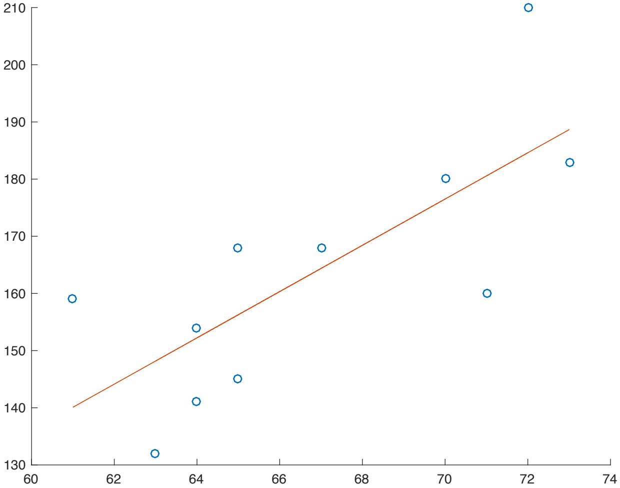
Choosing a mathematical model

- How a mathematical model should be:
  1. accurate (describes well the process at hand)
  2. computationally inexpensive (gives answers rapidly)
  3. general (can be applied to many different processes)

  Typically impossible to have all three!
- Two fundamentally different model building approaches:
  1. analytic: model each component of the system separately + their interactions, ($\approx$)accurate but hard to construct (need system access + technical knowledge)
  2. data-driven / synthetic: don't expect the model to closely match the underlying system, just to be simple and to ($\approx$)accurately reproduce its observed behaviour
- All models are approximate (the map is not the world), but for different reasons
- Analytic models: flexible shape, (relatively) few "hand-chosen" parameters
- Synthetic models: rigid shape, (very) many automatically chosen parameters
- Fitting: find the parameters of the model that best represents the phenomenon, clearly some sort of optimization problem (often a computational bottleneck)
- However, ML $\gg$ fitting: fitting minimizes training error $\equiv$ empirical risk, but ML aims at minimizing test error $\equiv$ risk $\equiv$ generalization error!

- A phenomenon measured by one number y is believed to depend on a vector $x = [\, x_1 \,, \, \ldots \,, \, x_n \,]$ of other numbers
- Available (hopefully, large) set of observations $(\, y^1 \,, \, x^1 \,), \ldots, (\, y^m \,, \, x^m \,)$
- Horribly optimistic assumption: the dependence is linear, i.e.,
  $y = \sum_{i=1}^{n} w_i x_i + w_0 = wx + w_0$
  for fixed $n+1$ real parameters $w = [\, w_0 \,, \, w_+ = [w_1 \,, \, \ldots \,, \, w_n \,] \,]$
- But $y^h = w_+ x^h + w_0$ for all $h = 1, \ldots, m$ is not really true for any w and w_0
- Find the w for which it is less untrue (Linear Least Squares):
  $$y = \begin{bmatrix} y^1 \\ \vdots \\ y^m \end{bmatrix} \;,\; X = \begin{bmatrix} 1 & x^1 \\ \vdots & \vdots \\ 1 & x^m \end{bmatrix} \;,\; \min_w \mathcal{L}(\, w \,) = \|\, y - Xw \,\|$$
- Minimize loss function $\mathcal{L}(\, w \,) = \|\, y - Xw \,\| \equiv$ empirical risk $\equiv$ how much the model fails the predict the phenomenon on the available observations
- Simple closed formula: $XX^T w = X^T y \implies w = (XX^T)^{-1} X^T y$


210
200
190
180
170
160
150
140
130
60
62
64
66
68
70
72
74



- In `Matlab`, this is just `c = y / X`
- Trade-off: very simple fitting for exceedingly crude model $\Longrightarrow$ high risk
- Then, of course Nonlinear Estimation ...

Example 2: Low-rank approximation

- A (large, sparse) matrix $M \in \mathbb{R}^{n \times m}$ describes a phenomenon depending on pairs (e.g., objects chosen from customers)
- Find "tall and thin" $A \in \mathbb{R}^{n \times k}$ and "fat and large" $B \in \mathbb{R}^{k \times m}$ ($k \ll n, m$) s.t. $M \approx AB$ $\equiv$ find a few features that describe most of users' choices

$$\boxed{M} \approx \boxed{A} \cdot \boxed{B} \quad , \quad \min_{A,B} \mathcal{L}(\, A\, ,\, B\,) = \|\, M - AB \,\|$$

- Minimize loss $\mathcal{L}(\, A\, ,\, B\,) = \|\, M - AB \,\| \equiv$ "amount of unexplained choices"
- Many applications (neural networks, community analysis, …)
- A, B can be obtained from eigenvectors of $M^T M$ and $M M^T$ …

Example 2: Low-rank approximation

- A (large, sparse) matrix $M \in \mathbb{R}^{n \times m}$ describes a phenomenon depending on pairs (e.g., objects chosen from customers)
- Find "tall and thin" $A \in \mathbb{R}^{n \times k}$ and "fat and large" $B \in \mathbb{R}^{k \times m}$ ($k \ll n, m$) s.t. $M \approx AB$ $\equiv$ find a few features that describe most of users' choices

$$\boxed{M} \approx \boxed{A} \cdot \boxed{B} \quad , \quad \min_{A,B} \mathcal{L}(\,A\,,\,B\,) = \big|\big|\,M - AB\,\big|\big|$$

- Minimize loss $\mathcal{L}(\,A\,,\,B\,) = \big|\big|\,M - AB\,\big|\big| \equiv$ "amount of unexplained choices"
- Many applications (neural networks, community analysis, ...)
- A, B can be obtained from eigenvectors of $M^T M$ and MM^T ...
  ... but that's a huge, possibly dense matrix
- Efficiently solving this problem requires:
  1. low-complexity computation (of course)
  2. avoiding ever explicitly forming $M^T M$ and MM^T (too much memory)
  3. exploiting structure of M (sparsity, similar columns, ...)
  4. ensuring the solution is numerically stable

Black/white image $\equiv M$ with color intensities $\in [0, 1]$

Original (512×512)

$k = 1$

$k = 10$

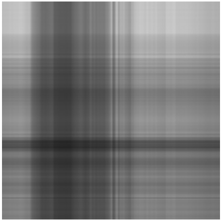
$k = 25$

$k = 50$

$k = 100$

Black/white image $\equiv M$ with color intensities $\in [0, 1]$


Original (512×512)


$k = 1$

$k = 10$

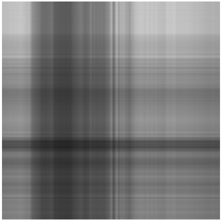
$k = 25$

$k = 50$

$k = 100$

Black/white image $\equiv M$ with color intensities $\in [0, 1]$

Original (512×512)



$k = 1$

$k = 10$

$k = 25$

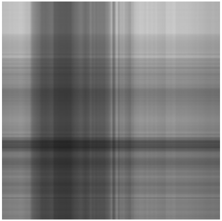
$k = 50$

$k = 100$

Low-rank approximation for image compression

Black/white image $\equiv M$ with color intensities $\in [0, 1]$

Original (512×512)



$k = 1$

$k = 10$

$k = 25$

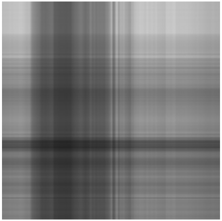
$k = 50$

$k = 100$

Low-rank approximation for image compression

Black/white image $\equiv M$ with color intensities $\in [0, 1]$

Original (512×512)



$k = 1$

$k = 10$

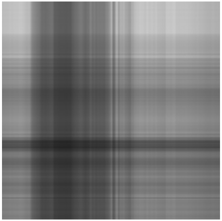
$k = 25$

$k = 50$

$k = 100$

Black/white image $\equiv M$ with color intensities $\in [0, 1]$

Original (512×512)



$k = 1$

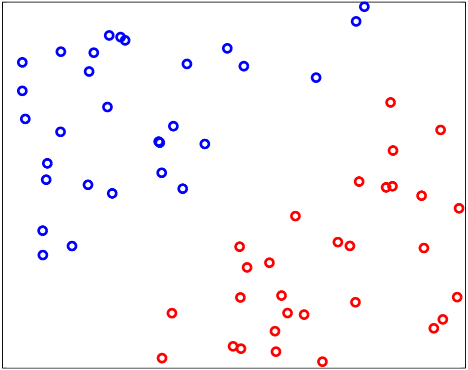
$k = 10$

$k = 25$

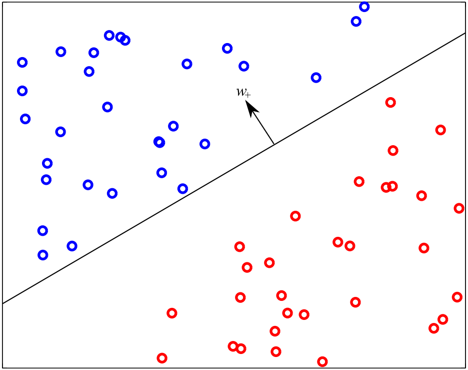
$k = 50$

$k = 100$

- Same setting as Example 1 but $y^h \in \{1, -1\}$ (have cancer or not)

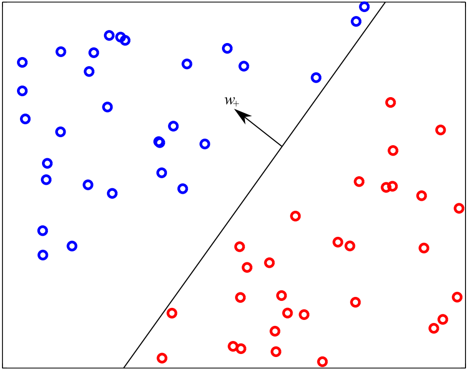


- Same setting as Example 1 but $y^h \in \{1, -1\}$ (have cancer or not)



- Want to linearly separate the two sets (diagnose the next patient)
- Countless many applications (medical diagnosis, OCR, spam filtering, fraud detection, marketing, image processing . . .)

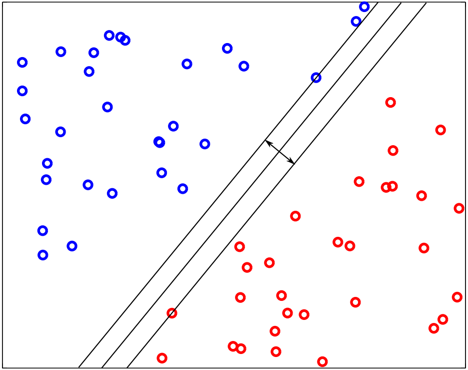
- Same setting as Example 1 but $y^h \in \{1, -1\}$ (have cancer or not)



- Want to linearly separate the two sets (diagnose the next patient)
- Countless many applications (medical diagnosis, OCR, spam filtering, fraud detection, marketing, image processing . . .)
- But which hyperplane do we choose?

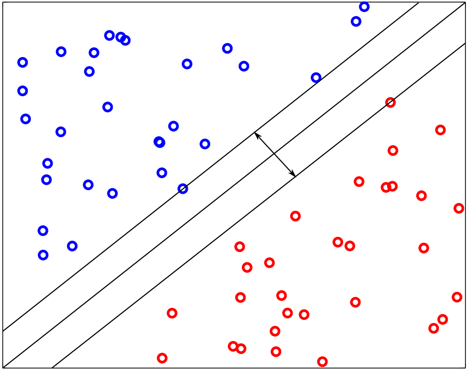
Example 3: Support Vector Machines

- Same setting as Example 1 but $y^h \in \{1, -1\}$ (have cancer or not)



- Want to linearly separate the two sets (diagnose the next patient)
- Countless many applications (medical diagnosis, OCR, spam filtering, fraud detection, marketing, image processing . . .)
- But which hyperplane do we choose?
- Intuitively, the margin is important (and theory supports the intuition)

- Same setting as Example 1 but $y^h \in \{1, -1\}$ (have cancer or not)



- Want to linearly separate the two sets (diagnose the next patient)
- Countless many applications (medical diagnosis, OCR, spam filtering, fraud detection, marketing, image processing . . .)
- But which hyperplane do we choose?
- Intuitively, the margin is important (and theory supports the intuition)
- Larger margin $\implies$ more "robust" classification

- Distance of // hyperplanes $(\, w_+ \,,\, w_0 \,)$ and $(\, w_+ \,,\, w_0' \,)$ is $|\, w_0 - w_0' \,| \,/\, ||\, w_+ \,||$
- We can always take the hyperplane in "the middle" + scale w
  $\implies\; w_+ x^h + w_0 \geq 1$ if $y^h = 1 \;\;,\;\; w_+ x^h + w_0 \leq -1$ if $y^h = -1$
- The maximum margin separating hyperplane is the solution of
  $$\min_w \{\, ||\, w_+ \,||^2 \;:\; y^h(w_+ x^h + w_0) \geq 1 \quad h = 1, \dots, m \,\}$$
  (margin $= 2 \,/\, ||w_+||$, "2" because I say so), assuming any exists
- What if it does not? Support Vector Machine
  $$\text{(SVM-P)} \quad \min_w \,||\, w_+ \,||^2 + C\mathcal{L}(\, w \,) = \textstyle\sum_{h=1}^m \max\{\, 1 - y^h(w_+ x^h + w_0) \,,\, 0 \,\}$$
  C weighs loss (of separation) against margin = regularization $R(\, w \,)$ (how?)
- $\mathcal{L}$ convex but nondifferentiable: reformulation
  $$\text{(SVM-P)} \quad \min_{w,\xi} \,||\, w_+ \,||^2 + C \textstyle\sum_{h=1}^m \xi_h$$
  $$y^h(w_+ x^h + w_0) \geq 1 - \xi_h \;\;,\;\; \xi_h \geq 0 \qquad h = 1, \dots, m$$
  "complex" but smooth (linear) constraints

- Equivalently, one can solve the dual problem (??? what ???)

$$\text{(SVM-D)} \quad \begin{array}{ll} \max_{\alpha} \ \sum_{h=1}^{m} \alpha_h - \frac{1}{2} \sum_{h=1}^{m} \sum_{k=1}^{m} \alpha_h \langle\, x^h \,,\, x^k \,\rangle \alpha_k & \\ \sum_{i=1}^{m} y^h \alpha_h = 0 & \\ 0 \leq \alpha_h \leq C & h = 1, \ldots, m \end{array}$$

  a convex constrained quadratic program, but with "simple constraints"

- Solve one problem by solving an apparently different one:
  α^* optimal for (SVM-D) $\implies$ $w_+^* = \sum_{h=1}^{m} \alpha_h^* y^h x^h$ optimal for (SVM-P)

- Dual formulation $\implies$ kernel trick: input space $\rightsquigarrow$ (larger) feature space

$$\langle\, x^h \,,\, x^k \,\rangle \ \rightsquigarrow \ \langle\, \phi(x^h) \,,\, \phi(x^k) \,\rangle$$

  where points are hopefully "more linearly separable"

- Feature space can be infinite-dimensional, provided that scalar product can be (efficiently) computed

- Efficient algorithms: (SVM-P) or (SVM-D) (or both), complexity, ...

Outline

Logistic

Motivation

Contents

Wrap up

- ...of Computational Mathematics for Learning and Data Analysis

- . . . of Computational Magic for Learning and Data Analysis

- There are two main quests in the course:
  1. get a general understanding of several different classes of numerical algorithms and their underlying mathematical principles
  2. be able to actually implement, debug, and tune a few of them
- Algorithms are mathematical objects $\Longrightarrow$ reasoning about algorithms often is proving theorems (+ some hand-waving)
- All the more when the algorithms deal with nontrivial mathematical objects
- This is (mostly) done in the optional "Mathematically speaking" slides
- Learning theorems' proofs by heart is not a subject of the exam, not even the few (very simple) ones we'll actually see in details during lectures
- But you will have a lot more fun if you face side quests seriously
- Exercises are there for the same reason

- Linear algebra and calculus background
- Unconstrained optimization and systems of equations
- Direct and iterative methods for linear systems and least-squares
- Numerical methods for unconstrained optimization
- Iterative methods for computing eigenvalues
- Constrained optimization and systems of equations
- Duality (Lagrangian, linear, quadratic, conic, ...)
- Numerical methods for constrained optimization
- Software tools for numerical computations (Matlab, Octave, ...)
- Sparse hints to AI/ML applications

- Slides prepared by the lecturers + recording of lectures
- `Matlab` programs + data
- L.N. Trefethen, D. Bau *Numerical Linear Algebra*, SIAM, 1997
- J. Demmel *Applied Numerical Linear Algebra*, SIAM, 1996
- S. Boyd, L. Vandenberghe *Convex optimization*, Cambridge Un. Press, 2008 (`http://web.stanford.edu/~boyd/cvxbook/`)
- L. Eldén *Matrix Methods in Data Mining and Pattern Recognition*, SIAM, 2007
- M.S. Bazaraa, H.D. Sherali, C.M. Shetty *Nonlinear programming: theory and algorithms*, Wiley & Sons, 2006
- D.G. Luenberger, Y. Ye *Linear and Nonlinear Programming*, Springer International Series in Operations Research & Management Science, 2008
- J. Nocedal, S. Wright *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, 2006
- Lecture notes in preparation, maybe available before the end of the course

Outline

Logistic

Motivation

Contents

Wrap up

- Learning as a computational, hence mathematical, process
- Mathematical foundations of many important learning processes
  $\equiv$ nonlinear optimization and numerical analysis techniques
- Easy problems (linear, quadratic, conic, convex) or local optima, because size is huge (hard because large, not hard because hard)
- Besides, in ML the global optimal solution can be bad!
- Emphasis on what can be done by linear algebra
- Focus on methods and software tools, theory only as needed to understand
- Applications to be seen in "Machine Learning" and/or "Data Mining" (in parallel, you can/are supposed to do it, we talk to each other)