Machine Learning: from Theory to Practice
Lecture 8
Kernels and margin bounds
Review of supervised learning methods

F. d'Alché-Buc and E. Le Pennec
email: florence.dalche@telecom-paristech.fr

Fall 2014

Outline

1. Positive Definite Kernels : applications

2. Margin Theory

Positive Definite Kernels : applications

Kernel design

- 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 $k_1, \dots, k_m$, 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)

Positive Definite Kernels : applications

Examples of kernels for non-vectorial data and applications

- Convolution kernels
- Fisher kernels
- Graph kernels
- Sequence kernels

Positive Definite Kernels : applications

Convolution kernels

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.

Positive Definite Kernels : applications

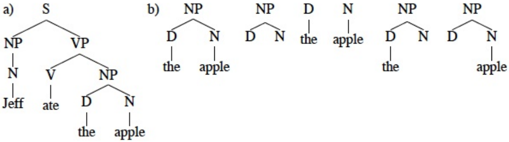
Kernels for Natural Language Processing

Learning task :

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

Positive Definite Kernels : applications

Kernel for Natural Language Processing



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 $h(T) = (h_1(T), \dots, h_m(T))^T$ where $h_i(T)$ is the number of occurrences of the i^{th} subtree.

Positive Definite Kernels : applications

Convolution Kernel for Natural Language Processing

Definition :

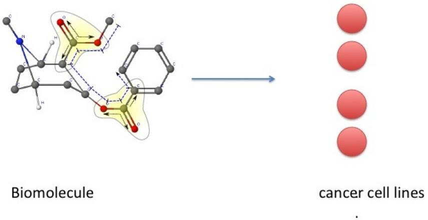
$$k_{conv}(T, T') = k(h(T), h(T'))$$

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.
References : Collins and Duffy, 2001 ; Suzuki et al. 2003

Positive Definite Kernels : applications

Kernel for labeled graphs

Motivation : a regression problem from structured data



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

Positive Definite Kernels : applications

Kernel for labeled graphs

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 $h(G) = (h_1(G), \ldots, h_m(T))^T$ where $h_i(T)$ is 1 if the i^{th} path appears in the labeled graph G, and 0 otherwise.

Positive Definite Kernels : applications

Kernel for labeled graphs

Definition 1 :

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

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.
Reference : Ralaivola et al. 2005, Su et al. 2011

Positive Definite Kernels : applications

Kernel between vertices in a graph

Let $x_1, \dots, x_n$, 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.
Reference : Kondor and Lafferty, 2003

Positive Definite Kernels : applications

Fisher kernel

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** : $k_{Fisher}(\mathbf{x}, \mathbf{x}') = \mathbf{u}_\theta(\mathbf{x})^T F_\theta \mathbf{u}_\theta(\mathbf{x})$

Applications

Classification of secondary structure of proteins, topic modeling in documents, image classification and object recognition, audio signal classification . . .

Positive Definite Kernels : applications

Multiclass classification with kernels

- Strategy (1) : one class against the other → build C classifiers, if C classes.
- Strategy (2) : a classifier for each pair of classes → take the largest vote
- Strategy (3) : multiple margin separators
- Strategy (4) : use vector-valued function and matrix-valued kernels (See semester 2)

Margin Theory

Outline

1. Positive Definite Kernels : applications

2. Margin Theory

Margin Theory

Generalization bounds

Question : learning guarantee

If we measure the empirical risk $R_S(h)$ associated to a classifier h, what can we say about its true risk $R(h)$?

Definition : **Shattering**
$\mathcal{H}$ is said to shatter a set of data points $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ if, for all the 2^n possible assignments of binary labels to those points, there exists a function $h \in \mathcal{H}$ such that the model h makes no errors when predicting that set of data points.

Definition : **VC-dimension**
The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$:

$$VCdim(\mathcal{H}) = max\{m : \exists(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathcal{X}^m \text{that are shattered by } \mathcal{H}\}$$

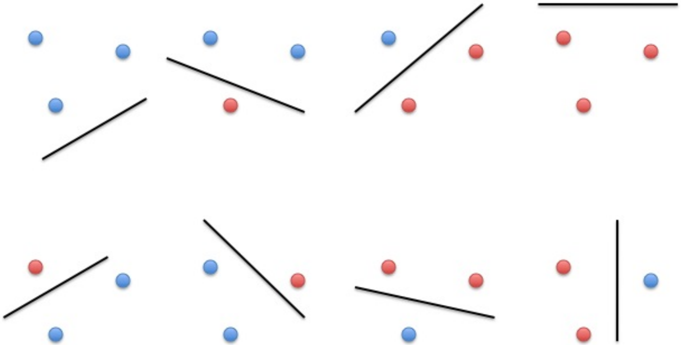
N.B. : if VCdim($\mathcal{H}$) = d, then there exists a set of d points that is fully shattered by $\mathcal{H}$, but this DOES NOT imply that all sets of dimension d or less are fully shattered !

What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?
Obviously VCdim($\mathcal{H}_2$) ≥ 2
Let us try with 3 points :

Margin Theory

VC-dimension of Hyperplanes

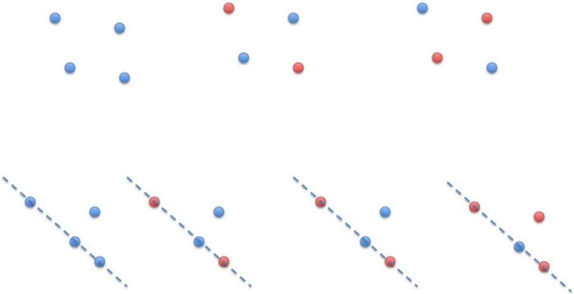
What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?
Let us consider the following triplet of points



Margin Theory

VC-dimension of Hyperplanes

What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ? For any set of 4 points, either 3 of them (at least) are aligned or no triplet of points is aligned.



We can show that it is not possible for $\mathcal{H}_2$ to shatter 4 points. Then $\text{VCdim}(\mathcal{H}_2) = 3$.

Margin Theory

VC-dimension of Hyperplanes

More generally, one can prove :

$$VCdim(\mathcal{H}_d) = d + 1$$

Margin Theory

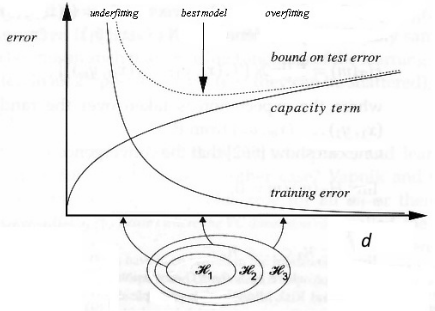
VC-dimension generalization bounds

Theorem **:**
Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, over a random sampling $\mathcal{S} \sim D^n$, the following holds for all $h \in \mathcal{H}$:

$$R_D(h) \leq R_{\mathcal{S}}(h) + \sqrt{\frac{2d \log(\frac{em}{d})}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$$

Margin Theory

Principle of Structural Risk Minimization



Vapnik proposed to replace empirical minimization principle by structural risk minimization, the underlying idea is to control the complexity of family $\mathcal{H}$ while reducing the empirical error.

Margin Theory

Application to optimal margin hyperplane : VC-dimension of canonical hyperplanes

Theorem : **VC-dimension**
Let $\mathcal{S} \subseteq \{\mathbf{x} : \|\mathbf{x}\| \leq r\}$. Then, the VC-dimension d of the set of canonical hyperplanes
$\{x \to \text{sgn}(\mathbf{w}^T\mathbf{x}) : min_{\mathbf{x} \in \mathcal{S}} |\mathbf{w}^T\mathbf{x}| = 1 \wedge \|\mathbf{w}\| \leq M\}$ verifies :

$$d \leq r^2 M^2.$$

N.B. : hard margin case.

Margin Theory
Proof

Assuming that d is the VC-dimension then there exists $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$ a set fully shattered by the canonical hyperplanes. Then, for all $\mathbf{y} = (y_1, \dots, y_d) \in \{-1, 1\}^d$, there exists a $\mathbf{w}$ such that :
$\forall i \in [1, d], 1 \leq y_i(w^T x_i)$ Summing up :

$$d \leq \mathbf{w}^T \sum_{i=1}^{d} y_i \mathbf{x}_i \leq \|w\| \|\sum_{i=1}^{d} y_i \mathbf{x}_i\| \leq M \|\sum_{i=1}^{d} y_i \mathbf{x}_i\|.$$

Because this is true for all $y_1 \dots, y_d$, it also works for the expectation taken over $\mathbf{y} = (y_1, \dots, y_d)$, the $y_i's$ i.i.d. from a uniform distribution

$$d \leq M \mathbb{E}_{\mathbf{y}}[\|\sum_{i=1}^{d} y_i \mathbf{x}_i\|]$$

Using Jensen's inequality : $d \leq M \mathbb{E}_{\mathbf{y}}[\| \sum_{i=1}^{d} y_i \mathbf{x}_i \|^2]]^{\frac{1}{2}}$

Margin Theory
Proof ctd'

By linearity of expectation :

$$\mathbb{E}_{\mathbf{y}}[\| \sum_{i=1}^{d} y_i \mathbf{x}_i \|^2]^{\frac{1}{2}} = (\mathbb{E}_{\mathbf{y}}[(\sum_{i,j=1}^{d} y_i \mathbf{x}_i)^T (\sum_{j=1}^{d} y_j \mathbf{x}_j)])^{\frac{1}{2}} = [\sum_{i,j=1}^{d} \mathbb{E}_{\mathbf{y}}[y_i y_j](\mathbf{x}_i^T \mathbf{x}_j)]^{\frac{1}{2}}$$

By independence property of $y_1, \ldots, y_d$ and because the distribution is uniform, we have :

$$\begin{aligned} \mathbb{E}_{\mathbf{y}}[\| \sum_{i=1}^{d} y_i x_i \|^2]]^{\frac{1}{2}} &= [\sum_{i \neq j}^{d} \mathbb{E}_{\mathbf{y}}[y_i] \mathbb{E}_{\mathbf{y}}[y_j](\mathbf{x}_i^T \mathbf{x}_j) + \sum_{i=1}^{d} E[y_i^2](\mathbf{x}_i^T \mathbf{x}_i)]^{\frac{1}{2}} \\ &= \sum_{i=1}^{d} (\mathbf{x}_i^T \mathbf{x}_i)]^{\frac{1}{2}} \leq [dr^2]^{\frac{1}{2}} = r\sqrt{d} \end{aligned}$$

Eventually, we have : $d \leq Mr\sqrt{d}$ Therefore $\sqrt{d} \leq rM$.

Margin Theory

Generalization bounds and Optimal Margin Hyperplane

Theorem **:**
For any $\delta > 0$, with probability at least $1 - \delta$, over a random sampling $\mathcal{S} \sim D^n$, and for canonical hyperplane h defined using $\mathcal{S}$, the following holds :

$$R_D(h) \leq R_{\mathcal{S}}(h) + \sqrt{\frac{2r^2M^2\log(\frac{en}{r^2M^2})}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$$

(with d replaced by the previous bound).

From the previous result on VC-dimension of canonical hyperplanes, we have :
controlling the norm **w** allows to control the VC-dimension of canonical hyperplanes and therefore reduces the second term of the bound. OMH (SVM) has been invented to implement SRM principle.

Margin Theory

Tightness of the generalization bounds

More tight bounds with Rademacher complexity (Koltchinskii, 1999)

Margin Theory
Empirical Rademacher complexity

Definition **:**
Let $\mathcal{G}$ a family of functions mapping a set $\mathcal{Z}$ to $[a, b]$ and $\mathcal{S} = \{z_1, \dots, z_n\}$ a fixed sample of size n with elements in $\mathcal{Z}$. Then, the *empirical Rademacher complexity* of $\mathcal{G}$ with respect to the sample $\mathcal{S}$ is defined as :

$$\begin{aligned} R_{\mathcal{S}}^{Rad}(\mathcal{G}) &= \mathbb{E}_\sigma[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)] \\ &= \mathbb{E}_\sigma[\sup_{g \in \mathcal{G}} \frac{< \sigma, g_{\mathcal{S}} >}{n}] \end{aligned}$$

where $\sigma = (\sigma_1, \dots, \sigma_n)^T$, with σ_i's independent uniform variables taking values in $\{-1, +1\}$ and $g_{\mathcal{S}} = (g(z_1), \dots, g(z_n))^T$. The random variables σ_i are called Rademacher variables.

Margin Theory

Empirical Rademacher complexity

$$R_{\mathcal{S}}^{Rad}(\mathcal{G}) = \mathbb{E}_{\sigma}[\sup_{g \in \mathcal{G}} \frac{< \sigma, g_{\mathcal{S}} >}{n}]$$

The empirical Rademacher complexity captures the richness of a family of functions by measuring the degree to which a set of hypotheses fit random noise. The complexity is data-dependent.

Margin Theory

Rademacher complexity

Definition **:**

Let D denote the distribution according to which samples are drawn. For any $n \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size n drawn according to D :

$$R_n^{Rad}(\mathcal{G}) = \mathbb{E}_{\mathcal{S} \sim D^n}[R_{\mathcal{S}}^{Rad}(\mathcal{G})]$$

Margin Theory

Rademacher complexity of kernel-based hypotheses

Theorem **:**
Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a PDS kernel and let $\Phi : \mathcal{X} \rightarrow \mathbb{R}$ be a feature map associated to k. Let $\mathcal{S} \subseteq \{x : k(x, x) \leq r^2\}$ be a sample of size n, and $\mathcal{F} = \{x \rightarrow \langle \mathbf{w}, \Phi(x) \rangle, ||\mathbf{w}||_{\mathcal{H}} \leq M\}$ for some $m \geq 0$. Then,

$$R_{\mathcal{S}}^{Rad}(\mathcal{F}) \leq \frac{M\sqrt{Tr(K)}}{n} \leq \sqrt{\frac{r^2 M^2}{n}}$$

N.B. : For Gaussian kernel, we have exactly : $Tr(K) = n$

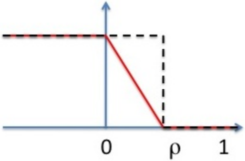
Definition : **Margin loss function**

$$\begin{aligned}\varphi_{\rho(x)} &= min(1, max(0, 1 - \frac{x}{\rho})) \\ &= min(1, [\frac{x}{\rho}]_+)\end{aligned}$$

Definition : **Empirical margin loss**

For any real-valued function f

$$R_{\mathcal{S}}^{\rho}(f) = \frac{1}{n}\sum_{i=1}^{n}\varphi_{\rho}(y_i f(\mathbf{x}_i)))$$



Margin Theory

Margin bounds for binary classification

Theorem **:**
Let $\mathcal{F}$ be a set of real-valued functions. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least (1 -δ), each of the following holds for all $f \in \mathcal{F}$:

$$R(f) \leq R^{\rho}_{\mathcal{S}}(f) + \frac{2}{\rho} R^{Rad}_{n}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$R(f) \leq R^{\rho}_{\mathcal{S}}(f) + \frac{2}{\rho} R^{Rad}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Reference : Cortès et al. ICML 2010

Theorem **:**
Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel with $\sup_{x \in \mathcal{X}} k(x,x) = r^2$ and $\Phi : \mathcal{X} \to \mathbb{R}$ be a feature map associated to k. Let $\mathcal{F} = \{x \to \langle \mathbf{w}, \Phi(x) \rangle, ||\mathbf{w}||_{\mathcal{F}} \leq M\}$. Fix $\rho > 0$. Then, for any $\delta > 0$, we have :

$$R(f) \leq R_{\mathcal{S}}^{\rho}(f) + 2\sqrt{\frac{\frac{r^2 M^2}{\rho^2}}{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Good generalization if $\frac{\rho}{r}$ is large and the empirical margin loss is small.

Reference : Cortès et al. ICML 2010

Margin Theory
References

- Foundations of Machine Learning, Morhi, Rostamizadeh, Talwalkar, MIT Press, 2012.
- A tutorial review in RKHS methods in Machine Learning, Hofman, Schoelpkof, Smola. (online pdf)
- Papers
  - Convolution kernels on discrete structure, D. Haussler, UCSC-CRL-99 (public technical report)
  - Convolution kernels for natural language, Collins and Duffy, NIPS 2001
  - Graph Kernels for chemical informatics, Ralaivola et al. 2005.Preprint Elsevier.
  - Structured output prediction of anti-cancer drug activity, Su et al. 2010, PRB 2010, online pdf.
  - Improving Fisher kernel for large scale image Classification, Perronin et al. 2010, PSM2010,online pdf.
  - Rademacher and gaussian complexities : risk bounds and structural results, Bartlett and Mendelson, JMLR 3(2002), online