Random features for kernel learning

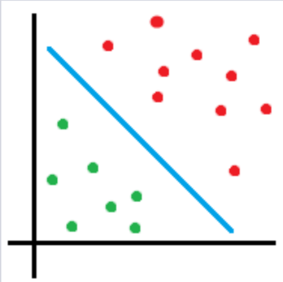
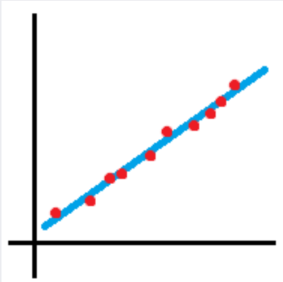
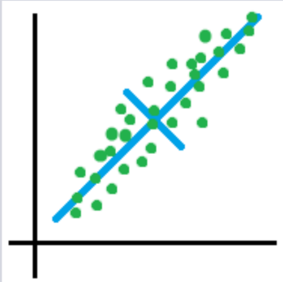
Purushottam Kar
Joint work with Harish Karnick

Department of Computer Science and Engineering
Indian Institute of Technology Kanpur

Mysore Workshop on Machine Learning, Aug 1-5, 2012

Kernel algorithms : an overview

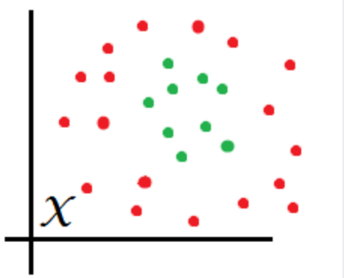
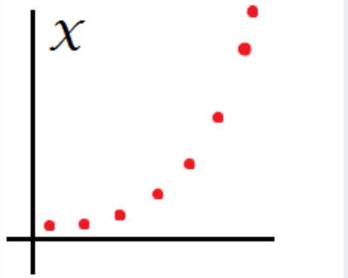
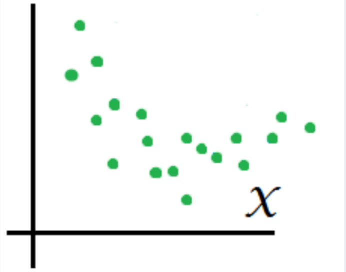
- Learning with simple linear models

Problem	Classification	Regression	PCA
Solution	Linear hyper-plane	Linear regression	Linear PCA
A thousand words			

- Linear models may under-fit data
  - Poor classification accuracy (think of the XOR problem)
  - Poor interpolation for regression
  - Data may appear near isotropic to linear model

Kernel algorithms : an overview

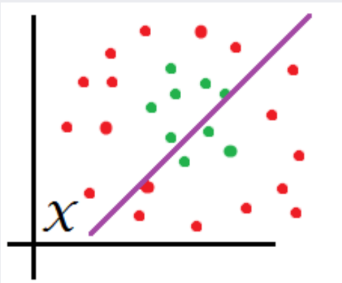
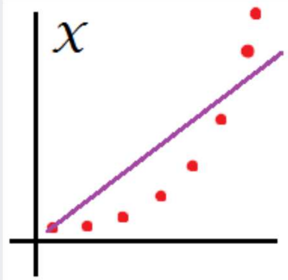
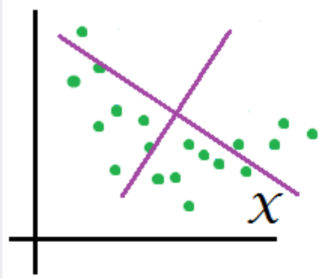
- Kernels exploit various invariances* in these algorithms
  - Use embedding $\Phi_K : \mathcal{X} \to \mathcal{H}_K$ where linear models are good !
  - $K(x, x') = \langle \Phi_K(x), \Phi_K(x') \rangle$

Problem	Classification	Regression	PCA
A thousand words			

* Obliviousness to the precise embedding : recall our favorite representer theorems !

Kernel algorithms : an overview

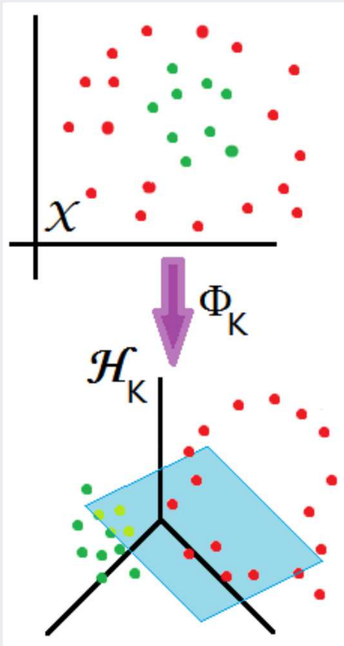
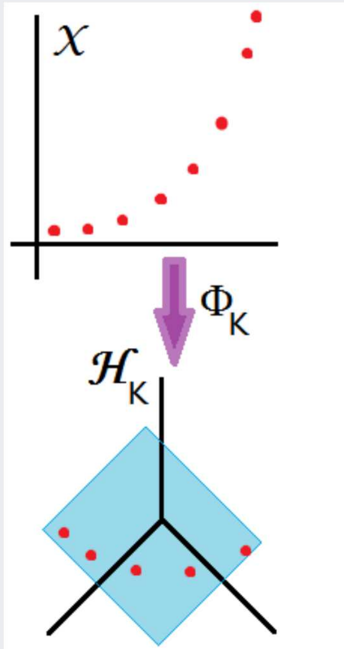
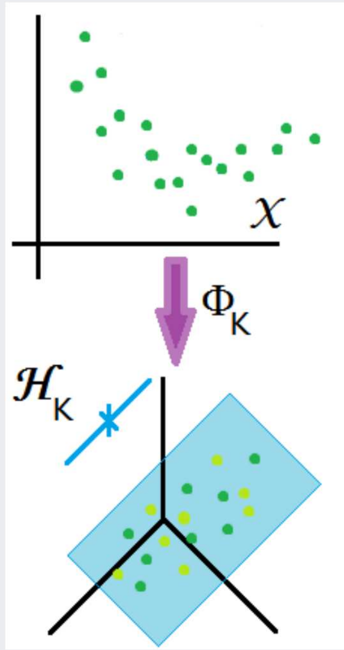
- Kernels exploit various invariances* in these algorithms
  - Use embedding $\Phi_K \colon \mathcal{X} \to \mathcal{H}_K$ where linear models are good !
  - $K(x, x') = \langle \Phi_K(x), \Phi_K(x') \rangle$

Problem	Classification	Regression	PCA
A thousand words			

* Obliviousness to the precise embedding : recall our favorite representer theorems !

Kernel algorithms : an overview

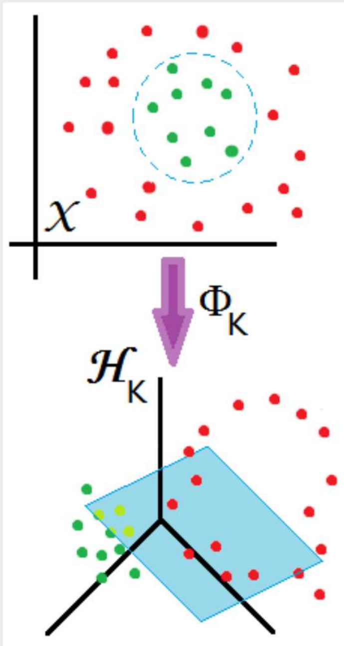
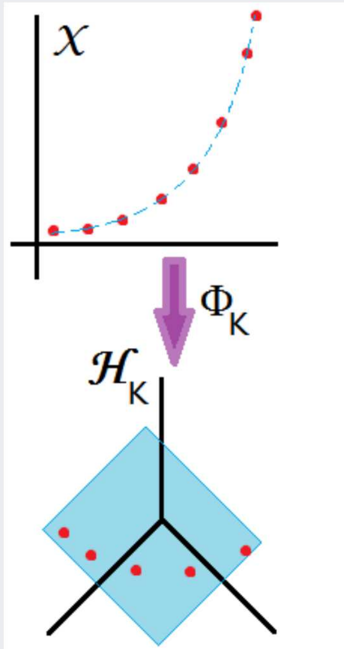
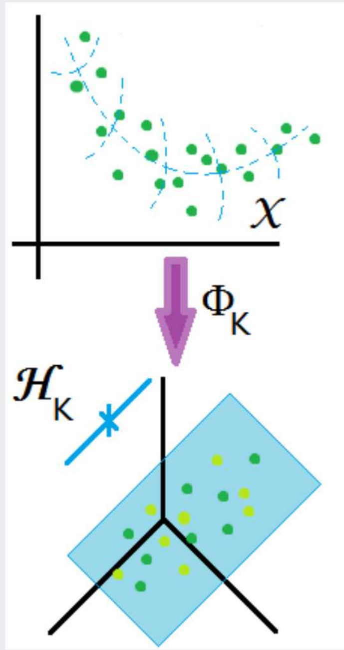
- Kernels exploit various invariances* in these algorithms
  - Use embedding $\Phi_K\colon \mathcal{X} \to \mathcal{H}_K$ where linear models are good !
  - $K(x, x') = \langle \Phi_K(x), \Phi_K(x') \rangle$

Problem	Classification	Regression	PCA
Two thousand words			

* Obliviousness to the precise embedding : recall our favorite representer theorems !

Kernel algorithms : an overview

- Kernels exploit various invariances* in these algorithms
  - Use embedding $\Phi_K\colon \mathcal{X} \to \mathcal{H}_K$ where linear models are good !
  - $K(x, x') = \langle \Phi_K(x), \Phi_K(x') \rangle$

Problem	Classification	Regression	PCA
Two thousand words			

* Obliviousness to the precise embedding : recall our favorite representer theorems !

The price of trickery

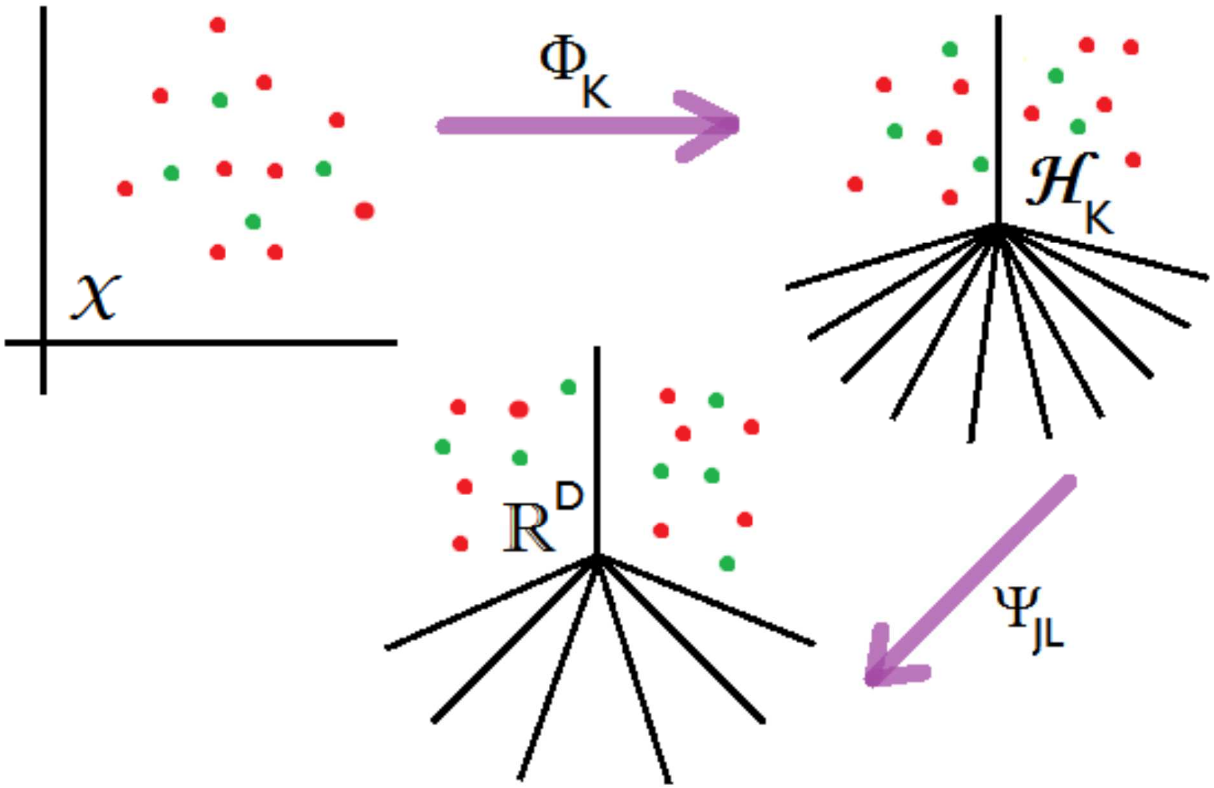
- Frequently, we choose complex kernels i.e. $\dim(\mathcal{H}_K) \gg 1$
  - Requires implicit representations for hypotheses

Problem	Hypothesis (explicit)	Hypothesis (implicit)
Classification	$\mathrm{sgn}(w^\top \Phi_K(x))$	$\mathrm{sgn}\left(\sum \alpha_i y_i K(x, x_i)\right)$
Regression	$w^\top \Phi_K(x)$	$\sum \alpha_i y_i K(x, x_i)$
PCA	$\left(V^k\right)^\top \Phi_K(x)$	$\sum \alpha_i^k K(x, x_i)$
Clustering	$\underset{k}{\mathrm{argmin}} \\|\Phi(x) - \mu_k\\|_{\mathcal{H}_K}^2$	$\underset{k}{\mathrm{argmax}} \left\{\sum \alpha_i^k K(x, x_i) + c_k\right\}$

- Summations frequently run over most of the training set
  - Provably a constant fraction in some cases [Steinwart '03]
  - Expensive test, training routines
  - Explicit forms much cheaper to work with

Dimensionality Reduction

- How about making $\mathcal{H}_K$ finite dimensional
  - JL Lemma : inner product preserving maps $\Psi_{JL}\colon \mathcal{H}_K \to \mathbb{R}^D$
  - Problem : "inductive" implementations require access to $\mathcal{H}_K$



Structure Theorems

- Characterizations for certain kernel families

Kernel family	Representation	Characterization
Translation invariant	$K(x, x') = f(x - x')$	Bochner's theorem over $(\mathbb{R}, +)$
Homogeneous	$K(x, x') = f\left(\frac{x}{x'}\right)$	Bochner's theorem over $(\mathbb{R} \setminus \{0\}, \times)$
Dot Product	$K(x, x') = f(\langle x, x' \rangle)$	Schoenberg's theorem

- Bochner's theorem : $f(x) = \int_\Gamma \gamma(x)\, d\mu(\gamma), \mu \geq 0$
  - $K(x, x') = \int_\Gamma \gamma(x - x') d\mu(\gamma) = \mathrm{E}\left[\gamma(x)\overline{\gamma(x')}\right]$
- Schoenberg's theorem : $f(x) = \sum_{n \geq 0} a_n x^n, a_n \geq 0$
  - $K(x, x') = \sum_{n \geq 0} a_n \langle x, x' \rangle^n = \mathrm{E}[\prod_{i \leq n} \langle \omega_i, x \rangle \prod_{i \leq n} \langle \omega_i, x' \rangle \,|n]$

Random Features

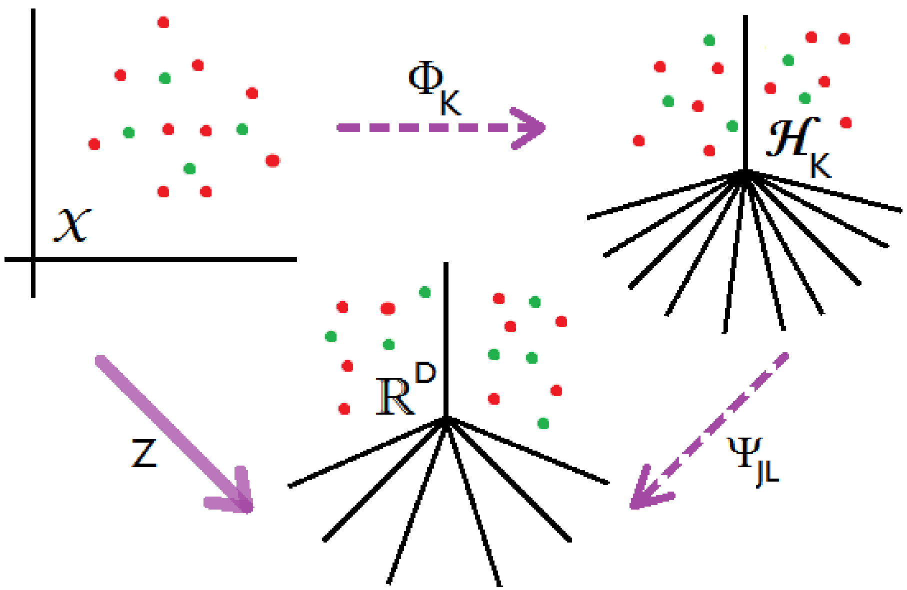
- General form

$$K(x,x') = \int_{\omega\in\Omega} K_\omega(x,x')d\mu(\omega) = \mathrm{E}[K_\omega(x,x')]$$

  - $K_\omega(x,x') = \langle\Phi_\omega(x),\Phi_\omega(x')\rangle$ for $\Phi_\omega : \mathcal{X} \to \mathbb{R}$ is rank-one
- A random such K_ω gives an unbiased estimate of K
  - Independent repetitions give us maps $Z : \mathcal{X} \to \mathbb{R}^D$
  - Think of Z as composing JL and Mercer maps $Z = \Psi_{JL} \circ \Phi_K$
- Guarantee on degree of approximation

If $\mathcal{X} \subset \mathbb{R}^d$ is "compact" and $D = \Omega\left(\frac{d}{\epsilon^2}\log\frac{1}{\epsilon\delta}\right)$ then with prob. $(1-\delta)$, we have $\sup_{x,x'\in\mathcal{X}} |K(x,x') - \langle Z(x), Z(x')\rangle| < \epsilon$

Random Features



Kernel Family	RF Construction
Translation Invariant *Gaussian, Laplacian*	[Rahimi-Recht NIPS'07]
Homogeneous *Chi-Square, Min*	[Vedaldi-Zisserman CVPR'10]
Radial Basis *Exp. Chi-Square*	[Vempati et. al. BMVC'10]
Dot Product *Poly, Exp DP*	[K.-Karnick AISTATS'12]

Random features

- Accelerated training and test routines
  - Comparable/increased accuracies
  - Most experimentation on classification tasks
    - SVMs [VZ,VVZJ,KK], ridge regression [RR]

Dataset	Cod-RNA	Adult	IJCNN	Cover-type
Exact SVM	acc = 95.2% trn = 91.5s tst = 17.1s	acc = 83.7% trn = 263s tst = 33.4s	acc = 98.4% trn = 136s tst = 30s	acc = 80.61% trn = 194s tst = 696s
RF	acc = 94.9% trn = 11.5s (8x) tst = 2.8s (6x) D = 500	acc = 82.9% trn = 40s (6.6x) tst = 14s (2.3x) D = 500	acc = 97.2% trn = 25s (5.5x) tst = 23s (1.3x) D = 1000	acc = 76.2% trn = 21s (9x) tst = 207s (3.6x) D = 1000
RF + (*)™	acc = 93.8% trn = 0.67s (136x) tst = 1.4s (12x) D = 50	acc = 84.8% trn = 7.2s (37x) tst = 9.4s (3.6x) D = 100	acc = 92.2% trn = 5.2s (26x) tst = 9s (3.3x) D = 200	acc = 75.5% trn = 3.7s (52x) tst = 80s (8.7x) D = 100