Binet-Cauchy Kernerls on Dynamical Systems and its Application to the Analysis of Dynamic Scenes

Vishwanathan S.V.N.
Smola Alexander J.
Vidal René

presented by
Tron Roberto

July 31, 2006

Support Vector Machines (SVMs)

Classification problem

Given m observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$
find $f : \mathcal{X} \to \mathcal{Y}$

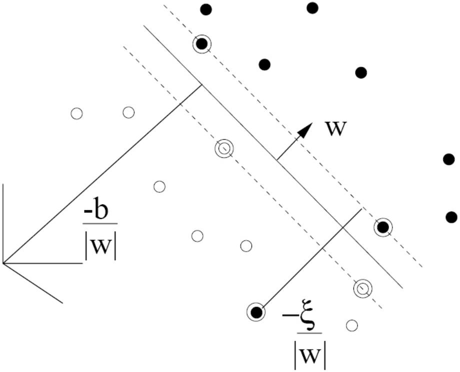
Support Vector Machines use

$$f = \text{sign}(\langle w, x \rangle + b) \tag{1}$$

The hyperplane $H(w, b) = \{x | \langle w, x, + \rangle b = 0\}$ splits $\mathcal{X}$ in two half spaces and is given by

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{subject to } & y_i(\langle w, x \rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & \forall i = 1, \ldots, m \end{aligned} \tag{2}$$

$H(w, b)$ depends only on some points (called **support vectors**)
$\Rightarrow$ efficient computations in training



Kernel methods

SVMs use the hypothesis of linearly separable clusters. What to do if the surface of separation is not linear?

Idea: use $\Phi : \mathcal{X} \to \mathcal{H}$ to map the points x_i into a higher (possible infinite) dimensional **feature space** where $\Phi(x_i)$ are linearly separable

f and $\|w\|^2$ can be expressed in term of the dot product $\langle \Phi(x), \Phi(x') \rangle$

$\Rightarrow$ avoid explicite computation of the mapping $\Phi(x)$

$\Rightarrow$ Kernel Trick: the classification can be done by replacing $\langle \Phi(x), \Phi(x') \rangle$ with the **kernel function** $k(x, x')$

Compound matrices

Let $A \in \mathbb{R}^{m \times n}$. Given $q \leq \min(m, n)$, define $I_q^n = \{\mathbf{i} = (i_1, i_2, \ldots, i_q) : 1 \leq i_1 \leq \ldots \leq i_q, i_i \in \mathcal{N}\}$ and likewise I_q^m.

The **compound matrix** of order q, $C_q(A)$ is defined as

$$[C_q(A)]_{\mathbf{i},\mathbf{j}} = det(A(i_k, j_l)) \text{ where } \mathbf{i} \in I_q^n \text{ and } \mathbf{j} \in I_q^m \tag{3}$$

The size of $C_q(A)$ is $\binom{n}{q} \times \binom{m}{q}$.

Two particular cases:

- if $q = m = n$, $C_q(A) = det(A)$
- if $q = 1$, $C_q(A) = A$

Binet-Cauchy kernels

Theorem (Binet-Cauchy)

Let $A \in \mathbb{R}^{l \times m}$ *and* $B \in \mathbb{R}^{l \times n}$. *For all* $1 \leq q \leq \min(m, n, l)$ *we have* $C_q(A^\top B) = C_q(A)^\top C_q(B)$

Theorem (Binet-Cauchy kernels)

Let $A, B \in \mathbb{R}^{n \times k}$. *For all* $1 \leq q \leq \min(n, k)$ *the two kernels*

$$k(A, B) = \mathrm{tr}(C_q(A^\top B)) \tag{4}$$

$$k(A, B) = \det(C_q(A^\top B)) \tag{5}$$

are well defined and positive semi-definite

The state of a dynamical system is defined by some $x \in \mathcal{H}$ where $\mathcal{H}$ is a RKHS with kernel $\kappa(\cdot, \cdot)$. The evolution of a linear dynamical system is denoted as:

$$x_{\mathbf{A}}(t) = \mathbf{A}(x, t) \text{ for } \; t \in \mathcal{T} \tag{6}$$

$\mathbf{A} : \mathcal{H} \times \mathcal{T} \rightarrow \mathcal{H}$ is a set of linear operators indexed by $\mathcal{T}$, the set of times of measurement.
The pair $(x, \mathbf{A})$ identifies the **trajectory** of a dynamical system, defined as the map:

$$\mathbf{Traj_A} : \mathcal{H} \rightarrow \mathcal{H}^{\mathcal{T}}, \; \mathbf{Traj_A}(x)(t) = x_{\mathbf{A}}(t) \tag{7}$$

Dynamical systems and trajectories
Comparison approaches

Two approaches:

- **Compare parameters:** useful when suitable parametrization exist. But systems with different parameters can behave almost identically.

Example

The two systems

$$x \leftarrow a(x) = |x|^p \text{ and } \; x \leftarrow b(x) = \min(|x|^p, |x|) \tag{8}$$

give identical trajectories as long as $p > 1$ and $|x| < 1$

- **Compare trajectories:** independent from parametrization, can take in account initial conditions. Requires a similarity measure.

Trace kernels on dynamical systems

Applying the trace kernel on trajectories involve the computation of

$$\operatorname{tr}\left(\mathbf{Traj}_{\mathbf{A}}(x)^{\top}\mathbf{Traj}_{\mathbf{A}'}(x')\right) = \sum_{t\in\mathcal{T}} \kappa(x_{\mathbf{A}}(t), x'_{\mathbf{A}'}(t)) \tag{9}$$

To assure convergence, a popular solution is to adopt discounting schemes:

$$\mu(t) = ce^{-\lambda t} \tag{10}$$

$$\mu(t) = \delta_{\tau}(t) \tag{11}$$

Definition (Trace kernel for trajectories)

The trace kernel on trajectories is defined as

$$k((x,\mathbf{A})(x',\mathbf{A}')) = \mathrm{E}_{t\sim\mu(t)}\left[\kappa(x_{\mathbf{A}}(t), x'_{\mathbf{A}'}(t))\right] \tag{12}$$

Trace kernels on dynamical systems

This kernel can be further specialized:

- on model parameters

$$k(\mathbf{A}, \mathbf{A}') = \mathrm{E}_{x,x'}[k((x, \mathbf{A})(x', \mathbf{A}'))] \tag{13}$$

- on initial conditions

$$k(x, x') = \mathrm{E}_{\mathbf{A},\mathbf{A}'}[k((x, \mathbf{A})(x', \mathbf{A}'))] \tag{14}$$

Determinant kernel on dynamical systems

Definition (Determinant kernel for trajectories)

Like the previous case, the **determinant kernel** is defined as

$$k((x, \mathbf{A})(x', \mathbf{A}')) = \mathrm{E}_{t\sim\mu(t)}\left[det(\mathbf{Traj_A}(x)^\top \mathbf{Traj_{A'}}(x'))\right] \quad (15)$$

For the determinant to exist, $\mu(t)$ must have finite support. Then, the kernel reduce to

$$k((x, \mathbf{A})(x', \mathbf{A}')) = det(\mathrm{K}) \quad (16)$$

where $\mathrm{K}_{i,j} = \kappa(x_{\mathbf{A}}(t_i), x'_{\mathbf{A}'}(t_j))$

This kernel, following an information-theoretic interpretation, measures the statistical dependence between two sequences

Linear dynamical systems - Discrete time

ARMA model

$$y_t = Cx_t + w_t \qquad \text{where} \quad w_t \sim \mathcal{N}(0, R), y \in \mathbb{R}^m \tag{17}$$

$$x_{t+1} = Ax_t + v_t \qquad \text{where} \quad v_t \sim \mathcal{N}(0, Q), x \in \mathbb{R}^n \tag{18}$$

In this case, closed form solution can be found for both trace and determinant kernels

Trace kernel on linear dynamical systems

With the assumption $\mu(t) = e^{-\lambda t}$, $\kappa(y_t, y'_t) = y_t^\top W y'_t$, W positive semi-definite, the trace kernel can be written

- with same realization of noise for both systems

$$k((x_0, A, C), (x'_0, A', C')) = x_0^\top M x'_0 + \frac{1}{1 - e^{-\lambda}} \operatorname{tr}(QM + WR) \tag{19}$$

- with different noise realizations

$$k((x_0, A, C), (x'_0, A', C')) = x_0^\top M x'_0 \tag{20}$$

where M satisfies the Sylvester's equation

$$M = e^{-\lambda} A^\top M A' + C^\top W C' \tag{21}$$

Other closed forms exist when independence from initial condition is desired and when the model is fully observable ($C = \mathbf{I}, R = 0$)

Determinant kernel on linear dynamical systems

With the same assumption used for the trace kernel, the determinant kernel can be expressed in a closed form

$$k((x_0, A, C), (x_0', A', C')) = \det(W) \det(C^\top M C') \qquad (22)$$

where

$$M = e^{-\lambda} A M (A')^\top + x_0 (x_0')^\top \qquad (23)$$

N.B.

1. k is uniformly zero if C or M do not have full rank (e.g. $m > n$)
2. if we consider $\mathrm{E}_{x_0, x_0'}[k]$ to obtain independence from initial conditions, a closed form is possible only for $m = n = 1$

Linear dynamical systems - Continous time

The evolution of an LTI system can be described with the model

$$\frac{d}{dt}x(t) = Ax(t) \tag{24}$$

which has solution

$$x(t) = \exp(At)x(0) \tag{25}$$

Using the exponential discount $\mu(t) = e^{-\lambda t}$, the trace kernel becomes

$$k((x_0, A), (x_0, A')) = x_0^\top M x_0' \tag{26}$$

where

$$(A - \frac{\lambda}{4}\mathrm{I})M + M(A - \frac{\lambda}{4}\mathrm{I}) = -W \tag{27}$$

Connections with existing kernels
Subspaces angles

Given an ARMA model without noise, the Henkel matrix

$$\mathcal{Z} = \begin{bmatrix} y_0 & y_1 & y_2 & \dots \\ y_1 & y_2 & \dots & \\ \vdots & \ddots & & \end{bmatrix} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \begin{bmatrix} x_0 & x_1 & x_2 & \dots \end{bmatrix}$$
$$= \mathcal{O} \begin{bmatrix} x_0 & x_1 & x_2 & \dots \end{bmatrix} \quad (28)$$

lives in a subspace spanned by the columns of $\mathcal{O}$
The kernel

$$k((A, C), (A', C')) = \prod_{i=1}^{n} \cos^2(\theta_i) = \det(Q^\top Q')^2 \quad (29)$$

measure the angle between subspaces

Connections with existing kernels

Cepstral coefficient

Let $H(z)$ be the transfer function of an ARMA model without noise. The **cepstrum** is defined as

$$log(H(z)H^*(z^{-1})) = \sum_{n \in \mathbb{Z}} c_n z^{-n} \tag{30}$$

The Martin kernel

$$k((A, C), (A', C')) = \sum_{n=1}^{\infty} n c_n^* c_n' \tag{31}$$

is equivalent to the Subspace Angles approach (De Cock and De Moor, 2002)

Random walk on a graph can be seen as an ARMA model without noise

$$y_t = Cx_t \qquad (32)$$

$$x_{t+1} = Ax_t \qquad (33)$$

Assume equiprobable distribution over starting node

$$k(G_1, G_2) = \frac{1}{|V_1||V_2|}\mathbf{1}^\top M \mathbf{1} \qquad (34)$$

$$M = e^{-\lambda} A_1^\top M A_2 + C_1^\top W C_2 \qquad (35)$$

Equivalent to the geometric graph kernel of Gärtner et al. (2003)

Connections with existing kernels
Diffusion on graphs

Continuous diffusion process can be seen as the continuous LTI system

$$\frac{d}{dt}x(t) = Lx(t) \tag{36}$$

Assume same Laplacian matrix, different initial conditions, $\mu(t) = \delta_\tau(t)$

$$[K]_{i,j} = k((L, e_i), (L, e_j)) = [\exp(L\tau)^\top \exp(L\tau)]_{i,j} \tag{37}$$

If undirected graph, $L = L^\top$

$$K = \exp(2L\tau) \tag{38}$$

Same as the diffusion kernel of Kondor and Lafferty (2002).

Burkhardt (2004) uses *feature functions* $\Phi(x, p)$ where p is a polygon and x is a vertex index

Dynamic of the system from $x \leftarrow (x \mod n) + 1$

Compare two polygons from their trajectories $(\Phi(1, p), \ldots, \Phi(n, p))$

Equiprobable distribution on initial conditions (starting points on each polygon)

$$k(p, p') = \frac{1}{nn'} \sum_{x,x'=1}^{n,n'} \langle \Phi(x, p), \Phi(x', p') \rangle \tag{39}$$

Equivalent to Burkhardt (2004) but can be generalized to any kernel

Non-linear dynamical systems

The time evolution of a fully observable non-linear dynamical systems is governed by

$$x_{t+1} = f(x_t) \tag{40}$$

Find a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ which linearize the system

$$\Phi(x_{t+1}) = A\Phi(x_t) \tag{41}$$

and apply the same tools.
N.B.

- the space of possible solutions have to be restricted by imposing constraints on the structure of A
- the same constraints appears on M while solving the Sylvester's equations

Dynamic textures
Setup

The proposed kernels have been applied to the problem of comparing dynamic textures, using ARMA models

System identification performed using the sub-optimal closed form solution of Doretto et al. (2003)

Dataset of 150 sequences (grayscale 8 bit, 115x170 px, 120 frames each)

Sequences from 65 to 80 cannot be modeled as dynamic textures

Dynamic textures
Results

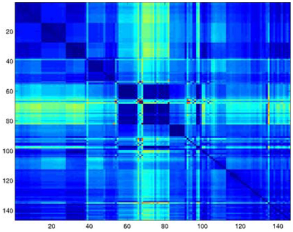


Figure: Binet-Cauchy kernel

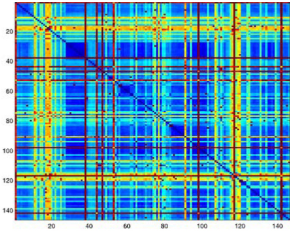


Figure: Martin kernel

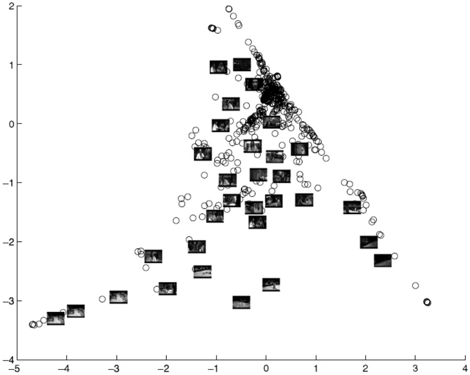
- Darker area $\Rightarrow$ low value kernel $\Rightarrow$ similar sequences
- The value of λ doesn't affect clustering
- Sequences 65-80 are recognized as novel

Dynamic textures
Parameters interpretation

- Initial condition of the system x_0: discriminate between sequences with same the foreground (dynamic behaviour) but different background (contained in the initial condition)
- Exponential decay λ: relative weights between short and long range interactions
- Dynamics (A, C): systems with A and A' with different dimensionality can be compared if the outputs y_t, y'_t belong to the same space. Can compare various levels of detail in the parametrization of the same system.

Video clips clustering

Clustering of random video clips (120 frames each)
The kernel has been used to compute the k-nearest neighborhood and LLE (Local Linear Embedding) has been used for clustering and embedding in the 2D space



Conclusion

- General framework which includes previous kernels as particular cases
- Can be applied to a wide class of dynamical systems (linear/non-linear, DT/CT)
- Computational-efficient closed form solution for linear dynamical systems
- Good results