Foundations of Machine Learning
Lecture 5: Part II

Mehryar Mohri (presented today by Marius Kloft)

Courant Institute

March 25, 2013

This Lecture: Part II

Outline

- Kernels on structured objects
- Multiple kernel learning (MKL)

This Lecture: Part II

Outline

- **Kernels on structured objects**
- Multiple kernel learning (MKL)

Motivation

Structured data ubiquitous in applied sciences:

- *Bioinformatics*
  e.g., DNA sequences and metabolic networks
- *Natural language processing*
  e.g., text documents and parse trees
- *Computer security*
  Network traffic and program behavior
- *Cheminformatics*
  molecule structures

Motivation

Structured data ubiquitous in applied sciences:

- *Bioinformatics*
  e.g., DNA sequences (strings) and metabolic networks (graphs)
- *Natural language processing*
  e.g., text documents (strings) and parse trees (trees)
- *Computer security*
  Network traffic and program behavior (strings, trees)
- *Cheminformatics*
  molecule structures (graphs)

Data can be modeled by disrete structures such as strings, trees, and graphs.

Motivation

Structured data ubiquitous in applied sciences:

- *Bioinformatics*
  e.g., DNA sequences (strings) and metabolic networks (graphs)
- *Natural language processing*
  e.g., text documents (strings) and parse trees (trees)
- *Computer security*
  Network traffic and program behavior (strings, trees)
- *Cheminformatics*
  molecule structures (graphs)

Data can be modeled by disrete structures such as strings, trees, and graphs.

Structured data $\neq$ vectors $\Rightarrow$ No machine learning possible?

Example of a String Kernel: Bag-of-Words Kernel

- Bag-of-words: characterization of strings using non-overlapping substrings ("words")

$x =$ `mary had a little lamb`

`mary` `had` `a` `little` `lamb`

- **Definition**: Let L be a language over an alphabet Σ and let $D \subset L$ be a set of delimiters. The *bag-of-words kernel* is defined by

$$\forall x, x' \in \Sigma^* : \qquad k(x, x') = \sum_{w \in L \setminus D} I_w(x) \cdot I_w(x'),$$

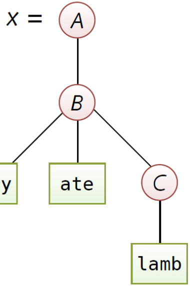
where I denotes the indicator function, i.e., $I_w(x) = 1$ if w is a substring of x, and $I_w(x) = 0$ otherwise.

- The BOW "kernel" is, indeed, a PDS kernel because $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ where

$$\Phi : \begin{array}{ccc} \Sigma^* & \to & \mathbb{R}^{|L \setminus D|} \\ x & \mapsto & (I_w(x))_{w \in L \setminus D} \end{array}.$$

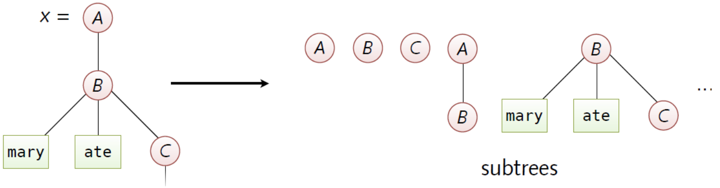
Example of a Tree Kernel: The Parse-Tree Kernel

- A *tree* $x = (V, E, v^*)$ is an acyclic graph (V, E) rooted at $v^* \in V$.
- A *parse tree* is a tree x derived from a grammar, such that each node $v \in V$ is associated with a production rule $p(v)$.
- **Example**: parse tree for "mary ate lamb" has production rules
  - $p_1 : A \to B$
  - $p_2 : B \to$ "*mary*" "*ate*" C
  - $p_3 : C \to$ "*lamb*"



Example of a Tree Kernel: The Parse-Tree Kernel

- Parse trees are common data structure in several application domains, e.g., natural language processing, compiler design, ...
- Characterization of parse trees using contained subtrees



- **Definition**: similar to the bag-of-words kernel, define the *parse-tree kernel* by

$$k(x, x') = \sum_{t \in T} I_t(x)\, I_t(x')\,.$$

Here: T = "set of all possible parse trees", and $I_t(x)$ returns the occurrence of subtree t in x

This Lecture: Part II

Outline

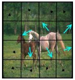
- Kernels on structured objects
- Multiple kernel learning (MKL)

This Lecture: Part II

Outline

- Kernels on structured objects
- **Multiple kernel learning (MKL)**

Motivation

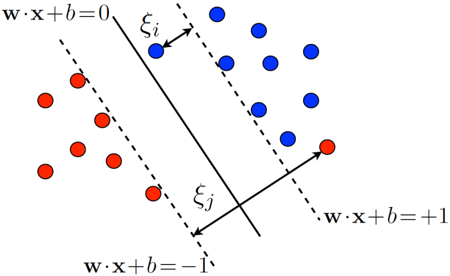
- Several important applications of ML come with multiple views of the data
- For example, in image analysis, an image can be described, in terms of, e.g.:
    - pixel colors 
    - shapes (gradients) 
    - local features 
    - spatial tilings 
- Each view gives rise to one or multiple kernels.

Recap: Support Vector Machine

- Constrained, convex optimization problem:

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|_{\mathbb{H}}^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \ \ y_i(\mathbf{w}\cdot\Phi(x_i)+b) \geq 1-\xi_i \ \ \wedge \ \ \xi_i \geq 0, \ \ i \in [1,m]$$



Multiple Kernel Learning

- Let $K_1, \ldots, K_d : X \times X \to \mathbb{R}$ be PDS kernels, associated with respective feature maps $\Phi_j : X \to \mathbb{H}_j$, $j \in [1, d]$
- Consider "weighted" Cartesian product feature space $\Phi_{\boldsymbol{\theta}} := \sqrt{\theta_1}\Phi_1 \times \cdots \times \sqrt{\theta_d}\Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights
  - corresponds to weighted kernel $K_{\boldsymbol{\theta}} := \theta_1 K_1 + \cdots + \theta_d K_d$

Multiple Kernel Learning

- Let $K_1, \ldots, K_d : X \times X \to \mathbb{R}$ be PDS kernels, associated with respective feature maps $\Phi_j : X \to \mathbb{H}_j$, $j \in [1, d]$
- Consider "weighted" Cartesian product feature space $\Phi_{\boldsymbol{\theta}} := \sqrt{\theta_1}\Phi_1 \times \cdots \times \sqrt{\theta_d}\Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights
  - corresponds to weighted kernel $K_{\boldsymbol{\theta}} := \theta_1 K_1 + \cdots + \theta_d K_d$
- SVM optimization problem:

$$\min_{\mathbf{w}, b, \xi} \; \frac{1}{2} \|\mathbf{w}\|_{\mathbb{H}}^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to} \;\; y_i(\mathbf{w} \cdot \Phi_{\boldsymbol{\theta}}(x_i) + b) \geq 1 - \xi_i \;\; \wedge \;\; \xi_i \geq 0, \;\; i \in [1, m]$$

Multiple Kernel Learning

- Let $K_1, \ldots, K_d : X \times X \to \mathbb{R}$ be PDS kernels, associated with respective feature maps $\Phi_j : X \to \mathbb{H}_j$, $j \in [1, d]$
- Consider "weighted" Cartesian product feature space $\Phi_{\boldsymbol{\theta}} := \sqrt{\theta_1}\Phi_1 \times \cdots \times \sqrt{\theta_d}\Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights
  - corresponds to weighted kernel $K_{\boldsymbol{\theta}} := \theta_1 K_1 + \cdots + \theta_d K_d$
- SVM optimization problem:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \underbrace{\|\mathbf{w}\|^2_{\mathbb{H}}}_{=\sum_{j=1}^d \|\mathbf{w}_j\|^2_{\mathbb{H}_j}} + C \sum_{i=1}^m \xi_i$$

$$\text{subject to} \quad y_i(\underbrace{\mathbf{w} \cdot \Phi_{\boldsymbol{\theta}}(x_i)}_{=\sum_{j=1}^d \sqrt{\theta_j}\mathbf{w}_j \cdot \Phi_j} + b) \geq 1 - \xi_i \ \wedge \ \xi_i \geq 0, \ i \in [1, m]$$

where $\mathbf{w} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_d^\top)^\top$

Multiple Kernel Learning

- Let $K_1, \ldots, K_d : X \times X \to \mathbb{R}$ be PDS kernels, associated with respective feature maps $\Phi_j : X \to \mathbb{H}_j$, $j \in [1, d]$
- Consider "weighted" Cartesian product feature space $\Phi_{\boldsymbol{\theta}} := \sqrt{\theta_1}\Phi_1 \times \cdots \times \sqrt{\theta_d}\Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights
  - corresponds to weighted kernel $K_{\boldsymbol{\theta}} := \theta_1 K_1 + \cdots + \theta_d K_d$
- SVM optimization problem:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \underbrace{\|\mathbf{w}\|_{\mathbb{H}}^2}_{=\sum_{j=1}^d \|\mathbf{w}_j\|_{\mathbb{H}_j}^2} + C \sum_{i=1}^m \xi_i$$

$$\text{subject to} \quad y_i(\underbrace{\mathbf{w} \cdot \Phi_{\boldsymbol{\theta}}(x_i)}_{=\sum_{j=1}^d \sqrt{\theta_j}\mathbf{w}_j \cdot \Phi_j} + b) \geq 1 - \xi_i \ \wedge \ \xi_i \geq 0, \ i \in [1, m]$$

where $\mathbf{w} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_d^\top)^\top$

- How to compute a "good" weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$?

Multiple Kernel Learning (MKL)

- MKL optimization problem:

$$\min_{\mathbf{w},b,\xi,\boldsymbol{\theta}:\boldsymbol{\theta}\geq 0\|\boldsymbol{\theta}\|\leq 1} \quad \frac{1}{2}\sum_{j=1}^{d}\|\mathbf{w}_j\|^2_{\mathbb{H}_j} + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i\Big(\sum_{j=1}^{d}\sqrt{\theta_j}\mathbf{w}_j\cdot\Phi_j(x_i)+b\Big)\geq 1-\xi_i \quad\wedge\quad \xi_i\geq 0$$

- Core idea:
  - Optimize over the kernel weights $\theta_1,\ldots,\theta_d$
  - Restrict $\|\boldsymbol{\theta}\|$ to avoid overfitting
    - ★ In the following $\|\boldsymbol{\theta}\|\equiv\|\boldsymbol{\theta}\|_p \stackrel{\text{def.}}{=} (\sum_{j=1}^{d}|\theta_j|^p)^{\frac{1}{p}}$ ("ℓ_p-norm")

Multiple Kernel Learning (MKL)

- MKL optimization problem:

$$\min_{\mathbf{w},b,\xi,\boldsymbol{\theta}:\boldsymbol{\theta}\geq 0 \|\boldsymbol{\theta}\|\leq 1} \quad \frac{1}{2}\sum_{j=1}^{d}\|\mathbf{w}_j\|^2_{\mathbb{H}_j} + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i\left(\sum_{j=1}^{d}\sqrt{\theta_j}\mathbf{w}_j\cdot\Phi_j(x_i)+b\right)\geq 1-\xi_i \quad \wedge \quad \xi_i\geq 0$$

- Core idea:
  - Optimize over the kernel weights $\theta_1,\ldots,\theta_d$
  - Restrict $\|\boldsymbol{\theta}\|$ to avoid overfitting
    - ★ In the following $\|\boldsymbol{\theta}\| \equiv \|\boldsymbol{\theta}\|_p \stackrel{\text{def.}}{=} \left(\sum_{j=1}^{d}|\theta_j|^p\right)^{\frac{1}{p}}$ ("ℓ_p-norm")

- Problem: OP is not convex because of the mixed products $\sqrt{\theta_j}\mathbf{w}_j$

Multiple Kernel Learning (MKL)

- Change of variables: $\mathbf{w}_j^{\text{new}} := \sqrt{\theta_j}\mathbf{w}_j^{\text{old}}$

Multiple Kernel Learning (MKL)

- Change of variables: $\mathbf{w}_j^{\text{new}} := \sqrt{\theta_j}\mathbf{w}_j^{\text{old}}$

  $\Rightarrow$ Equivalent MKL optimization problem:

$$\min_{\mathbf{w},b,\xi,\boldsymbol{\theta}:\boldsymbol{\theta}\geq 0\|\boldsymbol{\theta}\|_p\leq 1} \quad \frac{1}{2}\sum_{j=1}^{d}\frac{\|\mathbf{w}_j\|_{\mathbb{H}_j}^2}{\theta_j} + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i\Big(\underbrace{\sum_{j=1}^{d}\mathbf{w}_j\cdot\Phi_j(x_i)}_{=\mathbf{w}\cdot\Phi(x_i)}+b\Big)\geq 1-\xi_i \;\;\wedge\;\; \xi_i\geq 0,\;\; i\in[1,m]$$

Multiple Kernel Learning (MKL)

- Change of variables: $\mathbf{w}_j^{\text{new}} := \sqrt{\theta_j}\mathbf{w}_j^{\text{old}}$

  $\Rightarrow$ Equivalent MKL optimization problem:

$$\min_{\mathbf{w},b,\xi,\boldsymbol{\theta}:\boldsymbol{\theta}\geq 0 \|\boldsymbol{\theta}\|_p\leq 1} \quad \frac{1}{2}\sum_{j=1}^{d}\frac{\|\mathbf{w}_j\|_{\mathbb{H}_j}^2}{\theta_j} + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i\Big(\underbrace{\sum_{j=1}^{d}\mathbf{w}_j\cdot\Phi_j(x_i)}_{=\mathbf{w}\cdot\Phi(x_i)}+b\Big) \geq 1-\xi_i \quad \wedge \quad \xi_i \geq 0, \ \ i \in [1,m]$$

- Convex problem: because any function $(\mathbf{x}, y) \mapsto \frac{\mathbf{x}M\mathbf{x}}{y}$ with positive semi-definite M is convex for $y > 0$
- Convention: $0/0 := 0$ and $x/0 := \infty$ for $x \neq 0$

Rademacher Complexity of MKL

Theorem

Let $K_1, \ldots, K_d : X \times X \to \mathbb{R}$ be PDS kernels with associated feature mappings $\Phi_j : X \to \mathbb{H}_j$, $j \in [1, d]$. Let $S \subseteq \{x : K_j(x, x) \leq R^2, j \in [1, d]\}$ be a sample of size m, put $q := 2p/(p+1)$ and $q^* := q/(q-1)$, and let $H = \{x \mapsto \mathbf{w} \cdot \Phi(x) : \sum_{j=1}^d \|\mathbf{w}_j\|_{\mathbb{H}_j}^2 / \theta_j \leq \Lambda^2, \boldsymbol{\theta} \geq 0, \|\boldsymbol{\theta}\|_p \leq 1\}$. Then,

$$\widehat{\mathfrak{R}}_S(H) \leq \frac{\Lambda}{m}\sqrt{c\,\|(\mathrm{Tr}(\mathbf{K}_1), \ldots, \mathrm{Tr}(\mathbf{K}_d))\|_{\frac{q^*}{2}}} \leq \sqrt{\frac{c}{m}}\Lambda R d^{1/q^*}, \;\; c := \max(1, q^* - 1).$$

Proof.

First note that, $\min_{\boldsymbol{\theta} \geq 0, \|\boldsymbol{\theta}\|_p \leq 1} \sum_{j=1}^d \frac{a_j^2}{\theta_j} = \|(a_1, \ldots, a_d)\|_q^2$ with $q = 2p/(p+1)$ for any $a_1, \ldots, a_d \in \mathbb{R}$. Thus, denoting $\|\mathbf{w}\|_{2,q} := \left\|(\|\mathbf{w}_1\|_{\mathbb{H}_1}, \ldots, \|\mathbf{w}_d\|_{\mathbb{H}_d})\right\|_q$,

$$\widehat{\mathfrak{R}}_S(H) = \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\sum_{j=1}^d \|\mathbf{w}_j\|_{\mathbb{H}_j}^2/\theta_j \leq \Lambda^2 \\ \boldsymbol{\theta} \geq 0, \|\boldsymbol{\theta}\|_p \leq 1}} \mathbf{w} \cdot \sum_{i=1}^m \sigma_i \Phi(x_i)\right] = \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_{2,q} \leq \Lambda} \mathbf{w} \cdot \sum_{i=1}^m \sigma_i \Phi(x_i)\right]$$

$$\overset{(*)}{\leq} \frac{\Lambda}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i \Phi(x_i)\right\|_{2,q^*}\right] \overset{(**)}{\leq} \frac{\Lambda}{m}\left(\sum_{j=1}^d \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^m \sigma_i \Phi_j(x_i)\right\|_{\mathbb{H}_j}^{q^*}\right)^{1/q^*}$$

$$\overset{(***)}{\leq} \frac{\Lambda\sqrt{c}}{m}\left(\sum_{j=1}^d \left(\sum_{i=1}^m \|\Phi_j(x_i)\|_{\mathbb{H}_j}^2\right)^{q^*/2}\right)^{1/q^*} = \frac{\Lambda\sqrt{c}}{m}\sqrt{\|(\mathrm{Tr}(\mathbf{K}_1), \ldots, \mathrm{Tr}(\mathbf{K}_d))\|_{\frac{q^*}{2}}}$$

where $(*)$, $(**)$, and $(***)$, is by Hölder's, Jensen's, and Khintchine/Kahane's inequality. □