Lecture 2 - Mathematical tools for machine learning

Advanced course in Statistical Machine Learning:
Theory and Applications

Stéphane Canu
`stephane.canu@axiom.anu.edu.au`

`asi.insa-rouen.fr/~scanu`

National ICT of Australia

and

Australian National University

Overview

- vector spaces
  - we are learning functions
  - defining norms and dot products arrouns these function
  - a learning algo. provides a sequence of functions
- optimization
  - learning is optimizing some criterion
  - with some constrains
- probabilities
  - statistical learning theory
- matrices
  - for practical reason they are evrywhere

Thanks to Alex Smola and S.V.N. “Vishy” Vishwanathan for intial version of slides

(Real) vector space

$$(\mathcal{F}, +, \times)$$

- a set $\mathcal{F}$
- an internal operation +
- an external operation on $\mathbb{R}$: $\times$

required properties

1. $x + y = y + x$
2. $x + (y + z) = (x + y) + z$
3. $\forall x, y \in \mathcal{F}, \exists z \in \mathcal{F}$ such that $x + z = y$
4. $(\alpha\beta) \times x = \alpha(\beta \times x)$
5. $(\alpha + \beta) \times x = \alpha \times x + \beta \times x$
6. $\alpha(x + y) = \alpha x + \alpha y$
7. $1 \times x = x$

operator overloading for + and $\times$

Examples of real vector space

1. the real numbers $\mathbb{R}$
2. the set of all finite collections of real numbers (a vector) $\mathbb{R}^n$
3. the set of sequences $\mathbb{R}^\infty$
4. the set of sequences such that $\sum_{i=1}^{\infty} x_i^2 < \infty$
5. the set of continous functions $C^0(\Lambda)$ on a domain $\Lambda \subset \mathbb{R}^d$
6. the set of infinitely derivable functions $C^\infty(\Lambda)$ defined on $\Lambda \subset \mathbb{R}$
7. the set of all polynomials $\mathcal{P}$

Not a real vector space

1. the rational numbers (but it is a V.S. over $\mathbb{Q}$)
2. positive functions (defined through its domain)
3. $\{x < 1\}$

Some properties of vectorial spaces

- basis

  Distinguish the finite and the infinite case

  - **independence :** A finite familly of vectors $\mathcal{B} = \{x_1, ..., x_n\}$ is independent if $\sum_{i=1}^{n} \alpha_i x_i = 0 \Rightarrow \alpha_i = 0$ for all i
  - **independence :** An infinite familly of vectors $\mathcal{B}$ is independant if all of its finite sub collections are independent
  - **span :** the span of a familly of vectors is the set of all finite linear combinations of its members
  - **basis :** A family of vectors $\mathcal{B}$ is called a basis if it is independent and generative- $span\mathcal{B} = \mathcal{F}$

- vectorial sub space : the set spanned by some vectors
- dimension : minimum number of elements to get a basis : $\dim(E) = \text{card}(\mathcal{B})$

- finite - infinite - countable or not

distance and norm

Metric a two variable function from a set $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}^+$ is a metric if it satisfies $\forall x, y \in \mathcal{F}$

- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$ **(symmetric)**
- $d(x, y) \leq d(x, z) + d(z, y)$ **(Triangle inequality)**

Metric space is a pair $(\mathcal{F}, d)$, where $\mathcal{F}$ is a set and d is a metric

Norm a Function from a vector space $\mathcal{F}$ into $\mathbb{R}$ is a norm if it satisfies $\forall x \in \mathcal{F}$

- $||x|| = 0$ if and only if $x = 0$
- $||\alpha x|| = |\alpha| ||x||$ **(Scaling)**
- $||x + y|| \leq ||x|| + ||y||$ **(Triangle inequality)**

$\rightarrow$A norm not satisfying the first condition is called a pseudo norm

Normed space is a pair $(\mathcal{F}, ||\cdot||)$, where $\mathcal{F}$ is a vector space and $||\cdot||$ is a norm

A norm induces a metric via $d(\mathbf{x}, \mathbf{y}) := ||\mathbf{x} - \mathbf{y}||$

Example of distances and norms

$x = (x_1, x_2) \in \mathbb{R}^2 = \mathcal{F}$

- $\|x\|_1 = |x_1| + |x_2|$ **(city block distance)**
- $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$ **(Euclidean)**
- $\|x\|_p = (|x_1|^p + |x_2|^p)^{1/p}$, $1 < p < \infty$
- $\|x\|_\infty = \max\{|x_1|, |x_2|\}$

Let's have a look at the unit balls : $\left\{x \;\middle|\; \|x\| \leq 1\right\}$

$\mathcal{F} = C([\pi, \pi])$

- $\|x\|_1 = \int_{-\pi}^{\pi} |x(t)| dt$
- $\|x\|_2 = \sqrt{\int_{-\pi}^{\pi} x(t)^2 dt}$
- $\|x\|_\infty = \max_{t \in [\pi, \pi])}\{|x(t)|\}$

Let's have a look at the unit balls around function $\sin(t)$:

$\left\{x \;\middle|\; \|x - \sin\| \leq 1\right\}$

convergenceS

- a sequence $x_1, x_2, ..., x_n, ...$ converge to x
  - metric space

$$\forall \varepsilon > 0, \exists n_0 \text{ such that } \forall n > n_0 \Rightarrow d(x_n, x) \leq \varepsilon$$

  - normed space

$$\forall \varepsilon > 0, \exists n_0 \text{ such that } \forall n > n_0 \Rightarrow \|x_n - x\| \leq \varepsilon$$

$$\boxed{\lim_{n\to\infty} x_n = x \Rightarrow \lim_{n\to\infty} \|x_n - x\| = 0}$$

- for functions a sequence $f_1(t), f_2(t), ..., f_n(t), ...$ converge to $f(t)$
  - simple (no norm) - almost everywhere or pointwise

$$\forall t \in \Lambda, \;\; \lim_{n\to\infty} f_n(t) = f(t)$$

  - unifom $\|f\|_\infty$

$$\lim_{n\to\infty} \max_{t\in\Lambda} |f_n(t) - f(t)| = 0$$

Hilbert spaces and Scalar product

- **scalar product** a Function from a vector space $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}$ is a Scalar product if it satisfies $\forall x, y \in \mathcal{F}$
  - $\langle x, x \rangle \geq 0$ **(positivity)**
  - $\forall y, \langle x, y \rangle = 0$ if and only if $x = 0$ **(nondegenerate)**
  - $\langle x, y \rangle = \langle y, x \rangle$ **(symmetry)**
  - $\langle x, \alpha y + z \rangle = \alpha \langle x, y \rangle + \langle x, z \rangle$ **(Linearity)**

  $\rightarrow$A scalar product not satisfying the first condition is called an inner product

- induced norm $\|x\| := \sqrt{\langle x, x \rangle}$

  cool in the quadratic case : $\|x\|^2 = \langle x, x \rangle$

- Hilbert space is a pair $(\mathcal{F}, \langle \cdot, \cdot \rangle)$, where $\mathcal{F}$ is a vector space, $\langle \cdot, \cdot \rangle$ is a scalar product and $\mathcal{F}$ is complete with respect to the induced norm

a scalar product is bilinear

Examples of Hilbert spaces

- $\mathbb{R}^n$, (any finite dimensional v.s. Euclidian space) $\langle x, y\rangle = x^\top y$
- the set of square matrices of dim n, $\langle A, B\rangle = tr(A^\top B)$
- ℓ^2 the set of square sumable sequences $\langle x, y\rangle = \sum_{i=1}^\infty x_i y_i$
- $\mathcal{P}_k$ the set of polynomials of order lower or equals to k,
- $L^2(\Lambda)$, the set of square integrable functions $\langle x, y\rangle = \int x(t)y(t)dt$

$$\int_\Lambda f(t)^2 dt < \infty$$

Not a hilbert space

- L^1

$$\int_\Lambda |f(t)|dt < \infty$$

- the set of bounded functions L^∞ (**no scalar product**)
- Span$\{f(x_i), i \in \mathbb{N}\}$ (**not complete**)

When only the completion is missing, it is called pre-Hilbertian

How to “compare” objects

map $\mathcal{F} \longrightarrow \mathbb{R}$ or $\mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$

- topology — convegence structure
- distance — similarity
- norm — size (enregy)
- scalar product — correlation

- $||x|| := \sqrt{\langle x, x \rangle}$
- $d(x,y) := ||x - y||$
- $\mathcal{B}_x(r) := \{y | d(x,y) < r\}$

- $||x - y||^2 = ||x||^2 + ||y||^2 - 2\langle x, y \rangle \qquad = 2(1 - \langle x, y \rangle)$
- $|\langle x, y \rangle| \leq ||x|| ||y||$ **(Cauch Schwartz inequality)**

measure $\mathcal{F}$ objects through a map $\mathcal{F} \longrightarrow \mathbb{R}$
the set of all possible **linear and continuous** measures is the dual $\mathcal{F}'$
Example : what is the dual of $\mathbb{R}$?

An important example : the evaluation functional

- $f(t)$ have to mean something

$$f = g \Rightarrow \forall t \in \Lambda, f(t) = g(t)$$

$$\begin{array}{rccc} \delta_t : & \mathcal{F} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & \delta_t f = f(t) \end{array}$$

$L^2(\Lambda)$ is not ok ! $\boxed{\mathcal{F} \subset \mathbb{R}^\Lambda}$

- δ_t is a linear functional

$$\delta_t(\alpha f + g) = \alpha f(t) + g(t)$$

- if it is continous, represent δ_t by a function $k_t \in \mathcal{F}$

$$f(t) = \delta_t f = \langle f, k_t \rangle$$

Learning is functional optimization

- optimality principle
- convexity
  - unicity of the solution
  - efficent algorithms
- non convex
  - difficult problem
- minimization with constrains
  - lagrangian
  - KKT optimality conditions

Learning problems

in $x \in \mathbb{R}^n$

- objective function

$$\begin{array}{rccc} J: & \mathcal{F} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & J(f) \end{array}$$

- optimization (Weierstrass theorem) if Λ is compact, J derivable and convex :

$$\min_{x \in \Lambda \subset \mathbb{R}^n} J(x) \qquad \Leftrightarrow \qquad \text{find } x^* \text{ such that } \nabla J(x^*) = 0$$

- equality constraints $\left\{ \begin{array}{rl} \displaystyle\min_{x} & J(x) \\ \textbf{such that} & g_i(x) = 0, \quad i = 1, k \end{array} \right.$ **(Lagrange)**
- equality constraints $\left\{ \begin{array}{rl} \displaystyle\min_{x} & J(x) \\ \textbf{such that} & g_i(x) \leq 0, \quad i = 1, k \end{array} \right.$ **(KKT)**
- Both (**Karush Kuhn Tucker**)

convexity and derivatives

- derivatives (finite case)
  - f is not derivable : subdifferential at x (the set of subgradients)

$$\partial J(x) = \left\{ g \in \mathbb{R}^n \;\middle|\; J(y) \geq J(x) + g^\top (y - x) \right\}$$

- convexity
  - convex set, let $\mathcal{F}$ be a vector space, let $K \subset \mathcal{F}$. K is convex iff

$$\forall \lambda \in [0,1], \forall x, y \in K, \text{ we have } \lambda x + (1-\lambda) y \in K$$

    examples : unit ball, subdiffential...
  - convex function $f : \mathcal{F} \longrightarrow \mathbb{R}$

$$\forall \lambda \in [0,1], \forall x, y \in \mathcal{F}, \text{ we have } f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$$

    examples : linear functions, $\exp^x, x^2$, max, norms, log partition...
  - **convex (set + objective + constraints) $\Rightarrow$ unique solution exists**

Optimization : functional derivative

$\mathcal{F}$ a Hilbert space embeded with $\langle \cdot, \cdot \rangle$ and such that $f(t) = \langle f, k_t \rangle$

$$\begin{array}{rccc} J: & \mathcal{F} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & J(f) \end{array}$$

$$\boxed{\min_{f \in \mathcal{F}} J(f) \Leftrightarrow \text{ find } f^* \text{ such that } J'(f^*) = 0}$$

The gateau differential of the functional J in the direction g is the following limit if it exists

$$dJ(f, g) = \lim_{\alpha \to 0} \frac{J(f + \alpha g) - J(f)}{\alpha}$$

example

$$J(f) = \frac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$$

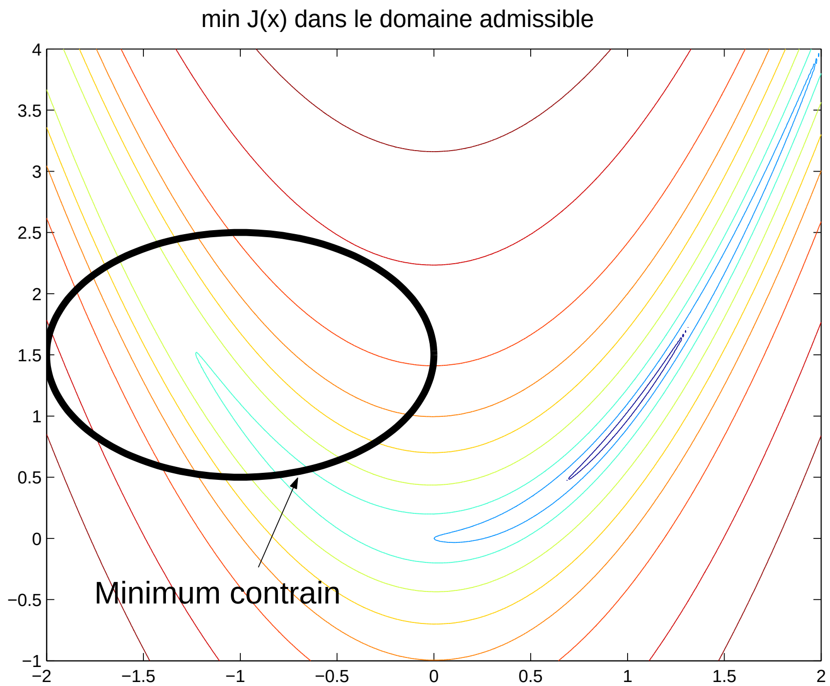
Example of functional derivative

- $J(f+\alpha g) = \frac{1}{2}\sum_{i=1}^{n}(f(x_i)+\alpha g(x_i)-y_i)^2 + \frac{\lambda}{2}\|f+\alpha g\|^2$
- $(f(x_i)+\alpha g(x_i)-y_i)^2 = (f(x_i)-y_i)^2+\alpha^2(g(x_i))^2+2\alpha(f(x_i)-y_i)g(x_i)$
- $\|f+\alpha g\|^2 = \|f\|^2+\alpha^2\|g\|^2+2\alpha\langle f,g\rangle$

$$\begin{aligned}
\frac{J(f+\alpha g)-J(f)}{\alpha} &= \sum_{i=1}^{n}(f(x_i)-y_i)g(x_i)+\lambda\langle f,g\rangle + \underbrace{\alpha((g(x_i))^2+\lambda\|g\|^2}_{\to 0} \\
&= \langle \underbrace{\sum_{i=1}^{n}(f(x_i)-y_i)k_{x_i}+\lambda f}_{J'(f)}, g\rangle
\end{aligned}$$

$$\boxed{J'(f)=0 \Leftrightarrow f(x)=\sum_{i=1}^{n}a_i k_{x_i}(x)},\ a_i = \frac{1}{\lambda}(f(x_i)-y_i)$$

minimizing with constraints : eliminate constraints



$$\begin{cases} & \min_{x \in \mathbb{R}^2} \quad J(x) \\ \text{such that} & A(x) = 0 \end{cases}$$

$$\Leftrightarrow$$

$$\min_{x} \max_{\lambda} \mathcal{L}(x, \lambda)$$ Lagrangien

$$\mathcal{L}(x, \lambda) = J(x) + \lambda A(x)$$

$$\begin{cases} \min_x \ J(x) \\ \text{such that } A(x) \leq 0 \end{cases} \Leftrightarrow \begin{cases} \nabla J(x) + \lambda^\top \nabla A(x) = 0 \\ \lambda^\top A(x) = 0, \quad \lambda > 0 \end{cases}$$ KKT conditions

λ represents the importance of the constraint in the solution

either $\lambda_i = 0$ or $A_i(x) = 0$

minimizing with constraints : dual formulation

- Optimality conditions : $x \in \mathbb{R}^n$

$$\begin{cases} \min_x \;\; J(x) \\ \text{such that } A(x) = 0 \end{cases} \Leftrightarrow \begin{cases} \min_x \max_\lambda \underbrace{J(x) + \lambda^\top A(x)}_{\color{blue}{Lagrangian}} \\ \end{cases}$$

- Phase 1

$$\nabla J(x) + \lambda^\top \nabla A(x) = 0 \qquad \Leftarrow \qquad \text{find a function } \Psi \text{ such that } x = \Psi(\lambda)$$

- phase 2 : $\lambda \in \mathbb{R}^k$

$$\max_\lambda J(\Psi(\lambda)) + \lambda^\top A(\Psi(\lambda))$$

- exemple $J(x) = x_1^2 - x_2$ and $A(x) = x_1^2 + x_2^2 - 1$

Probability

- set of events Ω : is it countable or not (discrete or continuous)
- discrete case : probability $\mathbb{P}(\omega)$
- continuous case : $\mathbb{P}(\omega) = 0$!
  - $\mathbb{P}(subset)$, e.g. $\Omega = \mathbb{R}$, $\quad F(x) = \mathbb{P}(\omega < x)$ cumulative function
  - no probability but density $f(x) = F'(x)$
- unified vue : measure

$$d\mu(x) = \begin{cases} \mathbb{P}(x) & \text{discrete case : probability} \\ f(x)dx & \text{continuous case : density} \end{cases}$$

Notation abuse - $\mathbb{P}(x)$ instread of $d\mu(x)$

Random variable

- functions : $X : \Omega \longrightarrow E = \mathbb{R}$ or $\mathbb{N}$ or $\{0,1\}$ or...
- E is a v.s. countable or not ?
- $\forall \mathcal{A} \subset E, \quad \mathbb{P}(X \in \mathcal{A}) := \mathbb{P}(X^{-1}(\mathcal{A}))$

- expectation - it is a linear operator from E to $\mathbb{R}$

$$\mathbb{E}(X) = \int x d\mu(x) = \begin{cases} \sum_i x_i \mathbb{P}(x_i) \\ \int x f(x) dx \end{cases}$$

discrete case : sum
continuous case : integral

- variance $V(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$

$$V(aX) = a^2 V(X)$$

Random variables

- joint law $\mathbb{P}(x, y)$ (discrete and/or continuous)
- Marginal $\mathbb{P}(x) = \int \mathbb{P}(x, y)dy$
- independance

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$$

- dependance : conditional laws and conditional expectation

$$\mathbb{P}(x|y) := \frac{\mathbb{P}(x, y)}{\mathbb{P}(y)} \qquad \mathbb{P}(y|x) := \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)}$$

$$\mathbb{E}(y|x) = \int y\mathbb{P}(y|x)dy = \frac{\int y\mathbb{P}(x, y)dy}{\mathbb{P}(x)}$$

- Bayes theorem

$$\mathbb{P}(y|x) = \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)}$$

example

AIDS-Test : We want to find out how likely it is that a patient *really* has AIDS (event X) if the test is positive (event Y)

- Roughly 0.1% of all Australians are infected ($\Pr(X) = 0.001$)
- The probability of a false positive is say 1% ($\Pr(Y|\overline{X}) = 0.01$ and $\Pr(Y|X) = 1$)
- By Bayes' rule

$$\begin{aligned}\Pr(X|Y) &= \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|\overline{X})\Pr(\overline{X})} \\ &= \frac{1 \times 0.001}{1 \times 0.001 + 0.01 \times 0.999} = 0.091\end{aligned}$$

- The probability of having AIDS even when the test is positive is just 9.1% !

Sample

- $X_1, X_2, ...X_n$ is i.i.d.
- problem : infer the law of X based on the sample
- model : the law of X is $\mathbb{P}(X|\theta)$ given
- bayesian choice θ is a random variable
- model : prior $\mathbb{P}(\theta)$ given
- bayesian choice - estimate the posterior :

$$\mathbb{P}(\theta|X_1, ...X_n) = \frac{\mathbb{P}(X_1, ...X_n|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X_1, ...X_n)}$$

Likelihood : $\mathbb{P}(X_1, ...X_n|\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i|\theta)$ given

$$\log \mathbb{P}(\theta|X_1, ...X_n) = \sum_{i=1}^{n} \log \mathbb{P}(X_i|\theta) + \log \mathbb{P}(\theta) - \log \mathbb{P}(X_1, ...X_n)$$

Convergence

X is a r.v. with $\mathbb{E}(X) = 0$ and $V(X) = 1$. $X_1, X_2, \ldots X_n$ is i.i.d.

- $\displaystyle\sum_{i=1}^{n} X_n \underset{n\to\infty}{\longrightarrow} \infty$
- $\displaystyle\frac{1}{n}\sum_{i=1}^{n} X_n \underset{n\to\infty}{\longrightarrow} 0$ LLN concentration
- $\displaystyle\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_n \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$ CLT speed
- $\displaystyle\sup_x \frac{1}{\sqrt{2n\log\log n}}\sum_{i=1}^{n} X_n \underset{n\to\infty}{\longrightarrow} 1$ LIL extreme events

Law of the large number, central limit theorem, Law of the iterated logarithm

what are you after ?

Matrices

- Mathematician, computer scientist, physicist
- linear mapping, tabular of real, set of linear equations
- singular, well defined
- singular values and eigen values