Tribulations of a similarity man in Kerneland

Lluís A. Belanche
belanche@lsi.upc.edu

Soft Computing Research Group

Dept. de Llenguatges i Sistemes Informàtics (Software department)

Universitat Politècnica de Catalunya

SESAAME Talk, November 2010

Tribulations of a similarity man in Kerneland

- The Support Vector Machine (SVM) was developed by Vapnik and his coworkers, initially for classification problems and has won great popularity as a tool for non-linear system identification

- A key idea in kernel machines is that of the *kernel*, but ...

- The SVM formulation does not include criteria to select a kernel function:

  - hand-design it, inspiration, trial-and-error ...

  - learn it (as in Multiple Kernel Learning)

Tribulations of a similarity man in Kerneland

A standard result for identifying kernels can be derived from Mercer's result:

Theorem 1 *A continuous and symmetric function* $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ *is a kernel if it satisfies the condition:*

$$\int_{\mathcal{H}\times\mathcal{H}} K(x,y)g(x)g(y)dxdy \geq 0$$

for any function g *such that* $\int_{\mathcal{H}}(g(x))^2dx < \infty$

- The function K can then be expressed as an absolutely and uniformly convergent series (finite or infinite), almost everywhere, in terms of eigenfunctions and (positive) eigenvalues:

$$K(x,y) = \sum_{j=1}^{F} \lambda_j \psi_j(x)\psi_j(y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$$

  where $F \in \mathbb{N} \cup \{\infty\}$.

Tribulations of a similarity man in Kerneland

- Except for specific cases, it may not be easy to check whether this condition is satisfied.

Another, equivalent, definition:

Theorem 2 *The function $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a kernel if for any finite subset $\{x_1, x_2, ..., x_n\} \in \mathcal{H}$ the associated kernel matrix $K_{n\times n} = (k_{ij})$, where $k_{ij} = K(x_i, x_j)$ is a symmetric positive semidefinite (PSD) matrix.*

Recall a real symmetric matrix $K_{n\times n}$ is PSD when for any $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i k_{ij} c_j \geq 0.$$

Tribulations of a similarity man in Kerneland

- To create a kernel function we find a space $\mathcal{H}$, a ψ mapping onto that space and define the kernel to be an inner-product in $\mathcal{H}$ using ψ,

- Many kernels have been developed in this way, for example the *String Kernel*

- One can take another approach: find a similarity measure that suits the application needs, and see if it is PSD (or can be "forced" to be).

Tribulations of a similarity man in Kerneland

- Similarities and kernels are two-place symmetric functions ...
- Are all kernels similarities? No (boundedness, transitivity, ...)
- Are all similarities kernels? No (PSD)
- Can similarities and kernels be identified in some contexts? Yes (Schölkopf's sheeps, many authors)

Tribulations of a similarity man in Kerneland

Do you know of a kernel that is a similarity?

$$k(x, y) = exp(-\frac{\|x - y\|^2}{\sigma^2})$$

Yes, the RBF kernel

Tribulations of a similarity man in Kerneland

What are the common problems?

- Handling non-standard data types
  - Categorical variables
  - Fuzzy variables
  - Set variables
  - Bitstring variables, ...
- Missing values
- Aggregation

Tribulations of a similarity man in Kerneland

Categorical variables

The basic similarity measure for these variables is the *overlap*:

$$k(x, y) = \mathbb{I}_{\{x=y\}} = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

The overlap kernel!

Proof. For any $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i=1}^{n} c_i^2 \geq 0.$$

Tribulations of a similarity man in Kerneland

Categorical variables

In binary $(+,-)$ variables the values are considered to *match* only if the state is present on both elements being compared $(++)$.

$$k(x,y) = \frac{a}{n-d},$$

where a (d) is the number of $++$ $(--)$ matches

Given two bitstrings of length n, we define A (resp., D) as $1/n$ times the overlap matrix for the $++$ (resp., $--$) matches

Both A, D are PSD, and $S = \frac{A}{1-D}$

Since $\lim_{n\to\infty}(1 + q + q^2 + \ldots + q^n) = \frac{1}{1-q}$, for $-1 < q < 1$,

we can re-write

$$S = \lim_{n\to\infty}[A + A * (D + D^2 + \ldots + D^n)].$$

Tribulations of a similarity man in Kerneland

Missing values

Missing information ...

- ... an old issue in statistical analysis
- Very common in Medicine and Engineering
- There are two basic ways of dealing with missing data:

  1. by *completing* the data description in a (hopefully) optimal way, or
  2. by *extending* the methods to work with incomplete descriptions.

Tribulations of a similarity man in Kerneland

Missing values

We present an approach that allows the *extension* of (almost) any kernel to one that is defined even in the presence of missing values.

Theorem 3 *Let K be a kernel in a set $\mathcal{H}$ (e.g. a similarity function) and p a probability density function in $\mathcal{H}$. Then the function*

$$\widehat{K}(x,y) = \begin{cases} K(x,y), & \text{if } x,y \neq \mathcal{X} \text{ ;} \\ \int_{\mathcal{H}} p(y)K(x,y)dy, & \text{if } x \neq \mathcal{X} \text{ and } y = \mathcal{X}; \\ \int_{\mathcal{H}} p(x)K(x,y)dx, & \text{if } x = \mathcal{X} \text{ and } y \neq \mathcal{X}; \\ \int_{\mathcal{H}} p(x) \int_{\mathcal{H}} p(y)K(x,y)dydx, & \text{if } x,y = \mathcal{X} \end{cases}$$

is a kernel in $\mathcal{H} \cup \{\mathcal{X}\}$.

Tribulations of a similarity man in Kerneland

Missing values

Theorem 4 *Let K be a kernel in $\mathcal{H}$ (e.g. a similarity function) and P a probability mass function in $\mathcal{H}$. Then the function*

$$\widehat{K}(x,y) = \begin{cases} K(x,y), & \text{if } x,y \neq \mathcal{X} \text{ ;} \\ \sum\limits_{y \in \mathcal{H}} P(y)K(x,y), & \text{if } x \neq \mathcal{X} \text{ and } y = \mathcal{X}; \\ \sum\limits_{x \in \mathcal{H}} P(x)K(x,y), & \text{if } x = \mathcal{X} \text{ and } y \neq \mathcal{X}; \\ \sum\limits_{x \in \mathcal{H}} P(x) \sum\limits_{y \in \mathcal{H}} P(y)K(x,y), & \text{if } x,y = \mathcal{X} \end{cases}$$

is a kernel in $\mathcal{H} \cup \{\mathcal{X}\}$.

Tribulations of a similarity man in Kerneland

Missing values

An example of Theorem 4. Consider a categoric feature that takes values in the finite set $\mathcal{V} = \{v_1, ..., v_l\}$.

The probability mass function f can be *estimated* in the usual way from the data set by the frequency of every modality among the values that are non-missing for this feature.

Then, for all $v_i, v_j \in \mathcal{V}$,

$$K(v_i, v_j) = \begin{cases} \mathbb{I}_{\{v_i = v_j\}}, & \text{if } v_i, v_j \neq \mathcal{X} \ ; \\ g(v_i), & \text{if } v_i \neq \mathcal{X} \text{ and } v_j = \mathcal{X}; \\ g(v_j), & \text{if } v_i = \mathcal{X} \text{ and } v_j \neq \mathcal{X}; \\ G, & \text{if } v_i = v_j = \mathcal{X} \text{ and } i \neq j; \\ 1 & \text{if } v_i = v_j = \mathcal{X} \text{ and } i = j \end{cases}$$

where $g(z) = \sum_{i=1}^{l} f(v_i)\mathbb{I}_{\{v_i = z\}} = f(z)$ and $G = \sum_{i=1}^{l} f(v_i)^2$, is a PSD kernel in $\mathcal{V} \cup \{\mathcal{X}\}$.

Tribulations of a similarity man in Kerneland

Aggregation

How to create a full kernel in $\mathcal{H} = \mathcal{H}_1 \times ... \times \mathcal{H}_t$ from a collection of *partial* kernels K_i defined in the sets $\{\mathcal{H}_i\}_{i=1\div t}$?

Theorem 5 *If $\{K_i\}_{i=1\div t}$ are kernels defined in the sets $\mathcal{H}_i$, the function:*

$$\mathcal{K}(x, y) = \frac{1}{t} \sum_{i=1}^{t} K_i(x_i, y_i) \tag{1}$$

is a kernel in the product space $\mathcal{H}$.

Proof. The sum of $t > 0$ PSD matrices is a PSD matrix; take any real $r > 0$ and a PSD matrix A, then rA is again PSD (in the present case, $r = 1/t$).

Tribulations of a similarity man in Kerneland

Aggregation

Theorem 6 *If* $\{K_i\}_{i=1\div t}$ *are kernels defined in the sets* $\mathcal{H}_i$*, the function:*

$$\mathcal{K}(x,y) = \frac{\sum_{i=1}^{t} \alpha_i K_i(x_i, y_i)}{\sum_{i=1}^{t} \alpha_i} \tag{2}$$

is a kernel in the product space $\mathcal{H}$ *if all* $\alpha_i \geq 0$ *(and at least one* $\alpha_i > 0$*).*

Proof. Call m the number of non-zero α_i. The sum of $0 < m < t$ PSD matrices is a PSD matrix; take any real $r > 0$ and a PSD matrix A, then rA is again PSD (in the present case, $r = \left(\sum_{i=1}^{t} \alpha_i\right)^{-1}$).

Tribulations of a similarity man in Kerneland

Aggregation

Question 1 *If* $\{K_i\}_{i=1 \div t}$ *are kernels defined in the sets* $\mathcal{H}_i$*, is the function:*

$$\mathcal{K}(x, y) = \frac{t}{\sum_{i=1}^{t} \frac{1}{K_i(x_i, y_i)}} \tag{3}$$

a kernel in the product space $\mathcal{H}$ *?*

My initial conjecture: yes. My second-thoughts answer: no.

Tribulations of a similarity man in Kerneland

Aggregation

Question 2 *If* $\{K_i\}_{i=1\div t}$ *are kernels defined in the sets* $\mathcal{H}_i$*, is the function:*

$$\mathcal{K}(x,y) = \left(\prod_{i=1}^{t} K_i(x_i, y_i)\right)^{\frac{1}{t}} \tag{4}$$

a kernel in the product space $\mathcal{H}$ *?*

My initial conjecture: yes. My second-thoughts answer: no.

Tribulations of a similarity man in Kerneland

Aggregation

Can we give a general answer for general means of the form:

$$\mathcal{K}(x,y) = \left(\frac{1}{t} \sum_{i=1}^{t} (K_i(x_i, y_i))^q \right)^{\frac{1}{q}}, \ t \in \mathbb{N}, q \in \mathbb{R} \tag{5}$$

My initial conjecture: no. My second-thoughts answer: yes.

Tribulations of a similarity man in Kerneland

Conclusions

- Now you understand the tribulations ...
- Moving force is to design kernels that are much more suited to a specific application than standard off-the-shelf kernels
- Not to mention kernels on discrete structures (Haussler's work)
- More efficient fuel for a high-performance engine