A kernel for time series

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Centre for Computational Biology
Ecole des Mines de Paris, ParisTech

Journée signal, reconnaissance des formes et machines à noyaux,
Telecom Paris, June 8, 2007.

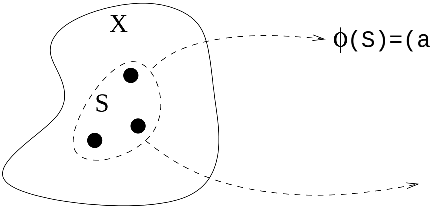
Outline

1. Motivations
2. An alignment kernel for time series
3. Experiments

Outline

1. Motivations
2. An alignment kernel for time series
3. Experiments

Kernel methods



φ(S)=(aatcgagtcac,atggacgtct,tgcactact)

$$K=\begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.6 \\ 0.3 & 0.6 & 1 \end{pmatrix}$$

- A positive definite kernel is a function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that any Gram matrix is positive semidefinite.
- Equivalently a p.d. kernel is an inner product after embedding $\mathcal{X}$ to a Hilbert space.
- Many algorithm for data analysis, called kernel methods, are based on p.d. kernels (SVMs, kernel PCA, kernel regression, ...)

Examples

Kernels for vectors

Classical kernels for vectors ($\mathcal{X} = \mathbb{R}^p$) include:

- The linear kernel

$$K_{lin}\left(\mathbf{x},\mathbf{x}'\right) = \mathbf{x}^\top \mathbf{x}' \,.$$

- The polynomial kernel

$$K_{poly}\left(\mathbf{x},\mathbf{x}'\right) = \left(\mathbf{x}^\top \mathbf{x}' + a\right)^d \,.$$

- The Gaussian RBF kernel:

$$K_{Gaussian}\left(\mathbf{x},\mathbf{x}'\right) = \exp\left(-\frac{\parallel \mathbf{x}-\mathbf{x}' \parallel^2}{2\sigma^2}\right) \,.$$

Examples

Kernels for vectors

Classical kernels for vectors ($\mathcal{X} = \mathbb{R}^p$) include:

- The linear kernel

$$K_{lin}\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{x}^\top \mathbf{x}' .$$

- The polynomial kernel

$$K_{poly}\left(\mathbf{x}, \mathbf{x}'\right) = \left(\mathbf{x}^\top \mathbf{x}' + a\right)^d .$$

- The Gaussian RBF kernel:

$$K_{Gaussian}\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\| \mathbf{x} - \mathbf{x}' \|^2}{2\sigma^2}\right) .$$

Examples

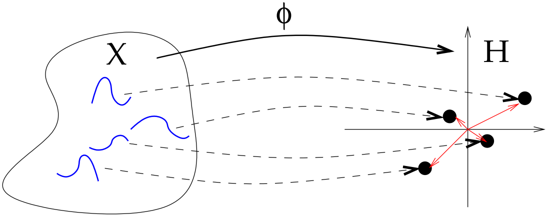
Kernels for vectors

Classical kernels for vectors ($\mathcal{X} = \mathbb{R}^p$) include:

- The linear kernel
$$K_{lin}\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{x}^\top \mathbf{x}' .$$
- The polynomial kernel
$$K_{poly}\left(\mathbf{x}, \mathbf{x}'\right) = \left(\mathbf{x}^\top \mathbf{x}' + a\right)^d .$$
- The Gaussian RBF kernel:
$$K_{Gaussian}\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\parallel \mathbf{x} - \mathbf{x}' \parallel^2}{2\sigma^2}\right) .$$

Kernel for time series

- Many problems in signal processing (in particular for speech recognition) involve finite-length discrete time series.
- In order to use kernel methods we need a kernel for time series.



Outline

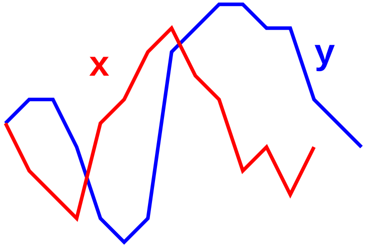
Notations

Time series

- $\mathcal{X}$ the set of observations ($\mathcal{X} = \mathbb{R}^d$)
- $\mathcal{X}^*$ the set of finite-length sequences of elements of $\mathcal{X}$
- $x = x_1 \dots x_m$ and $y = y_1 \dots y_n \in \mathcal{X}^*$ two finite-length sequences

Kernel

How to define a p.d. kernel $K(x, y)$ over $\mathcal{X}^*$?



Time series alignment

- How to compare 2 time series?

$$x = \mathtt{CGGSLIAMMW}$$
$$y = \mathtt{CLIVMMNRLMW}$$

- Find a good alignment:

```
CGGSLIAMMMMMMW
CCCCLIVMMNRLMW
```

- An alignment $\pi = (\pi_1, \pi_2)$ of length p is a pair of increasing p-tuples with no or unitary increments and no simultaneous repetitions

$$\pi_1 = (1,2,3,4,5,6,7,8,9,9,9,9,9,10)$$
$$\pi_2 = (1,1,1,1,2,3,4,5,6,7,8,9,10,11)$$

Time series alignment

- How to compare 2 time series?

$$x = \mathtt{CGGSLIAMMW}$$
$$y = \mathtt{CLIVMMNRLMW}$$

- Find a good alignment:

$$\mathtt{CGGSLIAMMMMMMW}$$
$$\mathtt{CCCCLIVMMNRLMW}$$

- An alignment $\pi = (\pi_1, \pi_2)$ of length p is a pair of increasing p-tuples with no or unitary increments and no simultaneous repetitions

$$\pi_1 = (1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 9, 9, 9, 10)$$
$$\pi_2 = (1, 1, 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$$

Time series alignment

- How to compare 2 time series?

$$x = \texttt{CGGSLIAMMW}$$
$$y = \texttt{CLIVMMNRLMW}$$

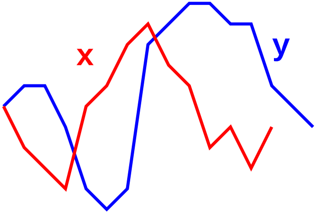
- Find a good alignment:

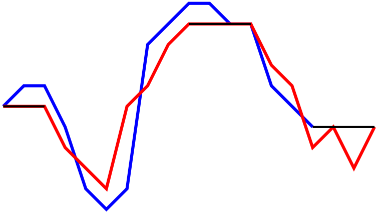
```
CGGSLIAMMMMMMW
CCCCLIVMMNRLMW
```

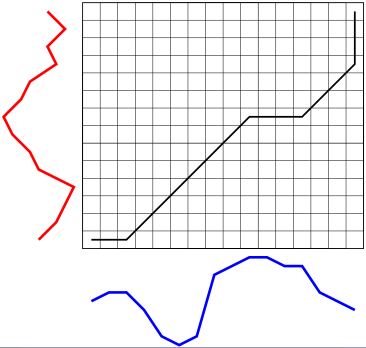
- An alignment $\pi = (\pi_1, \pi_2)$ of length p is a pair of increasing p-tuples with no or unitary increments and no simultaneous repetitions

$$\pi_1 = (1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 9, 9, 9, 10)$$
$$\pi_2 = (1, 1, 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$$

Illustration



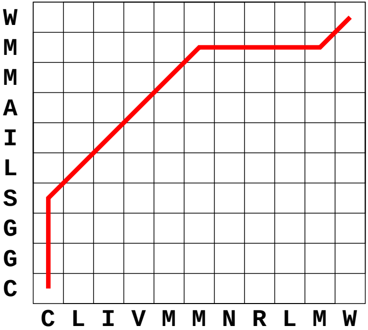




Dynamic time warping

$$S(\pi) = -\sum_{i=1}^{|\pi|} ||x_{\pi_1(i)} - y_{\pi_2(i)}||^2 \,.$$

$$\pi^* = \arg\max_{\pi} \frac{1}{|\pi|} S(\pi) \quad \text{in } O(|x| \times |y|)$$



DTW kernels

Related work

- Bahlmann et al. (2002):

$$K_{DTW1}(x, y) = e^{\frac{1}{|\pi^*|} S(\pi^*)} = \arg\max_{\pi} \exp\left(-\frac{1}{|\pi|} \sum_{i=1}^{|\pi|} ||x_{\pi_1(i)} - y_{\pi_2(i)}||^2 \right)$$

- Shimodaira et al. (2002)

$$K_{DTW2}(x, y) = \arg\max_{\pi} \frac{1}{|\pi|} \sum_{i=1}^{|\pi|} \exp\left(-\frac{1}{\sigma^2} ||x_{\pi_1(i)} - y_{\pi_2(i)}||^2 \right)$$

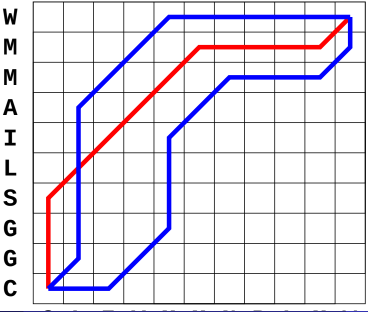
- Neither of them is p.d. in general.

A softmax DTW kernel

Definition

$$K_{softmax}(x,y) = \sum_{\pi} e^{S(\pi)}$$

$$= \sum_{\pi} \prod_{i=1}^{|\pi|} e^{-\beta||x_{\pi_1(i)} - y_{\pi_2(i)}||^2}$$



Positive definiteness of the softmax DTW kernel

Theorem

Let k be a p.d. kernel over $\mathcal{X}$ such that $\frac{k}{k+1}$ is also p.d. Then the softmax DTW kernel:

$$K(x,y) = \sum_{\pi} \prod_{i=1}^{|\pi|} k\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right)$$

is p.d. over $\mathcal{X}^*$.

Sketch of the proof

- Similar to convolution kernels (Haussler, 1999)
- Specific treatment to deal with the multiplicity of matchings when letters are repeated.
- Remark: slightly different from the local alignment kernel for strings (Saigo et al., 2004): gaps are replaced by repetitions.

Examples

Lemma

Let χ be a p.d. kernel such that $|\chi| < 1$. Then the kernel:

$$k = \sum_{i=1}^{\infty} \chi^i = \frac{\chi}{1-\chi}$$

is p.d. and $k/(k+1)$ is p.d. too.

Example

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

satisfies the conditions of the theorem.

Examples

Lemma

Let χ be a p.d. kernel such that $|\chi| < 1$. Then the kernel:

$$k = \sum_{i=1}^{\infty} \chi^i = \frac{\chi}{1 - \chi}$$

is p.d. and $k/(k+1)$ is p.d. too.

Example

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

satisfies the conditions of the theorem.

Computation

Dynamic programming

The softmax DTW kernel can be computed in $O(|x||y|)$ as:

$$K(x,y) = M_{n,m}$$

with:

$$\begin{aligned}
M_{0,0} &= 1 \\
M_{0,j} &= M_{j,0} = 0 \quad \text{for } j \geq 1\,, \\
M_{i,j} &= \left(M_{i,j-1} + M_{i-1,j} + M_{i-1,j-1}\right) k(x_i, y_j) \quad \text{for } i,j \geq 1\,.
\end{aligned}$$

Summary

- Taking the softmax instead of the max allows to quantify more subtle similarities
- Same computational complexity
- The resulting kernel is p.d., contrary to other DTW kernels.
- BUT: in practice, danger of diagonal dominance and "massaging" the Gram matrix might be required...

Outline

Data

- Isolated-word recognition experiments
- TI46 E-set database:
  - 3724 spoken letters: B,C,D,E,G,P,T,V,Z
  - Training: 1433 utterances
  - Testing: 2291 utterances
- 13-dimensional MFCCs (25 ms window, 10 ms shift)

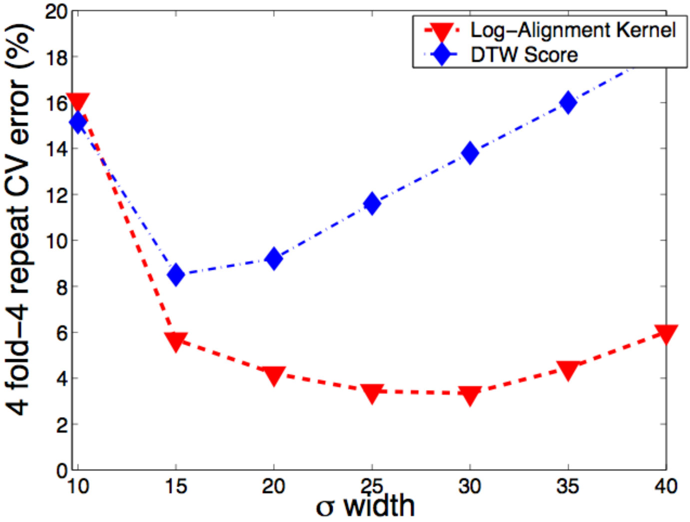
Algorithms

- HMM model
  - left-to-right model with 6 states and 5 mixtures
  - diagonal covariance
  - 39-dimensional feature vectors (13-dim. MFCCs, delta and acceleration coefficients)
- SVM + DTW kernel
- SVM + softmax DTW kernel

Results

Algo	Error rate
HMM	11.7%
SVM + DTW	11.5%
SVM + softmax DTW	5.4%

Effect of the kernel width



Conclusion

- A softmax version of DTW for time series
- Positive definite under mild assumption
- Excellent experimental results on simple speech recognition tasks

Acknowledgements

This is a joint work with

- Marco Cuturi (ISM, Tokyo)
- Øystein Birkenes (NTNU), Tomoko Matsui (ISM)

Availability

Paper and code available at:
`http://www.ism.ac.jp/ cuturi/articles/alignment2006arx`
Proceedings of ICASSP'07.