Advanced Machine Learning
Lecture 1 (a few notes)
Kernel methods and RKHS

F. d'Alché-Buc and E. Le Pennec

Fall 2015

- Learning set (x_i, y_i) with
  - x_i individual,
  - y_i value/label in $\mathcal{Y}$,

  are modeled as i.i.d. random sample of a pair (X, Y).
- Goal : predict the random variable Y from the random variable X through a function f.
- For sake of simplicity, we will assume that $X \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$ or $Y \in \mathbb{R}$.
- The first case is called classification, the second regression... but they are formally the same !

- To measure the quality of a prediction function f, one uses a loss function ℓ quantifying the error between y and $f(y)$.
- Examples :
  - Prediction loss : $\ell(y, f(x)) = \mathbf{1}_{y \neq f(x)}$
  - Quadratic loss : $\ell(y, f(x)) = |y - f(x)|^2$
- Probabilistic model is used to define an average error :

$$R(f) = \mathbb{E}\left[\ell(Y, f(X))\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\ell(Y, f(X))\right]\right]$$

- Examples :
  - Prediction loss : $\mathbb{E}\left[\ell(Y, f(X))\right] = \mathbb{P}\left\{Y \neq f(X)\right\}$
  - Quadratic loss : $\mathbb{E}\left[\ell(Y, f(X))\right] = \mathbb{E}\left[|Y - f(X)|^2\right]$

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Solve one of these equivalent problems : minimizing the empirical risk while controlling the complexity of the function. Let C be a constant to be chosen :

Pb1

$Min_f R_n(f)$ s.c $\Omega(f) \leq C$

Pb2

$Min_f \Omega(f)$ s.c $R_n(f) \leq C$

Pb3

$Min_f R_n(f) + \lambda\Omega(f)$

- Ω(f) : measures the complexity of a single function f

- Choose the space of functions to work in : RKHS induce by a kernel k
- $\Omega(f) = ||f||^2$ where the norm is the norm of the RKHS where we work.
- Apply kernel trick in a systematic way
- Use kernel methods for various problems (nonlinear decision, structured dataÃ

In the next two lectures, we visit the theory of Reproducing Kernel Hilbert Spaces (RKHS) and its use in Statistical learning. Most of the course will be on black board.

- ML 1
  - 25/09 - Advanced kernel methods 1 : Kernels and RKHS (F. A.)
  - 02/10 - Advanced kernel methods 2 : Kernels methods in RKHS (F. A.)
  - 09/10 - Advanced tree-based methods, including (bagging, boosting, exponential weights) (E. LP.)
  - 16/10 - Dimension reduction and feature design (E. LP.)
  - 23/10 - Penalization (E. LP.)
  - 06/11 - Advanced kernel methods 3 : gaussian processes for regression and classification (F.A.)
  - 13/11 - Advanced kernel methods 4 : large scale issues and applications (F.A.)
  - 20/11 - Exam 1

- ML 2
  - 27/11 - Deep learning 1 (N.L., Criteo)
  - 04/12 - Deep learning 2 (N.L., Criteo)
  - 11/12 - Deep learning 3 (NL/YO)
  - 18/12 - Collaborative filtering (E. LP.)
  - 08/01 - Ranking (E. LP.)
  - 15/01 - Semi supervised learning (F.A.)
  - 22/01 - Visualization (E. LP.)
  - 29/01 - Exam 2

- Exam 1 : 8 pts
- Exam 2 : 8 pts
- Challenge : 4 pts

To prepare the exams, you'll have a list of exercises to do.
You will get a correction but no mark for that. You are free to do these homeworks, they help you to prepare the exams.

Choose the space of functions to work in

In the next two lectures, we visit the theory of Reproducing Kernel Hilbert Spaces (RKHS) and its use in Statistical learning. Most of the course will be on black board.

- SVM and other kernel-based functions can be derived from the theory of Reproducing Kernel Hilbert Spaces (RKHS)
- RKHS theory is about working with the space of functions of the form $\sum_i \alpha_i k(\cdot, x_i)$ with a proper norm
- It offers a solid theoretical ground for **penalized regression** in such a space
- RKHS theory was introduced by Aronzajn in 1950.
- Wahba developed nonparametric estimation using RKHS theory and representer theorems in 1972

1. Define a PDS kernel : $k(\cdot, \cdot)$
2. Define a (unique) RKHS, $\mathcal{H}$ from k with an appropriate norm $|| \cdot ||_{\mathcal{H}}$
3. Define a loss functional with two terms : a local loss function ℓ and a penalty function Ω
4. Prove/use a representer theorem to get the form of the minimizer of this functional : $\sum_i \alpha_i h(\cdot, x_i)$
5. Solve the optimization problem with this minimizer

Problem in the primal space

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{under the constraints} \quad y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1 - \xi_i, i = 1, \ldots, n.$$

$$\xi_i \geq 0, i = 1, \ldots, n.$$

Problem in the dual space

$$\max_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{under the constraints} \quad 0 \leq \alpha_i \leq C \; i = 1, \ldots, n.$$

$$\sum_i \alpha_i y_i \; i = 1, \ldots, n.$$

with the definition :

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i$$

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i < \mathbf{x}, \mathbf{x}_i >$$

Definition

Let $\mathcal{X}$ be a non-emptyo : set. The function k :$\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$is Postive Definite Symmetric if (1) it is symmetric and (2), it satisfies one of the following condition : $\forall$m $\in \mathbb{N}^+$,{ $x_1, \ldots, \mathbf{x}_m\} \in \mathcal{X}$, and $\mathbf{c} \in \mathbb{R}^m$,
$\mathbf{c}^T K \mathbf{c} = \sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geq 0$

Nota Bene : a kernel can be seen as a similarity function on $\times\mathcal{X}$ with a special property (being SDP).

Definition (Reproducing Kernel Hilbert space - RKHS)

Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space if :

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (**reproducing property**).

In particular, for any x,y $\in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}$$

Theorem (Reproducing Kernel Hilbert space induced by a kernel)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite symmetric kernel. Then, there exists a Hilbert space $\mathcal{H}$ and a function $\varphi : \mathcal{X} \to \mathcal{H}$ such that :

$$\forall (x, x') \in \mathcal{X} \times \mathcal{X}, k(x, x') = < \varphi(x), \varphi(x') >_{\mathcal{H}}$$

Furthermore, $\mathcal{H}$ has the following reproducing property :

$$\forall f \in \mathcal{H}, \forall x \in \mathcal{X}, f(x) = < f(\cdot), k(\cdot, x) >$$

(blackboard)

Theorem

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite symmetric kernel and $\mathcal{H}_k$ be a Hilbert space built from k and $\mathcal{X}$, then $\mathcal{H}_k$ is unique.

Any Hilbert space $\mathcal{H}$ such that there exists $\varphi : \mathcal{X} \to \mathcal{H}$ with :
$\forall (x, x') \in \mathcal{X} \times \mathcal{X}, k(x, x') = < \varphi(x), \varphi(x') >$
is called a feature space associated with k and φ is called a feature map.

Theorem

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite symmetric kernel and $\mathcal{H}_k$, its corresponding RKHS, then, for any non-decreasing function $\Omega : \mathbb{R} \to \mathbb{R}$ and any loss function $\mathrm{L} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, any minimizer of :

$$J(f) = L(f(x_1), \ldots, f(x_n)) + \lambda \Omega(\|f\|_{\mathcal{H}}^2) \tag{1}$$

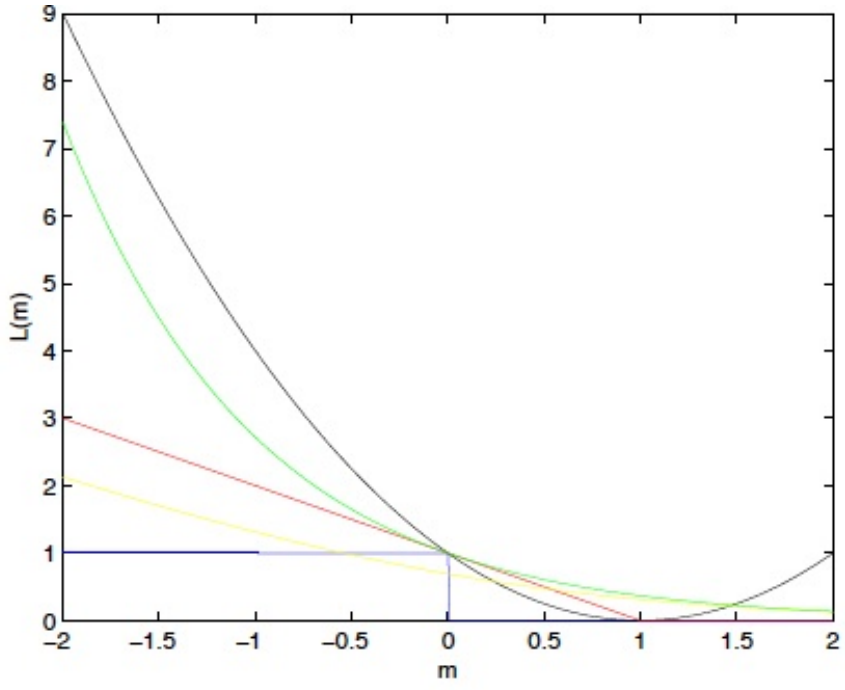
admits an expansion of the form :

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$

Moreover if Ω is strictly increasing, then any minimizer of 1 has exactly this form.

Let us define : $\mathcal{H}_1 = \text{span}\ \{k(x_i, \cdot), i = 1, \ldots, n\}$
Any $f \in \mathcal{H}$ writes as : $f = f_1 + f^{\perp}$, with $f_1 \in \mathcal{H}_1$ and $f^{\perp} \in \mathcal{H}_1^{\perp}$
where $\mathcal{H}=$ direct sum of$\mathcal{H}_1$ and $\mathcal{H}_1^{+}$.
By orthogonality, $\|f\|^2 = \|f_1\|^2 + \|f_1^{\perp}\|^2$
Hence, by property of Ω,
$\Omega(\|f\|^2) = \Omega(\|f_1\|^2) + \Omega(\|f_1^{\perp}\|^2) \geq \Omega(\|f_1\|^2)$
By the reproducing property, we get :
$f(x_i) =< f_1(\cdot) + f_1^{\perp}(\cdot), k(x_i, \cdot) >=< f_1(\cdot), k(x_i, \cdot) >= f_1(x_i)$
Hence, $L(f(x_1), \ldots, f(x_n)) = L(f_1(x_1), \ldots, f_1(x_n))$ and
$J(f_1) \leq J(f)$
To recap, if f is a minimizer of $J(f)$, then f_1 is also a minimizer of J. Moreover if Ω is stricly increasing, $J(f_1) < J(f)$, then any $f = f_1 + f_1^{\perp}$ exactly equals to f_1.

- $L(f(x_1), \ldots, f(x_n)) = \sum_i (y_i - h(x_i))^2$ and $\Omega(||f||) = ||f||^2$
  - Kernel ridge regression : $\hat{\alpha} = (K + \lambda Id)^{-1}\mathbf{y}$
- SVM without bias b
- $L(f(x_1), \ldots, f(x_n)) = max(0, 1 - y_i f(x_i))$ (hinge loss) and $\Omega(||f||) = ||f||^2$
  - If you want to introduce b, you need the refer to the semi-parametric representer theorem.


9
8
7
6
5
4
3
2
1
0
L(m)
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
m

- Use closure properties to build new kernels from existing ones
- Kernels can be defined for various objects :
  - **Structured objects** : (sets), graphs, trees, sequences, ...
  - Unstructured data with underlying structure : texts, images, documents, signal, biological objects (gene, mRNA,protein, ...)
- **Kernel learning** :
  - Hyperparameter learning : see Chapelle et al. 2002
  - Multiple Kernel Learning : given $k_1, \ldots, k_m$, learn a convex combination $\sum_i \beta_i k_i$ of kernels (see SimpleMKL Rakotomamonjy et al. 2008, unifying view in Kloft et al. 2010)

- Convolution kernels
- Fisher kernels
- Graph kernels
- Kernels between nodes in a graph

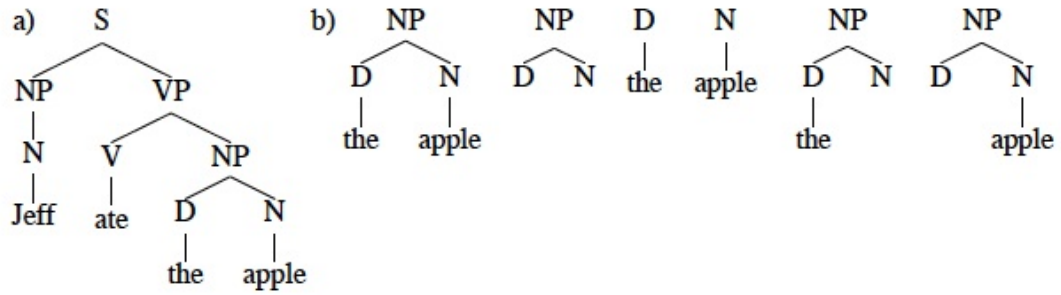
Definition :
Suppose that $x \in \mathcal{X}$ is a **composite structure** and $x_1, \ldots, x_D$ are its "parts" according a relation R such that $(R(x, x_1, x_2, \ldots, x_D)$ is true, with $x_d \in \mathcal{X}_d$ for each $1 \leq d \leq D$, D being a positive integer. k_d be a PDS kernel on a set $\mathcal{X} \times \mathcal{X}$, for all (x,x'), we define :

$$k_{conv}(x, x') = \sum_{(x_1,\ldots,x_d) \in R^{-1}(x), (x'_1,\ldots,x'_d) \in R^{-1}(x')} \prod_{d=1}^{D} k_d(x_d, x'_d)$$

$R^{-1}(x) =$ all decompositions $(x_1, \ldots, x_D)$ such that $(R(x, x_1, x_2, \ldots, x_D)$. k_{conv} is a PDS kernel as well. Intuitive kernel, used as a building principle for a lot of other kernels. Next, we will see two examples. **REF : Haussler, Convlolution kernels for discrete structure, UCSC tech report, 1999.**

Learning task :

- **Input** : sentence $\longrightarrow$ syntax tree
- **Output** : question class
- For instance, in economical news articles, classes are ORGANIZATION, LOCATION,



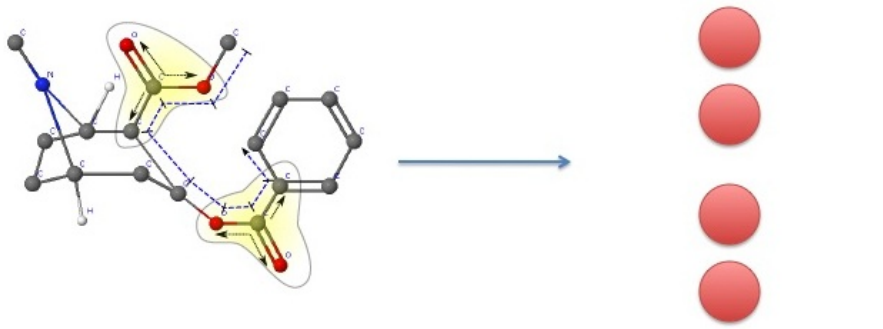
Let us first enumerate all tree fragments that occur in the training data. Let m be the size of this set. For a tree, define $v(T) = (v_1(T), \ldots, v_m(T))^T$ where $v_i(T)$ is the number of occurrences of the i^{th} subtree.

Definition :

$$k_{conv}(T, T') = k(v(T), v(T'))$$

NB : the kernel can be normalized. In NLP, k is often chosen as the linear kernel. Efficient implementations are available.
Sequences can be processed in the same way.
Ref : Collins and Duffy, 2001 ; Suzuki et al. 2003

Motivation : a regression problem from structured data



- **Inputs** : molecule (drug candidate)
- **Output** : activity on a cancer line

For a given length L, let us first enumerate all the paths of length $\ell \leq L$ in the training dataset (data are molecule = labeled graphs). Let m be the size of this (huge) set. For a graph, define $v(G) = (v_1(G), \ldots, v_m(T))^T$ where $v_i(T)$ is 1 if the i^{th} path appears in the labeled graph G, and 0 otherwise.

Definition 1 :

$$k_m(G, G') =< v(G), v(G') >$$

Tanimoto kernel

$$k_m^t(G, G') = \frac{k_m(G, G')}{k_m(G, G) + k_m(G', G') - k_m(G, G')}$$

idea : k_m^t calculates the ratio between the number of elements of the intersection of the two sets of paths (G and G' are seen as bags of paths) and the number of elements of the union of the two sets.
Refs : Ralaivola et al. 2005, Su et al. 2011

Let $x_1, \ldots, x_n$, n objects associated with a non oriented graph of size n and adjacency matrix W. Define the graph Laplacian : $L = D - W$, D is the diagonal matrix of degrees

$$K = \exp(-\lambda L)$$

We will see applications of this kernel in the unsupervised course.
Ref : Kondor and Lafferty, 2003

Combine the advantages of graphical models and discriminative methods

Let $\mathbf{x} \in \mathbb{R}^p$ be the input vector of a classifier.

- Learn a generative model $p_\theta(\mathbf{x})$ from unlabeled data $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- Define the Fisher vector as : $\mathbf{u}_\theta(\mathbf{x}) = \nabla_\theta \log p_\theta(\mathbf{x})$
- Estimate the Fisher Information matrix of p_θ : $F_\theta = \mathbb{E}_{\mathbf{x} \sim p_\theta}[\mathbf{u}_\theta(\mathbf{x})\mathbf{u}_\theta(\mathbf{x})^T]$
- **Definition** : $k_{Fisher}(\mathbf{x}, \mathbf{x}') = \mathbf{u}_\theta(\mathbf{x})^T F_\theta \mathbf{u}_\theta(\mathbf{x})$

Ref : Jaakola, Haussler, 1999 : Fisher kernel and HMM for protein modeling
Ref : structured kernels : Gartner, 2006 (a paper review)

Kernels

Exercices - Lecture 1

- Prove Theorem 3 ($\mathcal{F}$ is a RKHS off for all x, the evaluation functional is continuous.
- Prove unicity of a RKHS induced by a PSD kernel k
- Find some closure properties of the family of PSD kernels and prove them
- Practise in scikitlearn : KPCA, Kernel Ridge, SVR

- A maths course on RKHS by V. Paulsen (`http://www.math.uh.edu/~vern/rkhs.pdf`)
- Spline Models for observational data by G. Wahba, SIAM, 1990.
- A tutorial review of RKHS methods in Machine Learning, Hofman , Schoelkopf, Smola, 2005 (`Ahttps://www.researchgate.net/publication/228827159_A_Tutorial_Review_of_RKHS_Methods_in_Machine_Learning`)
- book : Foundations of Machine Learning, Mohri, Rostamizadeh, Talwalkar, MIT Press, 2012