Multiple Kernel Learning, Sparsity and Heterogeneous Data Fusion

Vladimir Koltchinskii
vlad@math.gatech.edu

Collaborators:
Ming Yuan, Pedro Rangel

Georgia Institute of Technology
School of Mathematics

Fodava Lead:
Dimension Reduction and Data Reduction
December 2009

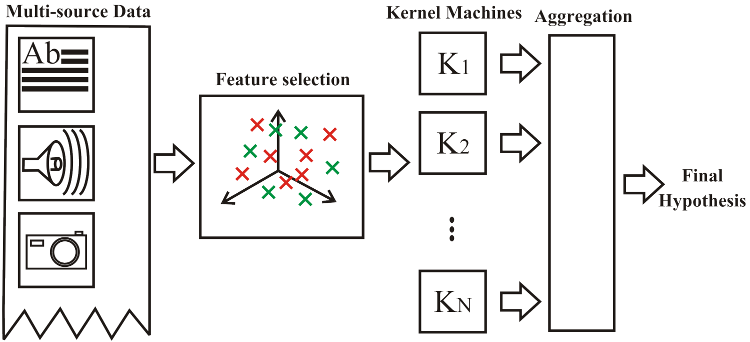
A framework for heterogeneous data fusion

- Lanckriet et al (2004), Bousquet and Herrmann (2003).
- Crammer et al (2003).
- Micchelli and Pontil (2005).
- Srebro and Ben-David (2006).
- ....
- Koltchinskii and Yuan (2008, 2009)

Applications of MKL

- Heterogeneous data fusion in bioinformatics: protein function prediction **Lanckriet et al (2004), ...**
- Text classification: classification of Reuter newswire stories **Lanckriet et al (2004), ...**
- Image annotation: **Harchaoui and Bach (2007), Siddiquie et al (2008), ...**
- Classification of activation patterns in fMRI: **Koltchinskii, Martinez-Ramon et al (2006, 2008)**

Combining multiple-source data using MKL



- Final hypothesis is given by $f(\cdot) = \sum_{j=1}^{n} \sum_{k=1}^{N} \alpha_j^{(k)} K_k(x_j, \cdot)$.

Prediction Problem

- (X, Y) a random couple.
- $X \in S$ a high dimensional "instance".
- $Y \in \mathbb{R}$ a "label" (to be predicted based on X).
- $f : S \mapsto \mathbb{R}$ a prediction rule.
- The risk of f :

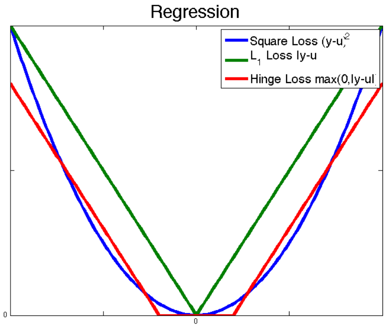
$$L(f) := \mathbb{E}\ell(Y; f(X))$$

where ℓ is a loss function.

Target Function: Optimal Prediction Rule

$$f_* := \operatorname{argmin}_{f:S\mapsto\mathbb{R}} L(f)$$

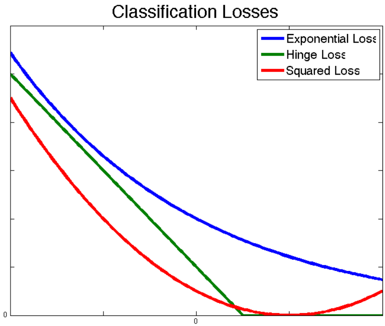
- Regression.



Target Function: Optimal Prediction Rule

$$f_* := \mathrm{argmin}_{f:S\mapsto\mathbb{R}} L(f)$$

- Regression.
- Large margin classification: boosting, kernel machines

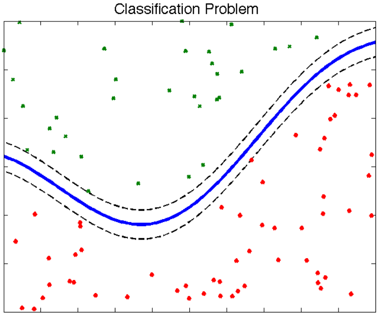


Kernel Trick

- S input data space.
- H feature space (a Hilbert space);
- $\phi : S \mapsto H$ a (nonlinear) embedding of S into H.
- $K(x, y) := \langle \phi(x), \phi(y) \rangle$ kernel;

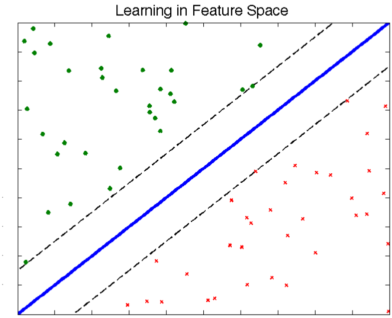
Target Function: Optimal Prediction Rule

- Nonlinear separation in classification problem.



Target Function: Optimal Prediction Rule

- Nonlinear separation in classification problem.
- In an adequate feature space will become a linearly separable classification problem



Kernel Machines: A Single Kernel

- $(X_1, Y_1), \dots, (X_n, Y_n)$ training data.
- $\varepsilon > 0$ a regularization parameter.
- ℓ a convex loss function.
- $L_n(f) := \frac{1}{n}\sum_{j=1}^{n} \ell(Y_j, f(X_j))$ empirical risk.
- K a symmetric nonnegatively definite function on $S \times S$ (a kernel).
- $\mathcal{H}_K$ the reproducing kernel Hilbert space (RKHS) generated by K.

Penalized Empirical Risk Minimization (ERM)

- Let us define

$$\hat{f} := \mathrm{argmin}_{f \in \mathcal{H}_K} \Big[L_n(f) + \varepsilon \|f\|_{\mathcal{H}_K}^{\alpha} \Big]$$

- Usually, $\alpha = 2$ or $\alpha = 1$.
- $\varepsilon > 0$ is a regularization parameter.

Multiple Kernel Learning (MKL): Aggregation of Kernel Machines

- $K_1, \dots, K_N$ kernels (for instance, representing different data sources).
- $\mathcal{H}_1, \dots, \mathcal{H}_N$ the corresponding RKHS.
- Boosting.
- penalized ERM with special penalties (often, based on convex optimization, in particular, semidefinite programming)
- Sparse Problems: the number of kernels N is very large, but only a small number of them is needed to represent the target function.

Two Approaches for sparse recovery in MKL: Infinite Dimensional **LASSO**

- **Koltchinskii and Yuan (2008)**
- $$(\hat{f}_1, \dots, \hat{f}_N) :=$$

$$\operatorname{argmin}_{f_j \in \mathcal{H}_j, j=1,\dots,N} \left[L_n(f_1 + \cdots + f_N) + \varepsilon \sum_{j=1}^{N} \|f_j\|_{\mathcal{H}_j} \right]$$

- Equivalent to (see, e.g., **Micchelli and Pontil (2005)**)

$$(\hat{f}, \hat{K}) := \arg \min_{f \in \mathcal{H}_K, K \in \mathcal{K}} \Big[L_n(f) + \varepsilon \|f\|_{\mathcal{H}_K} \Big],$$

- where

$$\mathcal{K} := \Big\{ \sum_{j=1}^{N} w_j K_j : w_j \geq 0, \sum_{j=1}^{N} w_j = 1 \Big\}$$

- The optimal kernel: $\hat{K} := \sum_{j=1}^{N} \hat{w}_j K_j$
- where the vector $\hat{w}_j \quad (j = 1, \dots, N)$ represents **"relative significance"** of kernels (data sources).

Double Penalization

- Koltchinskii and Yuan (2009). Related to **Sparse Additive Models: Ravikumar et al (2007), Meier et al (2008)**.

$$(\hat{f}_1, \ldots, \hat{f}_N) := \arg \min_{f_j \in \mathcal{H}_j, j=1,\ldots,N} \Big[L_n(f_1 + \cdots + f_N) +$$

$$\sum_{j=1}^{N} \Big(\hat{\varepsilon}_j \|f_j\|_{L_2(\Pi_n)} + \hat{\varepsilon}_j^2 \|f_j\|_{\mathcal{H}_j} \Big) \Big],$$

- where $\hat{\varepsilon}_j$ are data dependent regularization parameters defined in terms of spectra of kernel matrices $\hat{K}_j := \Big(n^{-1} K_j(X_k, X_l) \Big)_{k,l=1,n}$

Spectra and smoothness

- **Additive Representation:** $f = f_1 + \cdots + f_N, \; f_j \in \mathcal{H}_j$.
- Π **the design distribution** = the unknown distribution of X
- $T_{K_j} : L_2(\Pi) \mapsto L_2(\Pi)$ the integral operator with kernel K_j.
- $\{\lambda_k^{(j)}\}$ **distribution dependent eigenvalues** of the operator T_{K_j}.
- The smoothness of the components $f_j \in \mathcal{H}_j$ is related to the rate of decay of $\lambda_k^{(j)}, k \to \infty$.
- The spectrum of $\hat{K}_j$ is a statistical estimate of $\{\lambda_k^{(j)} : k \geq 1\}$.

Adaptive Regularization

- If, for all j,

$$\lambda_k^{(j)} \asymp k^{-2\beta_j}, \ \beta_j > 1/2,$$

  then the optimal choice of regularization parameter $\hat{\varepsilon}_j$ would be $\asymp n^{-\beta_j/(2\beta_j+1)}$

- **Koltchinskii and Yuan (2009)**: a data driven method of choosing $\hat{\varepsilon}_j$ that provides an adaptation to unknown smoothness of the components.

Mathematical Results

- Sparsity Oracle Inequalities:
- Show that, with a high probability, the empirical solution $\hat{f}$ provides the same approximation of the target function f_* as optimal "sparse oracles" up to an error term that depends on the degree of sparsity of the problem.
- Show that in "sparse" problems the empirical solution $\hat{f}$ is "approximately sparse" and its "sparsity pattern" mimics the sparsity pattern of "sparse oracles."

- Koltchinskii and Yuan (2009)
- Suppose

$$\lambda_k^{(j)} \asymp k^{-2\beta_j},\ \beta_j > 1/2$$

and denote

$$\varepsilon_j := n^{-\beta_j/(2\beta_j+1)}.$$

Support of $f : J_f := \{j : f_j \neq 0\}$.

- For all "oracles" $f = (f_1, \dots, f_N)$, with a high probability,

$$L(\hat{f}) - L(f_*) + c_1 \sum_{j=1}^{N} \left(\varepsilon_j \|\hat{f}_j - f_j\|_{L_2(\Pi)} + \varepsilon_j^2 \|\hat{f}_j\|_{\mathcal{H}_j} \right) \leq$$

$$2(L(f) - L(f_*)) + c_2 \sum_{j \in J_f} \varepsilon_j^2 (\gamma^2(J_f) + \|f_j\|_{\mathcal{H}_j}),$$

where $\gamma^2(J_f)$ is a geometric parameter of the dictionary that characterizes the degree of "independence" of spaces $\mathcal{H}_j$.

- Roughly, if the spaces $\mathcal{H}_j, j = 1, \ldots, N$ are "weakly dependent" and all the components are of the same smoothness $\beta_j = \beta$, then the error is controlled by $\mathrm{card}(J_f) n^{-\beta/(2\beta+1)}$, i.e., by the "sparsity" of the problem and by the "smoothness" of the components.
- Note that this learning algorithm relies on neither the knowledge of "sparsity"nor the knowledge of "smoothness". However, it is adaptive to both of them.

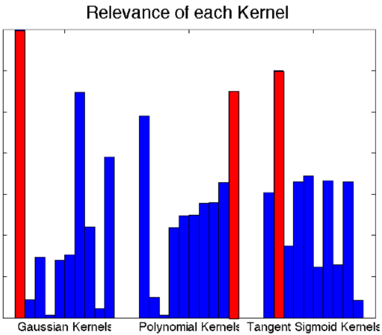
Experiments

- Let $K_{G(\sigma^2)}$ be a Gaussian kernel with variance σ^2 (i.e. $K_{G(\sigma^2)}(x_1, x_2) = \exp(-|x1 - x2|^2/\sigma^2))$.
- Let $K_{P(d)}$ be a polynomial kernel of degree d (i.e. $K_{P(d)} = (x1, x2) = \langle x1, x2\rangle^d)$.
- Let $K_{T(r)}$ be a tangent sigmoid kernel with parameter r (i.e. $K_{T(r)}(x1, x2) = \tanh(r\langle x1, x2\rangle))$.

- Let $f : \mathbb{R}^5 \to \mathbb{R}$ be an unknown function of form $f = f_G + f_P + f_T$, where $f_G \in \mathcal{H}_{K_{G(1)}}$, $f_G \in \mathcal{H}_{K_{P(10)}}$ and $f_G \in \mathcal{H}_{K_{T(2)}}$.
- Consider the problem of find f given a data set $\{\langle x_i, f(x_i) + N_i \rangle\}_{i=1}^{1000}$, where N_i is noise.
- Since f is unknown, we optimize over the linear combination $\mathcal{K}$ of the reproducing kernel Hilbert space of $K_{G(\sigma^2)}$, $K_{P(d)}$ and $K_{T(r)}$ for $\sigma^2 = 1, \ldots 10$, $d = 1, \ldots, 10$, $r = 1, \ldots 10$.
- So, we optimize over 30 different kernels, in such a way that f is a sparse model of $\mathcal{K}$.
- Each kernel represents a possible source of information. With a sparse additive model we are able to identify different sources.

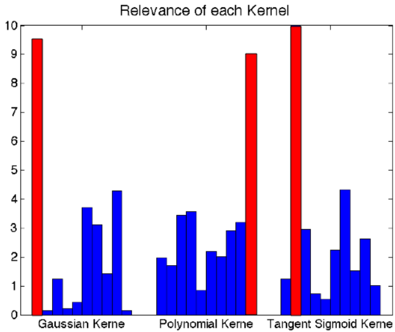
Identification of kernels

- Multiple Kernel Machine (no regularization)



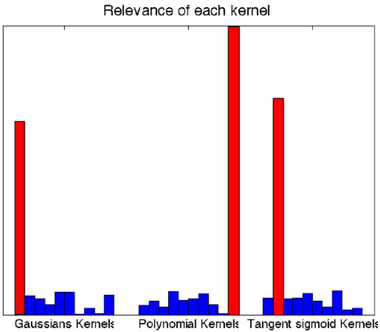
Identification of kernels

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.05$)



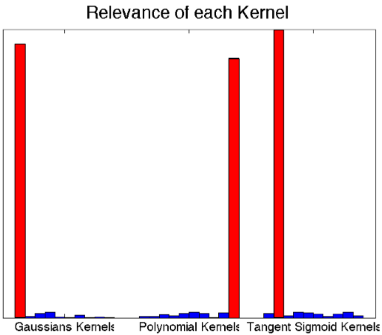
Identification of kernels

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.05$)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.1$)



Identification of kernels

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.05$)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.1$)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.325$)



Georgia Institute of Technology

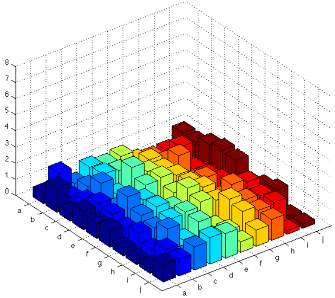
Wisconsin Diagnostic Breast Cancer (WDBC)

- a) radius (mean of distances from center to points on the perimeter).
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter ^{2} / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

- Let $K_{i,j}(x_1, x_2)$ be the dot product of the data points x_1 and x_2 restricted to the coordinates i and j.
- The main idea of choosing this set of kernels is to identify couple of features that characterize the learning problem.
- Using this approach, we identify: Perimeter vs. Smoothness, and Radius vs. Concave points.

Identification of correlated features

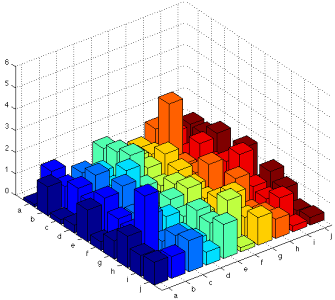
- Multiple Kernel Machine (no regularization)



Georgia Institute of Technology

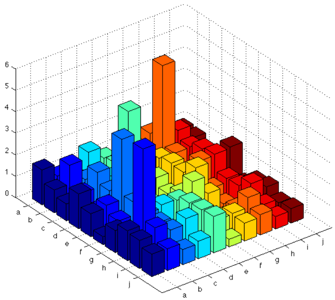
Identification of correlated features

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.2$).



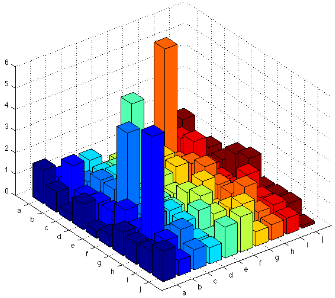
Identification of correlated features

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.2$).
- Sparse Multiple Kernel Machine ($\varepsilon = 0.4$).



Identification of correlated features

- Multiple Kernel Machine (no regularization)
- Sparse Multiple Kernel Machine ($\varepsilon = 0.2$).
- Sparse Multiple Kernel Machine ($\varepsilon = 0.4$).
- Sparse Multiple Kernel Machine ($\varepsilon = 0.6$).



Future Goals

- Computational aspects: convex optimization methods, backfitting, etc; jointly with **Haesun Park, Pedro Rangel**
- Statistical aspects: adaptation methods, feature selection, simulation studies, etc; jointly with **Pedro Rangel, Ming Yuan**
- Visualization: methods of visualizing large ensembles of prediction rules, aggregation maps and other approaches to visualizing relative significance of features