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 discrete 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 discrete structures such as **strings**, **trees**, and **graphs**.

Structured data ≠ vectors ⇒ No machine learning possible?
Example of a String Kernel: Bag-of-Words Kernel

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

  \[
  x = \text{mary had a little lamb}
  \]

  \[
  \text{mary had a little lamb}
  \]

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

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

  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

  \[
  \Phi : \Sigma^* \rightarrow \mathbb{R}^{|L \setminus D|}
  \]

  \[
  x \mapsto (I_w(x))_{w \in L \setminus D}
  \]
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

- $p_1 : A \rightarrow B$
- $p_2 : B \rightarrow "mary" \ "ate" C$
- $p_3 : C \rightarrow "lamb"$
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

\[
k(x, x') = \sum_{t \in T} l_t(x) l_t(x').
\]

Here: \( T = \text{“set of all possible parse trees”} \), and \( l_t(x) \) 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:

\[
\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} & \quad y_i (w \cdot \Phi(x_i) + b) \geq 1 - \xi_i \quad \land \quad \xi_i \geq 0, \quad i \in [1, m]
\end{align*}
\]

\[w \cdot x + b = 0\]
\[w \cdot x + b = +1\]
\[w \cdot x + b = -1\]
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
  
  $\Phi_\theta := \sqrt{\theta_1} \Phi_1 \times \cdots \times \sqrt{\theta_d} \Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights

  - corresponds to weighted kernel $K_\theta := \theta_1 K_1 + \cdots + \theta_d K_d$
Multiple Kernel Learning

- Let $K_1, \ldots, K_d : X \times X \rightarrow \mathbb{R}$ be PDS kernels, associated with respective feature maps $\Phi_j : X \rightarrow \mathbb{H}_j$, $j \in [1, d]$

- Consider “weighted” Cartesian product feature space
  $\Phi_{\theta} := \sqrt{\theta_1} \Phi_1 \times \cdots \times \sqrt{\theta_d} \Phi_d$ where $\theta_1, \ldots, \theta_d \geq 0$ are weights
    ▶ corresponds to weighted kernel $K_{\theta} := \theta_1 K_1 + \cdots + \theta_d K_d$

- SVM optimization problem:

\[
\min_{w, b, \xi} \quad \frac{1}{2} \|w\|_H^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to } y_i (w \cdot \Phi_{\theta}(x_i) + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, \quad i \in [1, m]
\]
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
  \( \Phi_\theta := \sqrt{\theta_1} \Phi_1 \times \cdots \times \sqrt{\theta_d} \Phi_d \) where \( \theta_1, \ldots, \theta_d \geq 0 \) are weights
  ▶ corresponds to weighted kernel \( K_\theta := \theta_1 K_1 + \cdots + \theta_d K_d \)

- SVM optimization problem:

\[
\begin{align*}
&\min_{w, b, \xi} \frac{1}{2} \left\| w \right\|_{\mathbb{H}}^2 + C \sum_{i=1}^m \xi_i \\
&\quad = \sum_{j=1}^d \left\| w_j \right\|_{\mathbb{H}_j}^2 \\
\text{subject to} \quad &y_i \left( \underbrace{w \cdot \Phi_\theta(x_i)}_{\Phi_\theta} + b \right) \geq 1 - \xi_i \land \xi_i \geq 0, \; i \in [1, m] \\
&= \sum_{j=1}^d \sqrt{\theta_j} w_j \cdot \Phi_j
\end{align*}
\]

where \( w = (w_1^\top, \ldots, w_d^\top)^\top \)
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
  \( \Phi_{\theta} := \sqrt{\theta_1} \Phi_1 \times \cdots \times \sqrt{\theta_d} \Phi_d \) where \( \theta_1, \ldots, \theta_d \geq 0 \) are weights
  - corresponds to weighted kernel \( K_{\theta} := \theta_1 K_1 + \cdots + \theta_d K_d \)

- SVM optimization problem:
  \[
  \min_{w, b, \xi} \begin{align*}
  & \frac{1}{2} \left\| w \right\|_{\mathbb{H}}^2 + C \sum_{i=1}^{m} \xi_i \\
  \text{subject to} & \quad y_i \left( w \cdot \Phi_{\theta}(x_i) + b \right) \geq 1 - \xi_i \wedge \xi_i \geq 0, \ i \in [1, m]
  \end{align*}
  \]

  where \( w = (w_1^\top, \ldots, w_d^\top)^\top \)

- How to compute a “good” weight vector \( \theta = (\theta_1, \ldots, \theta_d) \)?
Multiple Kernel Learning (MKL)

MKL optimization problem:

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

subject to

\[
y_i \left( \sum_{j=1}^{d} \sqrt{\theta_j w_j \cdot \Phi_j(x_i) + b} \right) \geq 1 - \xi_i \quad \land \quad \xi_i \geq 0
\]

Core idea:

- Optimize over the kernel weights \(\theta_1, \ldots, \theta_d\)
- Restrict \(\|\theta\|\) to avoid overfitting
  - In the following \(\|\theta\| \equiv \|\theta\|_p \overset{\text{def.}}{=} (\sum_{j=1}^{d} |\theta_j|^p)^{\frac{1}{p}}\) (“\(\ell_p\)-norm”)
Multiple Kernel Learning (MKL)

- MKL optimization problem:

\[
\min_{w,b,\xi,\theta: \theta \geq 0, \|\theta\| \leq 1} \frac{1}{2} \sum_{j=1}^{d} \|w_j\|_{H_j}^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to } y_i \left( \sum_{j=1}^{d} \sqrt{\theta_j} w_j \cdot \Phi_j(x_i) + b \right) \geq 1 - \xi_i \land \xi_i \geq 0
\]

- Core idea:
  - Optimize over the kernel weights \(\theta_1, \ldots, \theta_d\)
  - Restrict \(\|\theta\|\) to avoid overfitting

  ★ In the following \(\|\theta\| \equiv \|\theta\|_{\rho} \overset{\text{def.}}{=} (\sum_{j=1}^{d} |\theta_j|^p)^{\frac{1}{p}}\) (\(\ell_p\)-norm)

- Problem: OP is not convex because of the mixed products \(\sqrt{\theta_j} w_j\)
Multiple Kernel Learning (MKL)

- Change of variables: \( w_j^{\text{new}} := \sqrt{\theta_j} w_j^{\text{old}} \)
Multiple Kernel Learning (MKL)

- Change of variables: \( w_j^{\text{new}} := \sqrt{\theta_j} w_j^{\text{old}} \)

\[ \Rightarrow \text{Equivalent MKL optimization problem:} \]

\[
\min_{w, b, \xi, \theta: \theta \geq 0, \|\theta\|_p \leq 1} \quad \frac{1}{2} \sum_{j=1}^{d} \frac{\|w_j\|^2}{\theta_j} + C \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i \left( \sum_{j=1}^{d} w_j \cdot \Phi_j(x_i) + b \right) \geq 1 - \xi_i \land \xi_i \geq 0, \ i \in [1, m] \)

\[ = w \cdot \Phi(x_i) \]
Multiple Kernel Learning (MKL)

- Change of variables: \( w_{j}^{\text{new}} := \sqrt{\theta_j} w_{j}^{\text{old}} \)

\[ \Rightarrow \text{Equivalent MKL optimization problem:} \]

\[
\min_{w, b, \xi, \theta: \theta \geq 0 \|\theta\|_p \leq 1} \frac{1}{2} \sum_{j=1}^{d} \frac{\|w_j\|^2}{\theta_j} + C \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i \left( \sum_{j=1}^{d} w_j \cdot \Phi_j(x_i) + b \right) \geq 1 - \xi_i \land \xi_i \geq 0, \ i \in [1, m] \)

\[ = w \cdot \Phi(x_i) \]

- Convex problem: because any function \((x, y) \mapsto \frac{xMx}{y}\) 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 \( \Phi_j : X \to H_j \), \( j \in [1, d] \). 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 \( H = \{ x \mapsto w \cdot \Phi(x) : \sum_{j=1}^d \| w_j \|_{H_j}^2 / \theta_j \leq \Lambda^2, \theta \geq 0, \| \theta \|_p \leq 1 \} \). Then,

\[
\hat{R}_S(H) \leq \frac{\Lambda}{m} \sqrt{c \| (\text{Tr}(K_1), \ldots, \text{Tr}(K_d)) \|_{q^*}^{2}} \leq \sqrt{c \| \Lambda R d^{1/q^*}, \ c := \max(1, q^* - 1).}
\]

**Proof.**

First note that, \( \min_{\theta \geq 0, \| \theta \|_p \leq 1} \sum_{j=1}^d \frac{a_j^2}{\theta_j} = \| (a_1, \ldots, a_d) \|_q^2 \) with \( q = 2p/(p + 1) \) for any \( a_1, \ldots, a_d \in \mathbb{R} \). Thus, denoting \( \| w \|_{2,q} := \| (\| w_1 \|_{H_1}, \ldots, \| w_d \|_{H_d}) \|_q \),

\[
\hat{R}_S(H) = \frac{1}{m} \mathbb{E} \left[ \sup_{\theta \geq 0, \| \theta \|_p \leq 1} \mathbb{E} \left[ \sup_{\Sigma_{j=1}^d \| w_j \|_{H_j}^2 / \theta_j \leq \Lambda^2} \| w \cdot \sum_{i=1}^m \sigma_i \Phi(x_i) \|_{2,q^*} \right] \right] = \frac{1}{m} \mathbb{E} \left[ \sup_{\| w \|_{2,q} \leq \Lambda} \| w \cdot \sum_{i=1}^m \sigma_i \Phi(x_i) \|_{2,q^*} \right]
\]

\[
\leq \frac{\Lambda}{m} \mathbb{E} \left[ \| \sum_{i=1}^m \sigma_i \Phi(x_i) \|_{2,q^*} \right] \leq \frac{\Lambda}{m} \left( \sum_{j=1}^d \mathbb{E} \left[ \| \sigma_i \Phi_j(x_i) \|_{H_j} \right] \right)^{q^* / 2}
\]

\[
\leq \frac{\Lambda \sqrt{c}}{m} \left( \sum_{j=1}^d \left( \sum_{i=1}^m \| \Phi_j(x_i) \|_{H_j}^2 \right) q^* / 2 \right)^{1/q^*}
\]

where \((*)\), \((***)\), and \((***)\), is by Hölder’s, Jensen’s, and Khintchine/Kahane’s inequality.