Kernels for discrete objects

Jean-Philippe.Vert@mines.org
5~ETBg3X2=3X8&5f=j
%P%$%*%$%s%U%)%/%F%#%/%9%;%s%?!<6b5W8&

2001G/117n21F|

Motivations

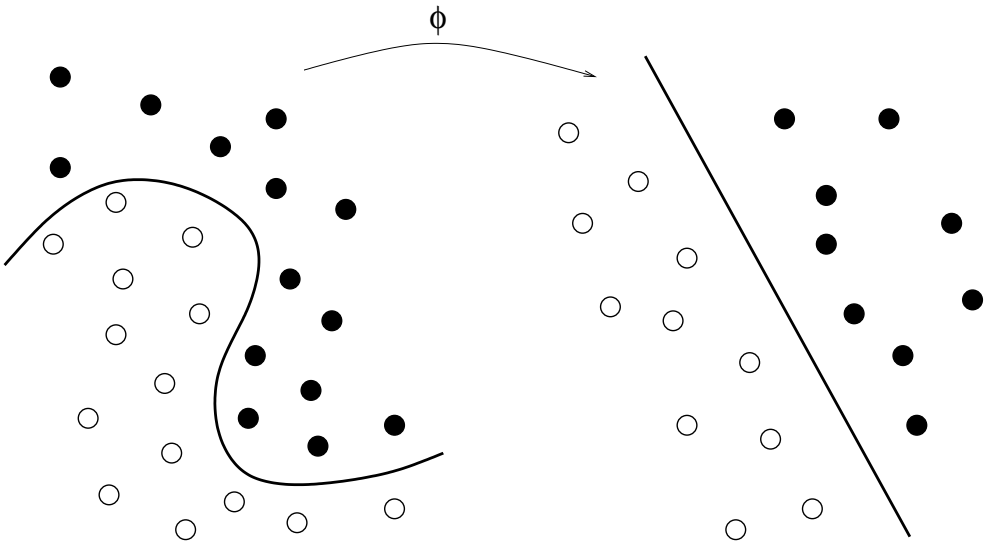
Using kernel-based methods:

- Support vector machines (SVM)
- kernel principal component analysis (KPCA)
- kernel clustering (...)

to analyse biological data:

- sequences
- trees, graphs

Example : SVM (1)



- Objects $x \in \mathcal{X}$ are mapped by $\Phi(x)$ to the feature space

- The linear separation with largest margin is found in the feature space

Example : SVM (2)

- The resulting classifier is:

$$f(x) = sgn\left(\sum_i \lambda_i \Phi(x).\Phi(x_i)\right) \qquad (1)$$

  where the λ_i solve a constrained optimization problem.

- All computation only involve Φ through the products:

$$K(x, y) = \Phi(x).\Phi(y) \qquad (2)$$

- $K(.,.)$ is called a kernel

Famous kernels

- Polynomial:

$$K(x, y) = (x.y + c)^d$$

- Gaussian kernel (or radial kernel)

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

- Sigmoid kernel:

$$K(x, y) = \tanh\left(\kappa x.y + \theta\right)$$

... but the object space must be $\mathcal{X} = R^m$

Alternative kernels for finite $\mathcal{X}$

- map $\mathcal{X}$ to R^m by direct feature extraction and use the famous kernels

- Fisher kernel (Jaakkola, Haussler): use a parametric statistical model:

$$\left\{P_\theta \in \mathcal{M}_+^1(\mathcal{X}), \theta \in \Theta \subset R^m\right\}$$

and map any object $x \in \mathcal{X}$ to the score vector:

$$\nabla_\theta \log p_\theta(x) \in R^m$$

Alternative kernels for finite $\mathcal{X}$ (2)

If p is a probability distribution on $\mathcal{X}$ then two basic kernels can be built:

- Product kernel:

  $$K_{prod}(x, y) = p(x)p(y)$$

  then $m = 1$, $\Phi(x) = p(x)$, the resulting SVM classifies a new example based on its probability only (larger or smaller than a threshold). *Two objects are close if they have close probabilities.*

- Diagonal kernel:

  $$K_{diag}(x, y) = p(x)\delta(x, y)$$

  then $m = |\mathcal{X}|$, $\Phi(x) = \sqrt{p(x)}e_x$, the resulting SVM checks whether the object has been seen in the training set. *Two objects are close if they are the same.*

Interpolated kernel (1)

- Suppose $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, i.e. objects $x = (x_1, x_2)$ are composite.

- Interpolated kernel:

$$K(x, y) = K_{prod}(x_1, y_1) K_{diag}(x_2, y_2)$$

- Two objects are close if
  - x_1 and y_1 have close probability (loose),
  - x_2 and y_2 are the same (strong)

Interpolated kernel (2)

- More generally $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$
- A set of index subsets is given: $V = \{I_1, \ldots, I_v\}$ where each $I_i \subset \{1, \ldots, p\}$
- Interpolated kernel:

$$K(x, y) = \frac{1}{|V|} \sum_{I \in V} K_{prod}(x_I, y_I) K_{diag}(x_{\bar{I}}, y_{\bar{I}})$$

- x and y are close if they share rare common subsets

Implementation

- Not efficient for general p and V, but factorization possible for particular choices.

- Example: $V = \mathcal{P}(\{1, \ldots, N\})$, product distribution $p(x) = \prod_{j=1}^{N} p_j(x_j)$. Then:

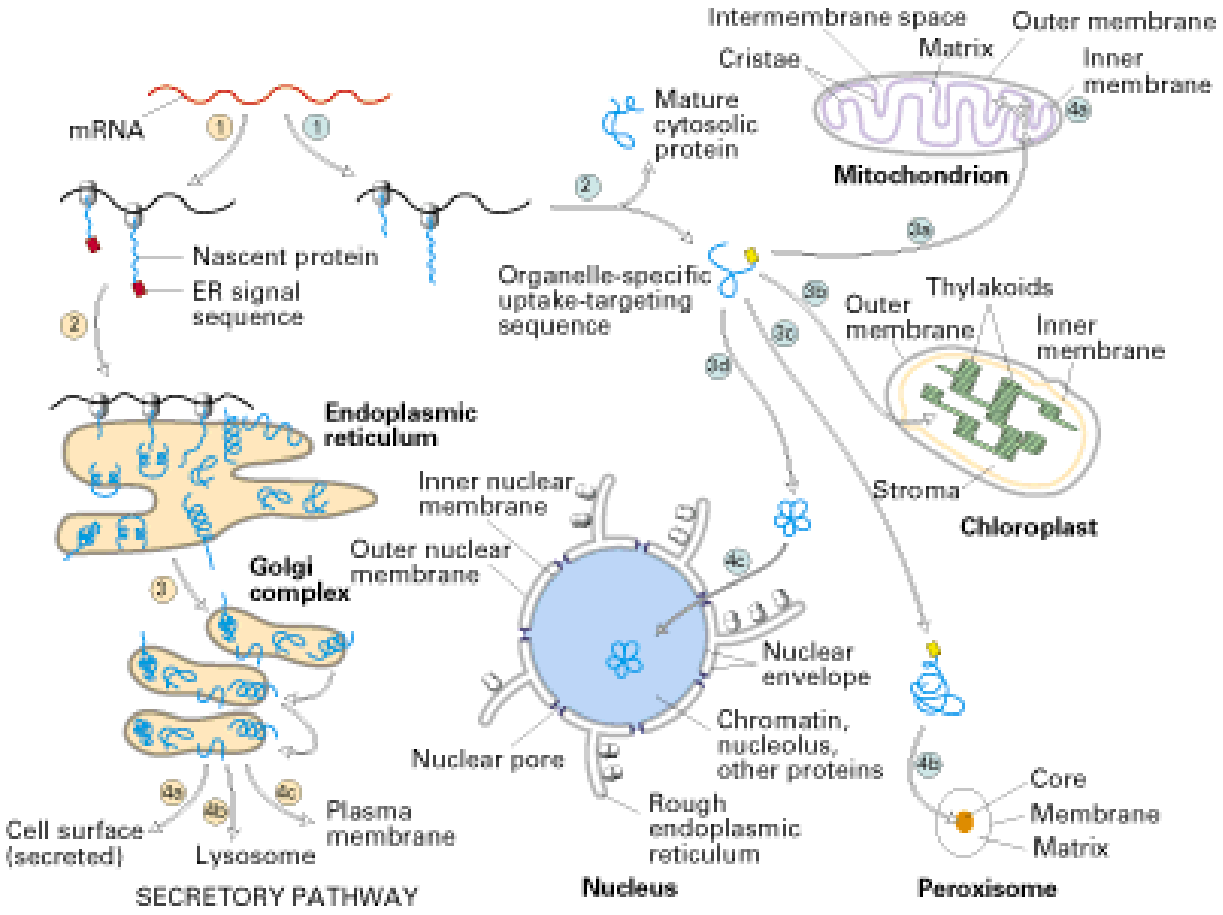
$$K_{\mathcal{V}}(x, y) = \frac{1}{2^N} \prod_{i=1}^{N} \phi_i(x_i, y_i), \qquad (3)$$

with:

$$\phi_i(x_i, y_i) = p_i(x_i) + p_i(x_i)^2 \text{ if } x_i = y_i, \quad (4)$$

$$\phi_i(x_i, y_i) = p_i(x_i) p_i(y_i) \text{ if } x_i \neq y_i \qquad (5)$$

Example : Protein sorting



(from "Molecular cell biology", *Lodish et al.*, 1999)

Signal peptide recognition (1)

- After (or during) synthesis in the cytoplasm proteins are sorted according to their final destination

- Plasma membrane, lysosomal or secreted proteins follow the secretory pathway, through the endoplasmic reticulum and the Golgi and post-Golgi complex.

- A nascent protein is targeted to the secretory pathway if it contains particular signal peptides near the N-terminal region.

Signal peptide recognition (2)

Protein	-1 +1
(1)	MKANAKTIIAGMIALAISHTAMA EE...
(2)	MKQSTIALALLPLLFTPVTKA RT...
(3)	MKATKLVLGAVILGSTLLAG CS...

(1):Leucine-binding protein, (2):Pre-alkaline phosphatase, (3)Pre-lipoprotein

- One or more positively charged aminoacids following M
- Then 6-12 hydrophobic residues (in red)
- (-3,-1) : small uncharged residues

Signal peptide recognition (3)

- Challenge : classification of aminoacids windows of length 10, positive if cleavage occurs between -1 and +1:

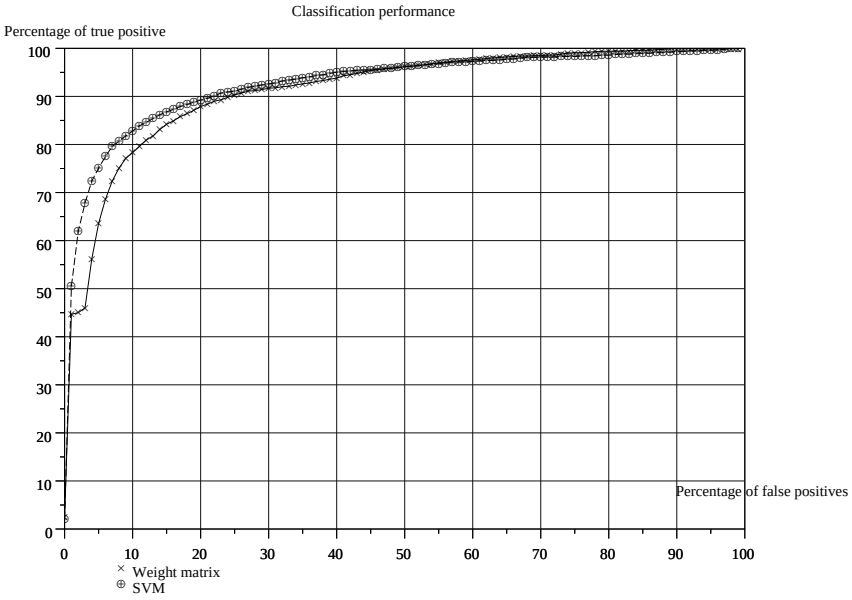
$$[x_{-8}, x_{-7}, \ldots, x_{-1}, x_1, x_2]$$

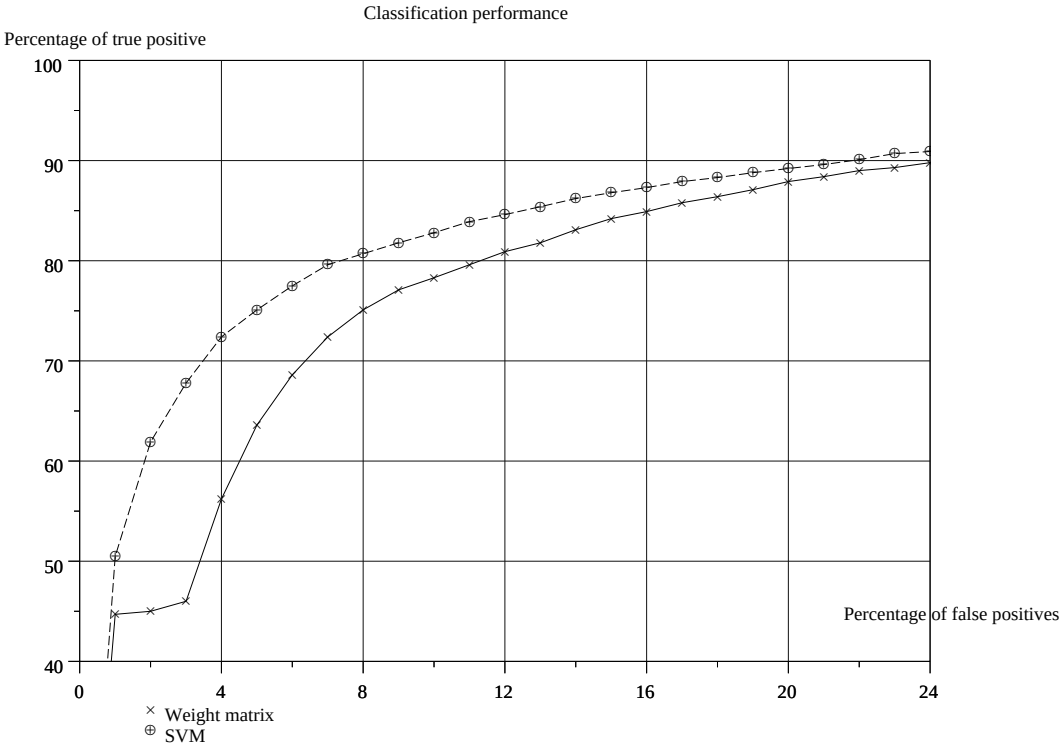
- 1,418 positive examples, 65,216 negative examples

- Computation of a weight matrix (naive Bayes classifier)

- Computation of the interpolated kernel from the weight matrix, and SVM classifier

Results : ROC curves





Conclusion

- An other way to derive a kernel from a probability distribution
- Based on the decomposable nature of the objects
- Encouraging results on real-world application