Classification of biological sequences with kernel methods

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

Center for Computational Biology
Ecole des Mines de Paris

CMLA symposium, Ecole normale supérieure de Cachan, May 18, 2006

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods

2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

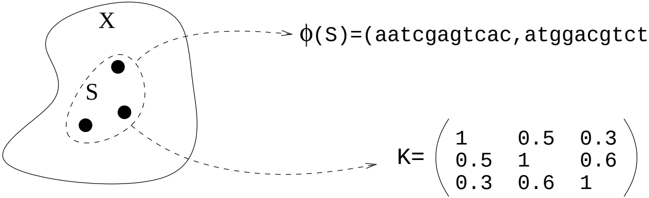
Kernels and Kernel Methods

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Motivations

- Develop versatile algorithms to process and analyze data
- No hypothesis made regarding the type of data (vectors, strings, graphs, images, ...)
- Instead we study methods based on pairwise comparisons.



$\phi(S)=(\texttt{aatcgagtcac},\texttt{atggacgtct},\texttt{tgcactact})$

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

Positive Definite Kernels

Definition

A positive definite (p.d.) kernel on the set $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric:

$$\forall \left(\mathbf{x}, \mathbf{x}'\right) \in \mathcal{X}^2, \quad K\left(\mathbf{x}, \mathbf{x}'\right) = K\left(\mathbf{x}', \mathbf{x}\right),$$

and which satisfies, for all $N \in \mathbb{N}$, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathcal{X}^N$ et $(a_1, a_2, \ldots, a_N) \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 0.$$

Examples

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:

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

Kernels as Inner Products

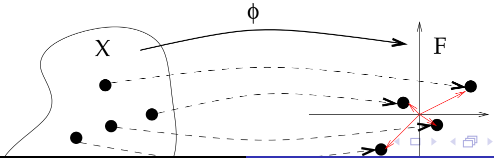
Theorem (Aronszajn, 1950)

K is a p.d. kernel on the set $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping

$$\Phi : \mathcal{X} \mapsto \mathcal{H} \,,$$

such that, for any $\mathbf{x}, \mathbf{x}'$ in $\mathcal{X}$:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = \left\langle \Phi\left(\mathbf{x}\right), \Phi\left(\mathbf{x}'\right) \right\rangle_{\mathcal{H}} \;.$$



Reproducing Kernel Hilbert Space

- To each p.d. kernel on $\mathcal{X}$ is associated a unique Hilbert space of function $\mathcal{X} \to \mathbb{R}$, called the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$.
- Typical functions are:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \;,$$

with norm

$$\| f \|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \;.$$

Reproducing property

- For any $\mathbf{x} \in \mathcal{X}$ let $K_{\mathbf{x}} : \mathcal{X} \rightarrow \mathbb{R}$ be defined by:

$$K_{\mathbf{x}}\left(\mathbf{x}'\right) = K\left(\mathbf{x}, \mathbf{x}'\right), \quad \forall \mathbf{x}' \in \mathcal{X} .$$

- In the RKHS it holds that:

$$f\left(\mathbf{x}\right) = \left\langle f, K_{\mathbf{x}} \right\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}, \mathbf{x} \in \mathcal{X} .$$

- Reproducing property:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = \left\langle K_{\mathbf{x}}, K_{\mathbf{x}'} \right\rangle_{\mathcal{H}}, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} .$$

Smoothness functional

By Cauchy-Schwarz we have, for any function $f \in \mathcal{H}$ and any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$:

$$\begin{aligned} \left| f(\mathbf{x}) - f\left(\mathbf{x}'\right) \right| &= \left| \langle f, K_{\mathbf{x}} - K_{\mathbf{x}'} \rangle_{\mathcal{H}} \right| \\ &\leq \| f \|_{\mathcal{H}} \times \| K_{\mathbf{x}} - K_{\mathbf{x}'} \|_{\mathcal{H}} \\ &= \| f \|_{\mathcal{H}} \times d_K\left(\mathbf{x}, \mathbf{x}'\right) . \end{aligned}$$

The norm of a function in the RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the geometry defined by the kernel. Small norm $\Longrightarrow$ slow variations.

Examples

- Linear kernel:

$$\begin{cases} f(\mathbf{x}) & = w^\top x \,, \\ \| f \|_{\mathcal{H}} & = \| w \|_2 \,. \end{cases}$$

- Gaussian RBF kernel

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(- \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{2\sigma^2} \right) ,$$

$$\| f \|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \exp\left(- \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{2\sigma^2} \right)$$

$$= \int \left| \hat{f}(\omega) \right|^2 e^{\frac{\sigma^2 \omega^2}{2}} d\omega \,.$$

Examples

- Linear kernel:

$$\begin{cases} f(\mathbf{x}) & = w^\top x \,, \\ \| f \|_{\mathcal{H}} & = \| w \|_2 \,. \end{cases}$$

- Gaussian RBF kernel

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(- \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{2\sigma^2} \right) \,,$$

$$\| f \|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \exp\left(- \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{2\sigma^2} \right)$$

$$= \int \left| \hat{f}(\omega) \right|^2 e^{\frac{\sigma^2 \omega^2}{2}} d\omega \,.$$

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Pattern recognition and regression

- Input variables $\mathbf{x} \in \mathcal{X}$
- Output $y \in \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$ (pattern recognition) or $\mathcal{Y} = \mathbb{R}$ (regression)
- Training set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
- Goal: learn the mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$

Kernel methods

1. Define a loss function $L(y, \hat{y})$
2. Solve the problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f\left(\mathbf{x}_i\right)\right) + \lambda \| f \|_{\mathcal{H}}^2 \, .$$

λ controls the trade-off between fitting the data and being a smooth function.

Examples

- Support vector machines for classification:

$$L_{hinge}(y, \hat{y}) = \max(0, 1 - y\hat{y}) \; .$$

- Kernel logistic regression

$$L_{logit} = \log\left(1 + e^{-y\hat{y}}\right) \; .$$

- Kernel ridge regression

$$L_{square}(y, \hat{y}) = (y - \hat{y})^2 \; .$$

Summary

- A kernel defines an implicit geometry on the space of data, although data do not need to have any prior geometric/algebric structure
- Kernel methods learn functions that tend to be smooth with respect to this geometry
- Kernel engineering is the problem of designing specific kernel for specific data and specific tasks. Good place to put prior knowledge!
- We will now see on a practical examples different technical tricks to design kernels.

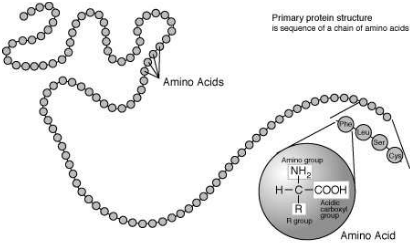
Kernels for Biological Sequences

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods

2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Protein sequence



A : Alanine	V : Valine	L : Leucine
F : Phenylalanine	P : Proline	M : Méthionine
E : Acide glutamique	K : Lysine	R : Arginine
T : Threonine	C : Cysteine	N : Asparagine
H : Histidine	V : Thyrosine	W : Tryptophane
I : Isoleucine	S : Sérine	Q : Glutamine
D : Acide aspartique	G : Glycine	

Challenges with protein sequences

- A protein sequences can be seen as a variable-length sequence over the 20-letter alphabet of amino-acids, e.g., insuline:
  FVNQHLCGSHLVEALYLVCGERGFFYTPKA
- These sequences are produced at a fast rate (result of the sequencing programs)
- Need for algorithms to compare, classify, analyze these sequences
- Applications: classification into functional or structural classes, prediction of cellular localization and interactions, ...

Kernels for protein sequences

- Kernel methods have been widely investigated since Jaakkola et al.'s seminal paper (1998).
- What is a good kernel?
  - it should be mathematically valid (symmetric, p.d. or c.p.d.)
  - fast to compute
  - adapted to the problem (give good performances)

Kernel engineering for protein sequences

- Define a (possibly high-dimensional) feature space of interest
  - Physico-chemical kernels
  - Spectrum, mismatch, substring kernels
  - Pairwise, motif kernels
- Derive a kernel from a generative model
  - Fisher kernel
  - Mutual information kernel
  - Marginalized kernel
- Derive a kernel from a similarity measure
  - Local alignment kernel

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Physico-chemical kernels

How to embed explicitly a sequence $\mathbf{x} \in \mathcal{X}$ into a vector $\Phi(\mathbf{x}) \in \mathbb{R}^n$?

Extract relevant features, such as:

- length of the sequence
- time series analysis of numerical physico-chemical properties of amino-acids along the sequence (e.g., polarity, hydrophobicity), using for example:
  - Fourier transforms (Wang et al., 2004)
  - Autocorrelation functions (Zhang et al., 2003)

$$r_j = \frac{1}{n-j} \sum_{i=1}^{n-j} h_i h_{i+j}$$

Substring indexation

Alternatively, index the feature space by fixed-length strings, i.e.,

$$\Phi(\mathbf{x}) = (\Phi_u(\mathbf{x}))_{u \in \mathcal{A}^k}$$

where $\Phi_u(\mathbf{x})$ can be:

- the number of occurrences of u in $\mathbf{x}$ (without gaps) : spectrum kernel (Leslie et al., 2002)
- the number of occurrences of u in $\mathbf{x}$ up to m mismatches (without gaps) : mismatch kernel (Leslie et al., 2004)
- the number of occurrences of u in $\mathbf{x}$ allowing gaps, with a weight decaying exponentially with the number of gaps : substring kernel (Lohdi et al., 2002)

Substring indexation in practice

- Implementation in $O(|\mathbf{x}| + |\mathbf{x}'|)$ in memory and time for the spectrum and mismatch kernels (with suffix trees)
- Implementation in $O(|\mathbf{x}| \times |\mathbf{x}'|)$ in memory and time for the substring kernels
- The feature space has high dimension ($|\mathcal{A}|^k$), so learning requires regularized methods (such as SVM)

Dictionary-based indexation

- Chose a dictionary of sequences $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
- Chose a measure of similarity $s(\mathbf{x}, \mathbf{x}')$
- Define the mapping $\Phi_{\mathcal{D}}(\mathbf{x}) = (s(\mathbf{x}, \mathbf{x}_i))_{\mathbf{x}_i \in \mathcal{D}}$

This includes:

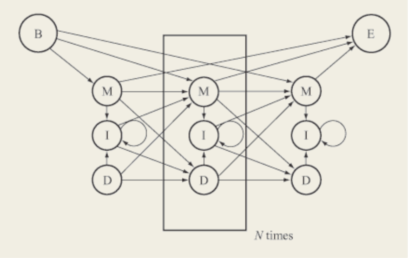
- Motif kernels (Logan et al., 2001): the dictionary is a library of motifs, the similarity function is a matching function
- Pairwise kernel (Liao & Noble, 2003): the dictionary is the training set, the similarity is a classical measure of similarity between sequences.

Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Probabilistic models for sequences

Probabilistic modeling of biological sequences is older than kernel designs. Important models include HMM for protein sequences, SCFG for RNA sequences.



A model is a family of distribution

$$\{P_\theta, \theta \in \Theta \subset \mathbb{R}^m\} \subset \mathcal{M}_1^+(\mathcal{X})$$

Fisher kernel

- Fix a parameter $\theta_0 \in \Theta$ (e.g., by maximum likelihood over a training set of sequences)
- For each sequence $\mathbf{x}$, compute the Fisher score vector:

$$\Phi_{\theta_0}(\mathbf{x}) = \nabla_\theta \log P_\theta(\mathbf{x})|_{\theta=\theta_0} \ .$$

- Form the kernel (Jaakkola et al., 1998):

$$K\left(\mathbf{x},\mathbf{x}'\right) = \Phi_{\theta_0}(\mathbf{x})^\top I(\theta_0)^{-1} \Phi_{\theta_0}(\mathbf{x}') \ ,$$

where $I(\theta_0) = E_{\theta_0}\left[\Phi_{\theta_0}(\mathbf{x})\Phi_{\theta_0}(\mathbf{x})^\top\right]$ is the Fisher information matrix.

Fisher kernel in practice

- $\Phi_{\theta_0}(\mathbf{x})$ can be computed explicitly for many models (e.g., HMMs)
- $I(\theta_0)$ is often replaced by the identity matrix
- Several different models (i.e., different θ_0) can be trained and combined
- Feature vectors are explicitly computed

Mutual information kernels

- Chose a prior $w(d\theta)$ on the measurable set Θ
- Form the kernel (Seeger, 2002):

$$K(\mathbf{x}, \mathbf{x}') = \int_{\theta \in \Theta} P_\theta(\mathbf{x}) P_\theta(\mathbf{x}') w(d\theta) \,.$$

- No explicit computation of a finite-dimensional feature vector
- $K(\mathbf{x}, \mathbf{x}') = < \phi(\mathbf{x}), \phi(\mathbf{x}') >_{L_2(w)}$ with

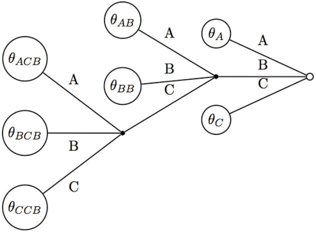
$$\phi(\mathbf{x}) = (P_\theta(\mathbf{x}))_{\theta \in \Theta} \,.$$

The context-tree kernel

Consider a variable-memory Markov chain:

$$P_{\mathcal{D},\theta}(\mathbf{x}) = P_{\mathcal{D},\theta}\left(x_1 \ldots x_D\right) \prod_{i=D+1}^{n} P_{\mathcal{D},\theta}\left(x_i \mid x_{i-D} \ldots x_{i-1}\right)$$

- $\mathcal{D}$ is a suffix tree
- $\theta \in \Sigma^{\mathcal{D}}$ is a set of conditional probabilities (multinomials)



The context-tree kernel (cont.)

- For particular choices of priors, the context-tree kernel:

$$K\left(\mathbf{x},\mathbf{x}'\right) = \sum_{\mathcal{D}} \int_{\theta\in\Sigma^{\mathcal{D}}} P_{\mathcal{D},\theta}(\mathbf{x}) P_{\mathcal{D},\theta}(\mathbf{x}') w(d\theta|\mathcal{D})\pi(\mathcal{D})$$

  can be computed in $O(|\mathbf{x}| + |\mathbf{x}'|)$ with a variant of the Context-Tree Weighting algorithm (Cuturi et al., 2004).
- This is a valid mutual information kernel.
- The similarity is related to information-theoretical measure of mutual information between strings.

Marginalized kernels

- For any observed data $\mathbf{x} \in \mathcal{X}$, let a latent variable $\mathbf{y} \in \mathcal{Y}$ be associated probabilistically through a conditional probability $P_{\mathbf{x}}(d\mathbf{y})$.
- Let $K_{\mathcal{Z}}$ be a kernel for the complete data $\mathbf{z} = (\mathbf{x}, \mathbf{y})$
- Then the following kernel is a valid kernel on $\mathcal{X}$, called a marginalized kernel (Tsuda et al., 2002):

$$\begin{aligned} K_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right) &:= E_{P_{\mathbf{x}}(d\mathbf{y}) \times P_{\mathbf{x}'}(d\mathbf{y}')} K_{\mathcal{Z}}\left(\mathbf{z}, \mathbf{z}'\right) \\ &= \int \int K_{\mathcal{Z}}\left((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')\right) P_{\mathbf{x}}(d\mathbf{y}) P_{\mathbf{x}'}(d\mathbf{y}') \ . \end{aligned}$$

Marginalized kernels in practice

- Spectrum kernel on the hidden states of a HMM for protein sequences (Tsuda et al., 2002)
- Kernels for RNA sequences based on SCFG (Kin et al., 2002)
- Kernels for graphs based on random walks on graphs (Kashima et al., 2004)
- Kernels for multiple alignments based on phylogenetic models (Vert et al., 2005)

Outline

1. Kernels and kernel methods
    - Kernels
    - Kernel Methods

2. Kernels for biological sequences
    - Motivations
    - Feature space approach
    - Using generative models
    - Derive from a similarity measure
    - Application: remote homology detection

Sequence alignment

How to compare 2 sequences?

$$\mathbf{x}_1 = \texttt{CGGSLIAMMWFGV}$$
$$\mathbf{x}_2 = \texttt{CLIVMMNRLMWFGV}$$

Find a good alignment:

```
CGGSLIAMM----WFGV
|...|||||....||||
C---LIVMMNRLMWFGV
```

Alignment score

In order to quantify the relevance of an alignment π, define:

- a substitution matrix $S \in \mathbb{R}^{\mathcal{A} \times \mathcal{A}}$
- a gap penalty function $g : \mathbb{N} \to \mathbb{R}$

Any alignment is then scored as follows

```
CGGSLIAMM----WFGV
|...|||||....||||
C---LIVMMNRLMWFGV
```

$$s_{S,g}(\pi) = S(C,C) + S(L,L) + S(I,I) + S(A,V) + 2S(M,M) \\ + S(W,W) + S(F,F) + S(G,G) + S(V,V) - g(3) - g(4)$$

Local alignment kernel

- The widely-used Smith-Waterman local alignment score is defined by:

$$SW_{S,g}(\mathbf{x},\mathbf{y}) := \max_{\pi \in \Pi(\mathbf{x},\mathbf{y})} s_{S,g}(\pi).$$

- It is symmetric, but not positive definite...
- The local alignment kernel:

$$K_{LA}^{(\beta)}(\mathbf{x},\mathbf{y}) = \sum_{\pi \in \Pi(\mathbf{x},\mathbf{y})} \exp\left(\beta s\left(\mathbf{x},\mathbf{y},\pi\right)\right),$$

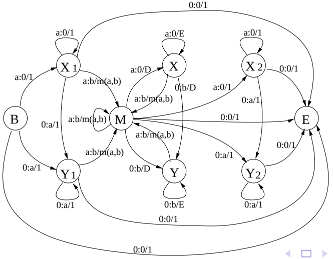
is symmetric positive definite (Vert et al., 2004).

LA kernel in practice

- LA kernel is p.d. because it is a convolution kernel (Haussler, 1999):

$$K_{LA}^{(\beta)} = \sum_{n=0}^{\infty} K_0 \star \left(K_a^{(\beta)} \star K_g^{(\beta)}\right)^{(n-1)} \star K_a^{(\beta)} \star K_0 .$$

- Implementation by dynamic programming in $O(|\mathbf{x}| \times |\mathbf{x}'|)$

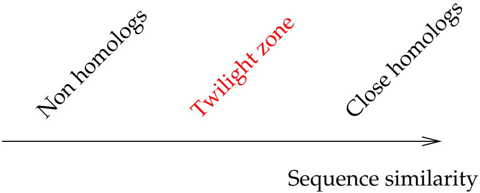


Outline

1. Kernels and kernel methods
   - Kernels
   - Kernel Methods

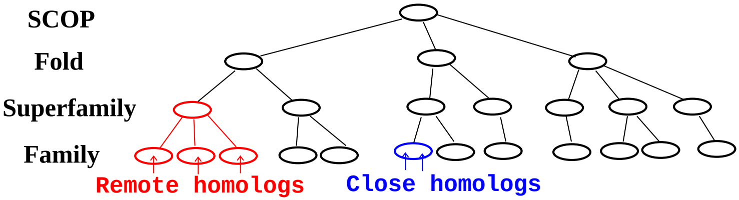
2. Kernels for biological sequences
   - Motivations
   - Feature space approach
   - Using generative models
   - Derive from a similarity measure
   - Application: remote homology detection

Remote homology



- Homologs have common ancestors
- Structures and functions are more conserved than sequences
- Remote homologs can not be detected by direct sequence comparison

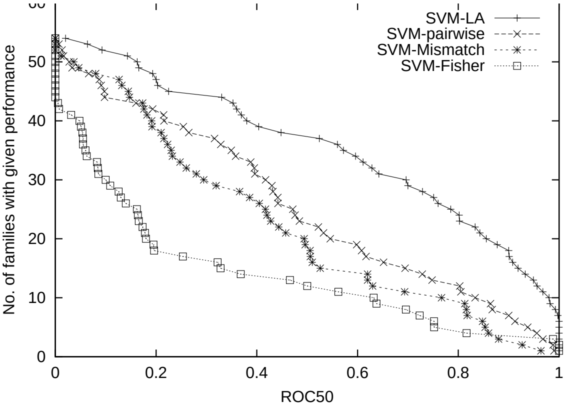
SCOP database



A benchmark experiment

- Goal: recognize directly the superfamily
- Training: for a sequence of interest, positive examples come from the same superfamily, but different families. Negative from other superfamilies.
- Test: predict the superfamily.

Difference in performance



Performance on the SCOP superfamily recognition benchmark (from Vert et al., 2004).

Summary

- Kernel methods offer interesting opportunities for non-vectorial and structured data.
- Good kernel design is important for each data and each task. Performance is not the only criterion.
- Still an art, although principled ways have started to emerge.
- Latest trends: semi-supervised kernels, combination of kernels.